Author:
Smith, Helen

Title:
Artificial intelligence use in clinical decision-making

allocating ethical and legal responsibility

Artificial intelligence use in clinical decision-making: allocating ethical and legal responsibility

Helen Smith

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Health Sciences.

Centre for Ethics in Medicine, Bristol Medical School,
Population Health Sciences.

January 2022

Word Count: 85,936

Abstract

Artificial intelligence use in clinical decision-making: allocating ethical and legal responsibility

Advances in computer science have resulted in the development of artificially intelligent systems (AISs) designed for deployment in healthcare environments. There is a potential risk of patient harm eventuating if an AIS dispenses an output which is inappropriate for a patient and a clinician's decision-making is influenced by that output. Because of this potential risk, the ethical and legal consequences of AIS used must be considered and planned for prior to AIS deployment.

My literature review noted neither case law nor legislation in the law of England and Wales specific to negligence in the use of AISs in clinical decision-making. This informs two research questions:

- How, according to current law in England and Wales, will legal liability be allocated between clinicians and software developing companies (SDCs) when AISs are used in clinical decision-making?
- How can ethical responsibility for the consequences of the use of AIS in clinical decision-making be determined and allocated?

My legal analysis finds that clinicians risk shouldering the burden of a negligence claim despite the SDCs actions of supplying the AIS. Using ethical theory, I determine that it is unfair for clinical users to solely shoulder responsibility as an SDC is also causally responsible for harms resulting from the use of their AIS's outputs.

To achieve a fair balance of responsibility between the clinician and the SDC when AISs are used in clinical decision-making, I propose a shared model of responsibility informed by contractarian theories.

To exemplify this approach, I present the concept of risk pooling. This solution: 1) addresses the problem of clinicians being used as moral and legal 'crumple zones'; 2) offers SDCs the opportunity to proactively accept responsibility for the effects of their AISs on a clinician's decision-making; and 3) makes provision for patients who may be harmed as a result of AIS use.

Dedication

For Mr ChrisP.

What a long, strange journey it has been.

Acknowledgements

This thesis was, most certainly, not written in solitude. The following few words of acknowledgement do not do justice to the gratitude which I feel to each and every person who helped me to complete this work.

The greatest thanks go to Professor Jonathan Ives, Dr Giles Birchley, and Professor Andrew Charlesworth. It was their generous giving of knowledge, wisdom, feedback, advice, support, and unbounded kindness which made this project even faintly possible.

Kit Fotheringham's shared interest in AI permitted the fun venture of our first joint publication (featured in chapter 5); our work prompted the ethical discussion which fuelled the remainder of the thesis.

Dr Helen Tunbridge showed me the PhD path, inspired me to apply, and helped push me forward.

Dr Verity Jones's support, commitment, encouragement, interest, and enthusiasm in my work helped carry me through.

My gratitude for the limitless love and support from my husband, Chris, is beyond expression.

These and many others helped me along this journey; without you all this would have been impossible.

I thank you all.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

Signed: Helen Smith

Dated: 17th January 2022

Work published in connection with this thesis

1. Smith, H., 2020. If they asked you to jump off a cliff? In: Paglia, V., & Pegoraro, R. (eds), 2020*. AI and clinical decision making. The "Good" Algorithm?: Artificial Intelligence: Ethics, Law, Health. XXVI General Assembly of Members 2020.* Rome: Pontifical Academy of Life
2. Smith, H., 2021. Clinical AI: opacity, accountability, responsibility and liability. *AI & Society.* 36(2): 535-545

   This published paper is a version of chapter 4's literature review.

   Wording from this paper is also used in chapters 1 and 3.
3. Smith, H. & Fotheringham, K., 2020. Artificial intelligence in clinical decision-making: Rethinking liability. *Medical Law International* 20(2): 131-154

   This published paper is a version of chapter 5's legal analysis and the basis of some of chapter 7's solutions.

   Wording from this paper is also used in chapter 1.

   The first draft of this paper was completed by HS. The second draft of this paper was completed by KF. HS and KF completed further iterative redrafting together. Major revisions from the journal's reviewers were firstly addressed by HS and then checked for accuracy, language issues and further additional refinements by KF.
4. Smith, H., 2020. The challenge of fairly allocating ethical and legal responsibility for the outcomes of using artificial intelligence to inform clinical decision-making. *EACME Newsletter*, Number 56, December 2020.
5. Smith, H., 2021. Artificial Intelligence to Inform Clinical Decision Making: A Practical Solution to An Ethical and Legal Challenge. *IJCAI 2020 Workshop on AI for Social Good.* Harvard CRCS. https://crcs.seas.harvard.edu/publications/artificial-intelligence-inform-clinical-decision-making-practical-solution-ethical

   Wording from this paper is used in chapters 6 and 7.
6. Smith, H., Birchley, G., & Ives, J., 2022. Artificial intelligence in clinical decision-making: rethinking personal moral responsibility – abstract accepted for the 16th Word Congress of Bioethics in 2022

   This paper is based on the ethical analysis of chapter 6's ethical analysis. The first draft of this paper was completed by HS. Comments, additions, and refinements were made by GB and JI. All authors agreed on the final version prior to submission to WBC.
7. Smith, H. & Fotheringham, K., 2022. Exploring Remedies for Defective Artificial Intelligence Aids in Clinical Decision Making in post-Brexit England and Wales. *Medical Law International.*

22(1):p. 33-51 27 DOI: 10.1177/09685332221076124

The first draft of this paper was completed by HS. The second draft of this paper was completed by KF. HS and KF completed further iterative redrafting together. Major revisions from the journal's reviewers were firstly addressed by HS and then checked for accuracy, language issues and further additional refinements by KF.

Signed: Helen Smith

Dated: 17th January 2022

Contents

Abstract 2

Dedication 3

Acknowledgements 4

Author's declaration 5

Work published in connection with this thesis 6

Chapter 1: Introduction 13

Scene setting 13

Intentional limitations of this thesis 17

Included stakeholder groups 18

Excluded stakeholder groups 19

Assumptions of this thesis 21

Thesis approach and outline 22

Conclusion 25

Chapter 2: How do clinicians currently make decisions? 27

Models of clinical decision-making 28

Some problems with human decision-making 31

Uncertainty 31

Bias 31

Too much information 33

Too much medicine 35

Individuality 36

How might AI be employed to help address these issues in clinical decision-making? 37

Conclusion 38

Chapter 3: What is AI? 39

Defining and Describing AI 39

Opacity 42

What Are the Ethical/Legal/Professional Challenges of Using AIS in Clinical Decision-Making? 47

Opacity 47

Consequences of AIS Opacity: Accountability, Responsibility, and Liability 48

Conclusion 50

Chapter 4: Literature Review 52

Aims 52

Methods 52

Step 1: Formulate the review question 52

Step 2: Identify all the literature that meets the eligibility criteria 53

Step 3: Extract and synthesise data by the allocation of pertinent points to theme headings.... 56
Step 4: Derive and present results organised by themes 56
Findings 56
Why is accountability important? 57
How does opacity interfere with accountability? 59
Examples of opaque AI scenarios 61
Responsibility 63
Should multiple stakeholders hold responsibility for a system's use rather than an individual? 65
Liability 66
In summary 67
Limitations 68
Updates to this literature review 69
Conclusion 75
Chapter 5: Concerning the tort of negligence 77
A note on vicarious liability 80
Clinical negligence 81
Software developer company's negligence 84
Duty of Care 85
Breach of Duty 93
Causation 96
Using volenti as a defence 100
Preparing should the worst happen 102
Conclusion 102
Chapter 6: How does ethical theory inform the allocation of responsibility for the use of AIS in clinical decision-making? 106
A note on ethics and morals 107
Why do ethics and morality matter to the conduct of law? 107
Current theories of ethics 109
Utilitarian approaches 110
Egalitarian approaches 113
Communitarian approaches 116
Contractarian approaches 119
"*But-for*" when determining causation 130
But-for is over-inclusive 131
But-for is imprecise, difficult to prove, and can result in pre-emption 134
But-for risks resulting in overdetermination 135

Addressing the problems of *but-for* causation 137

Linking morally significant causation and moral responsibility 137

Personal moral responsibility........ 138

Prospective personal moral responsibility........ 140

Retrospective personal moral responsibility 141

Voluntariness and Foreseeability........ 142

Practical application of ethical theory 146

Technologist’s deployment of an AIS which is not perfectly accurate 147

User error........ 148

Identifying the ‘moral crumple zone’........ 152

How might the legal position be challenged using ethical theory?........ 153

Conclusion........ 155

Chapter 7: Solutions........ 157

Current practical models of responsibility are inadequate 157

Regulating stakeholders........ 158

Regulating the AIS........ 162

Retrospective negligence claims........ 164

Introducing a shared model of responsibility 166

Mixed prospective/retrospective approaches incorporating a shared model of responsibility 169

New Zealand- “No Fault” system. 171

Risk pooling 172

Next steps: consultation before AIS adoption 193

Consultation........ 194

Negotiation 196

Review........ 196

Empirical work 196

Conclusion........ 198

Chapter 8: Conclusion 199

Finally 201

References 202

Cases 202

All other references 204

Appendix A: A compilation of biases, failed heuristics, and cognitive dispositions to respond........ 232

Appendix B: The 26 items of literature included in chapter 4’s literature review 237

List of Tables

Table 1: Key models of clinical decision-making ... 29

List of Figures

Figure 1: The red thread running through this thesis ... 26

Figure 2: Extremely simplified AI process flowchart ... 40

Figure 3: The AI process as a black box ... 44

Figure 4: PRISMA diagram ... 55

Figure 5: The (non-exhaustive) field of regulation, legislation, and case law regarding AIS use in clinical decision making as discussed in this thesis ... 79

Figure 6: Speculated negligence claim pathway ... 104

Figure 7: Risk Pooling ... 185

List of Abbreviations used

AI	Artificial Intelligence
AIS	Artificially Intelligent System
CQC	Care Quality Commission
GMC	General Medical Council
HCPC	Health and Care Professions Council
IBM	International Business Machines Corporation
MHRA	Medicines and Healthcare products Regulatory Agency
NHS	National Health Service
NMC	Nursing and Midwifery Council
NICE	National Institute for Clinical Excellence
SDC	Software Developing Company

Chapter 1: Introduction

The aim of this thesis is to explore the ethical and legal allocation of responsibility when artificially intelligent systems (AIS) are used in clinical decision-making. This thesis will argue that the burden of ethical and legal responsibility for the use of AISs should be shared between those who develop and deploy AISs and those who use them. Practical suggestions are offered on how this shared model of responsibility could be achieved.

This introductory chapter will identify the potential future opportunities for the rising use of AISs in healthcare, explain why this thesis's focus was chosen, and supply key definitions, limitations, and assumptions of this thesis. A brief overview of the approach and structure of this thesis, outlining the upcoming chapters, concludes this chapter. In-depth critical discussion of both AI and clinical decision-making are avoided in this chapter as these are located in chapters two and three.

Scene setting

The domain of healthcare could benefit from developments in AI due to its claimed potential to "improve diagnostic accuracy, improve efficiency in provider workflow and clinical operations, facilitate better disease and therapeutic monitoring, and improve procedure accuracy and overall patient outcomes" (Kaul *et al,* 2020, p.807). Advances in computer science have resulted in AISs being developed by software developing companies (SDCs) for deployment in healthcare. For example: medical image analysis in Cambridge (Microsoft, 2021), diagnosis of coronary heart disease in Birmingham (Open Access Government, 2019) and investigations or referrals for suspected cancer in Sutton (NHS England, undated a). The previous United Kingdom's (UK's) Secretary of State for Health and Social Care displayed a very public interest in AI technologies (Hambury, 2018). This enthusiasm has led to notable initiatives such as the creation of the organisation NHSX (NHSX, undated) to guide digital transformation within the service, and multimillion pound investments in the NHS AI Lab with the aim to "improve the health and lives of patients" (Downey, 2019). This activity was further bolstered by an additional funding boost for new projects targeted to diagnosis and care via the AI in Health and Care award (Department of Health and Social Care, 2021a). These activities cumulatively indicate that the UK's healthcare environment is being prepared for AISs to be usefully developed and deployed to aid clinical decision-making.

Whilst funding and development are essential components of AIS implementation, the consequences of the use of this technology must be considered prior to its deployment. As healthcare is something which will be needed at some point in every human's lifetime, it is not outrageous to claim that the widespread adoption of AISs within the clinical environment could potentially affect each individual member of an entire nation in some way. For this reason, the consequences of AIS use in healthcare are everyone's concern.

This thesis is addressed to not just those who develop, deploy, and use AIS in the clinical environment, but all others who may be involved in or affected by its adoption by the National Health Service (NHS). Specific interested stakeholder groups could include (but are not limited to) patients and their significant others, anyone who is involved in commissioning or determining healthcare policy (for example: political and policy leaders, NHS leaders for AI adoption (e.g., NHSX), and dispute resolution (e.g., NHS Resolution), purchasing NHS trusts) as well as SDCs who develop the AISs and end-user clinicians. If this body of work is considered *prior* to the widespread adoption of AISs in clinical decision-making, it may inform stakeholders of possible consequences of such implementation: specifically, the possible harmful consequences of AIS use and the subsequent ethical and legal repercussions which could impact on patients, SDCs and the clinical users.

Clinical care directly affects the health and wellbeing of the individuals who access it; good clinical decision-making may aid a patient's recovery, whereas poor clinical decision-making might risk the prospects of a patient's improvement. Thus, the introduction of AISs in aiding clinical decision-making has the potential to have a high impact on the individuals on which it is used. If an AIS advises the clinician correctly the patient will benefit, but if an AIS advises a clinician wrongly and the clinician fails to spot the error the patient might suffer harm due to the clinician following that wrong advice.

Clinical decision-making has historically been in the hands of clinicians, and SDCs are creating space for themselves as new actors in this field. SDCs aim for clinicians to use the AISs which they have made to aid clinical decision-making; if AISs are adopted by clinicians, a new relationship between clinicians and SDCs results. The clinician and the SDC are inextricably linked when AIS is adopted in clinical decision-making; without the clinician the SDC's AIS cannot reach the patient, while without the SDC there is no AIS to offer the clinician to use in their clinical decision-making.

This is not a close association, though. Whilst AISs are positioned to influence clinicians making decisions about patient care, the SDCs are situated far from the bedside. This distance results in the relationship between the SDC and the clinical user being so subtle that it is nigh invisible. However, such a lack of proximity is not unusual; with other technological applications in healthcare the creators of the technology are not necessarily expected to be present when it is being used. For example, the practical use of a syringe driver is not routinely directly overseen by the organisation which developed and deployed it for use; the manufacturer will not directly advise a clinician when to start or stop a specific medication on a particular patient. This contrasts to SDCs which develop and deploy AISs designed to aid clinical decision-making: whilst SDCs are not present to directly influence how a clinician uses their AIS in the clinical environment, the AISs are designed to *directly* influence the clinician's decision-making for their patient; therefore, the SDC will have indirectly influenced the patient's subsequent care via their AIS.

The distance between the SDC and the patient risks creating a disconnect between the SDC and the effects of their technology on the patient; the clinician is readily available and visible to the patient, whereas the SDC (whose AIS has influenced the clinician) is not. The relationship between the SDC and the clinician may also be invisible to the clinician themselves, due to a lack of direct contact. The clinician might spend an entire career using a piece of technology in healthcare and have no interaction with its developer. However, just because that relationship may be unapparent, does not mean that it does not exist.

Demonstrable examples help to frame the issues addressed in this thesis, as employing a practical example gives context of how using AIS in healthcare decision-making can be problematic. However, this is still an area in development and there is currently no widespread use of AISs in clinical decision-making in England and Wales. Instead, two accounts of one AIS will be outlined which originate from outside of England and Wales. These accounts provide an illustration of how it is possible for an AIS to advise a clinician wrongly and how an SDC can attempt to distance itself from the effects their AIS can have on patient care.

The AIS in both accounts is IBM's Watson for Oncology (Watson). This system was developed in the USA and has been used in several other countries (Ross and Swetlitz 2017). Watson accepts information about the clinician's desired patient, processes that information, and then makes a recommendation for the patient's cancer treatment (Tupasela & Di Nucci, 2020). In principle, this AIS would help clinicians identify the next treatment steps for a patient, however there is still need for caution when deploying and using such AISs. The first problematic account is from Ross and Swetlitz (2017) who have reported an incident in South Korea of Watson inappropriately recommending a drug for an oncology patient which was not indicated. This erroneous AIS output was noted by an experienced oncologist so the patient was not exposed to a risk of harm from unnecessary drug administration, but Ross and Swetlitz (2017) identify that the same AIS is used in another hospital in Mongolia without such specialist oversight. This demonstrates that AISs are not necessarily infallible, and that it is possible for an AIS to make it all the way from development to deployment for use in clinical decision-making whilst being still capable of error. The second problematic account stems from the dissonance between expectations of academic commentators and the statements of the SDC. This is illustrated by a quote from Hengstler *et al* (2016) who interviewed an unnamed IBM executive as part of an investigation regarding applied AI and trust:

"Watson does not make decisions on what a doctor should do. It makes recommendations based on hypothesis and evidence based [sic.]"

Hengstler et al, 2016, p.115

The term 'recommendations' demonstrates how Watson is deliberately positioned to influence the clinical user, whilst IBM simultaneously dissociates itself from how Watson's own outputs are finally utilised in clinical practice. Clinicians are generally held to owe a duty of care (as exemplified in English law in *Barnett v. Chelsea and Kensington Hospital Management Committee* and *Darnley v. Croydon Health Services NHS Trust,* discussed in chapter 5), but the quote captured by Hengstler *et al* (2016) suggests that SDCs have attempted to draw a clear line to avoid owing a comparable duty.

Both of the above accounts draw attention to the possible consequences of the use of AISs. If it is possible for an AIS to make a recommendation which is erroneous, it is possible that a clinician may not detect that error, and will be influenced by, and use, that recommendation - leading to patient harm. If the SDC has positioned their AIS as a 'recommender' and not a 'decision-maker', there is an indication that the SDC intends to deflect both legal and ethical responsibility for the consequences of the use of their AIS to the clinical user. This arrangement may be beneficial to the SDC, but unfavourable to the clinical user.

The arrival of SDCs as novel actors in the clinical decision-making space and the novel relationship between clinicians and SDCs via the use of AISs may alter the allocation of ethical and legal responsibility for harms stemming from clinical decision-making. Over the time it has taken to develop and construct this thesis,[1] a lot of thought, writing, and action has taken place globally concerning the use of AIS in the clinical environment. The words 'responsibility' and 'accountability' are often cited when discussing AIS adoption, but without comprehensively and specifically addressing how either of these notions ought to be considered or addressed. For example, NHSX's 2020 "Buyer's Guide to AI in Health and Care" states that a "culture of ethical responsibility" ought to be built and maintained around an AI project and that "it's important to be clear on who has responsibility should anything go wrong." This advice lacks any depth and specificity, and without extensive and definite guidance the clinical adopter is left to find their own way. To make up for the lack of guidance, multiple clinical organisations (such as NHS trusts) may perform their own ethical and legal analysis regarding the allocation of responsibility to guide their activities; but the duplication of this work by different organisations may result in their differing conclusions. Such irregularities may result in unequal adoption of AIS's in every NHS organisation and thus varying benefits or inconveniences for any

[1] I commenced in March 2017

stakeholder group. As noted by Huxtable (2020) “clarity, consistency and fairness may best be served by authoritative national ethical guidance.” Whilst Huxtable drafted these words in the context of the COVID-19 pandemic, the same is true when considering the adoption of AISs. A lack of robust ethical and legal analysis will hinder the construction of national guidance for the use of AISs in clinical decision-making. This thesis addresses this knowledge gap by raising, and attempting to answer, the following questions:

- How, according to current law in England and Wales, will legal liability be allocated between clinicians and SDCs when AISs are used in clinical decision-making?
- How can ethical responsibility for the consequences of the use of AIS in clinical decision-making be determined and allocated?
- What could be a fair balance of responsibility between the clinician and the SDC when AISs are used in clinical decision-making?
- How could a fair balance be practically achieved?

By answering these questions, the outcomes of this work may serve to inform future thinking which guides stakeholders towards understanding their ethical and legal obligations to others, as well as speculatively demonstrating how responsibility can be shared in an ethical manner.

Intentional limitations of this thesis

The scope of this work has five significant limitations:

1) the discussion will be mainly limited to just two specific key actors, the clinician and the SDC,
2) the geographical area where this thesis is set will be limited to England and Wales,
3) the clinical environment discussed will be restricted to that of the English and Welsh NHS,
4) discussions involving liability are constrained to the tort of negligence,
5) this thesis will not look at analogous fields.

Limitations 2), 3), and 4) can both be easily justified for reasons of space; no thesis could reasonably consider every jurisdiction, every possible application of liability, and every clinical environment worldwide.

The reasoning for limitation 5) is that, whilst the potential applications for AI in other areas are vast, the discussion in sectors other than healthcare predominantly refers to the scenario that the AIS is taking over a role from a human. As an example, the Automated and Electric Vehicles Act 2018 (in the absence of contributory negligence) applies when an autonomous vehicle is driving itself and a human driver is neither controlling nor monitoring its operation:

autonomous vehicle → injured person

If there is no driver for the vehicle, then there is no human whose decision-making is being influenced by an AIS which has resulted in harm to another (for example, a driver permitting a vehicle to collide with a pedestrian). Instead, in this thesis, the clinician occupies the role of 'driver'; this is when the clinician's decision-making is being influenced by the AIS:

AIS → clinician → harmed patient

The AIS outputs information to the clinician, who then chooses whether to use it or not in the patient's care. As such, there is still a human involved in the decision-making. Given that all clinical decisions in healthcare are uniquely made for the patient, there is a level of precision and personalisation that is not present in other sectors. Returning to the example of transportation, applications of AIS on the road take place in structured environments which are rule-orientated and in which actors are compelled to follow the mandatory Highway Code (Department for Transport, 2022).[2] Whilst healthcare is structured, in so far as it is often informed by authoritative guidelines (see chapter 2), a clinician is still needed to individualise that care to the patient rather than it being entirely decided by an AIS. Limitation 1) is quite specific. The following introduces key stakeholders and explains why certain purposeful limitations were made.

Included stakeholder groups

Whilst multiple stakeholder groups will be affected should AISs be introduced to the clinical environment, to explore them all would be impossible in *any* thesis. Thus, the number of stakeholder groups has been limited to a purposefully chosen few; the following introduces the two on which this thesis focuses; the clinician and the SDC.

Clinicians are identified as persons who practice as healthcare professionals and are registered and deemed fit to practice with a professional regulatory body. The non-specific term 'clinician' has been purposefully chosen as the range of specialties in the modern clinical environment is broad and truly multidisciplinary. The wide range of clinical professionals are recognisable by their affiliation with key organisations such as the General Medical Council (GMC), Nursing and Midwifery Council (NMC), or the Health and Care Professions Council (HCPC).[3] Whilst it is tempting to focus on a single clinical professional group, such a medicine, the issues which concern this thesis are transferable to all.

[2] The same can be similarly said for areas such as aviation in which actors manoeuvre in highly structured and rule-mandated environments.

[3] Other clinical professionals also need to be registered with their respective regulator to practice in England and Wales, for example dentists and osteopaths. However, the GMC, NMC, and HCPC are the three regulators of greatest prominence and hold the largest proportion of practitioners, thus focussed upon.

Recognising the variety of clinical professionals does not take away from considerations made in this thesis and serves to increase generalisability.

Clinicians can be accessed by patients using private funding routes, but, in the United Kingdom (UK), the most common route to healthcare is publicly funded and supplied by the NHS, thus NHS care is the chosen setting for this thesis. The term 'clinicians' is used throughout this work as they are the agents who make clinical decisions based on their assessment, planning, implementation, and evaluation of patients in their care. Depending on the context of care, clinicians can perform their roles either alone or as part of a larger team; yet, there is a notable distinction between individual and corporate interests. The existence of the NHS's aforementioned NHSX and NHS AI lab demonstrates a greater organisational interest in welcoming the opportunity of using AIS in clinical decision-making. Whilst individual clinicians may be consulted before an AIS is chosen, it is the NHS organisational bodies rather than individual clinicians who would purchase access to AISs and roll them out into the clinical environment. This means that clinicians may be presented with an AIS to use and not be involved in its choice or procurement. These users could be wary of adopting AISs into their clinical practice as each clinician is held individually professionally responsible for their own actions, or conversely AI may be casually welcomed into the clinical space when more caution may be advisable. Unless otherwise stated, this thesis chiefly considers the individual clinician rather than the views of NHS organisations, as whilst an NHS organisation may provide the clinical environment and tools for healthcare and be regulated by the Care Quality Commission (CQC) to ensure standards of quality and safety (Care Quality Commission, 2017), it is the clinician who remains accountable for their own practice and it is the clinician who makes the final decisions regarding the selection of tools used (e.g., and AIS) to inform the care of individual patients.

SDCs are the organisations which create and deploy AISs. In this thesis the term 'SDCs' shall be dominant as, generally, an organisation would create and deploy an AIS for the clinical setting, rather than a single technologist operating alone. However, similarly to the distinction between clinicians and the organisations which they work for, there will be times where it will be pertinent to consider the actions of a single technologist exclusively to the SDC for which they work. When this is the case, the term 'technologist' will be employed to indicate an individual actor.

Excluded stakeholder groups

Two stakeholder groups have been broadly excluded from this discussion: patients and the AISs themselves. The following explains why.

Patients are those individuals who need clinical assistance and are the subject of clinical decision-making; for this reason, they are the most important stakeholder of all in healthcare debate. When

AIS is used by a clinician, current norms around consent would still be observed, and so the patient role in this will not be substantially different to what it is now. Patients rely upon the wealth of the clinician's expertise to inform them of options, risks, and benefits regarding their healthcare. For as long as the interaction remains between the patient and clinician, their relationship remains the same; it is only the clinician's position that changes, and that change is because the clinician is using an AIS as another source of expert information. When AISs interact directly with patients then this dynamic changes as there is no clinician involvement. This is already the case in two hospital trusts in England where an AI powered tool is used directly by patients to help identify those with potential COVID-19 (Downey, 2020). The dynamic which exists when AIS and patients interact directly, whilst worthy of in-depth study, is outside of the scope of this thesis; this work's focus remains on the decision the clinician makes and what informs those decisions.

The term AIS refers to the computer systems which can receive information, process it, and then deliver an output determined by the information which it has processed. Whilst AISs have been identified as moral agents (Floridi and Sanders, 2004), the introduction of special status for 'electronic persons' has been considered by the Committee on Legal affairs of the European Parliament (2017), and one AIS has even been given honorary Saudi Arabian citizenship (Griffin, 2017), AISs are intentionally not identified as stakeholders in this thesis, as giving them such status is problematic. Machines do not think, therefore cannot carry moral responsibility or liability for the outcomes of their actions (Panel for the Future of Science and Technology, 2020). Bryson *et al* (2017) explain that legal fictions of personhood when considering such electronic persons are morally and legally troublesome. Their most compelling argument describes how these digital artefacts can be used as liability shields which would absorb the legal responsibilities of human actors; that there is no entity to hold responsible for harms caused. If electronic persons were permitted, then the more advanced AISs become, the more difficult it would be to blame a human for the effect of the AISs use. Bryson *et al* (2017) prefer the view that when allocating responsibility for harms done, we look for the real person who ought to answer for the harm, rather than the artificial person presented to manage the liability on their behalf. Everett (2021) also observes that people have commissioned the development of systems, set their parameters, and employed them into practice; he further suggests that we ought to look to these persons rather than the system when things go wrong and blame needs to be allocated. Winfield carried the same sentiment when he wrote that we should ensure that, regarding AISs, "responsibility and accountability remains with their human designers or operators" (Winfield, 2019, p.47). Indeed, the European Group on Ethics in Science and New Technologies (2018) noted that "the ability and willingness to take and attribute moral responsibility is an integral part of the conception of the person on which all our moral, social and legal institutions are based" and concluded that “moral responsibility, in whatever sense, cannot be allocated or shifted to ‘autonomous’ technology” (European Group on Ethics in Science and New Technologies, 2018, p.10). UNESCO (2021) underlined this by stating that, whilst humans may use AIS in their decision-making and acting, member states should ensure that ethical and legal responsibility for AIS use should remain with people or existing legal entities and not be allocated to AISs. For these reasons, AISs will not be considered as stakeholders that can be allocated ethical and legal responsibility for their actions. Instead, the clinician and the SDCs have been specifically identified as the two key actors in this work. It is the SDCs who create and deploy the AISs in question, and the clinician who would eventually use it at the point of care.

As well as stipulating limitations on the actors who will be considered, this thesis also makes a number of assumptions about those actors, which will now be set out.

Assumptions of this thesis

This thesis assumes that clinicians and SDCs are separate stakeholder groups. Whilst it is accepted that clinicians may have some technological knowledge, and that those who work in SDCs may bring clinical knowledge to their roles, when considering scenarios in this thesis, I assume that the actions of these two groups are distinct.

This thesis additionally assumes that the SDC and the clinician aim to do no harm and to actively do good for patients. This is not an unreasonable assumption to make of a clinician, as a central and crucial aim of healthcare is to help people. Additionally, non-maleficence (doing no harm) and beneficence (doing good) are identified as key principles in biomedical ethics in Beauchamp and Childress’s (2013) lauded account of medical ethics. It is also reasonable to assume that the SDC who has taken actions to attempt to improve patient care and outcomes through their work has similar aims. In the same vein, an SDC and clinician may both be financially rewarded for their contributions to clinical care. Payment for work performed is the dominant way by which people find the means to meet their financial needs, and, although money may provide personal motivation, it will not immediately be assumed that financial gain interferes with their presumed goal of patient wellbeing and provision of best possible care.

Finally, this thesis generally assumes that the patient assents to AISs being consulted when their care is being deliberated by their clinical provider. Given that the patient is the subject of decision-making, this is a large assumption to make as it places the opinions of the patient firmly outside of the scope of this thesis; however, consideration of the patient’s perspective regarding the use of AIS in their care is a standalone research project in its own right. Future research regarding stakeholder consultation is explored at the end of this thesis.

Having now identified the key limitations and assumptions in this thesis, I shall now illustrate how the research was conducted and outline the thesis by chapter.

Thesis approach and outline

The following describes the path taken to identify, explore, deliberate, and discuss the challenges around the allocation of ethical and legal responsibilities between the SDC and the technologists it employs and the clinician when AI systems are used in clinical decision-making.

Chapter 2 opens this thesis by identifying how human clinicians make decisions. Humans are fallible actors, and the clinical professions are no exception (Berner and Graber 2008; Makary and Daniel, 2016; Hogan *et al* 2012). The nature of clinical decision-making is complex and challenging to describe (Castelvecchi, 2016; Croskerry, 2002), plagued with uncertainty (Farnan *et al*, 2008), and affected by biases held by the decision-maker themselves (Croskerry, 2002; Stiegler and Tung, 2014). These issues are made more challenging when one considers that clinicians wrangle impractically large volumes of information to stay up to date with the current best evidence in their field (Alper *et al*, 2004; Allen and Harkins, 2005). Knowing this makes it understandable that clinicians then practice defensively; for example, ordering more tests to inform their decision-making, which might result in a contradictory effect on patient wellbeing by increasing the risk to the patient (Pearce *et al,* 2012) and increase the costs of care (Reschovsky and Saiontz-Martinez, 2017), whilst not necessarily changing the outcome (Brito *et al,* 2014).

Of note, due to the specific location of the jurisdiction of this thesis, Ortashi *et al* (2013) surveyed 204 hospital doctors practicing in the United Kingdom. Of those, 78% reported practicing defensive medicine in various ways including ordering unnecessary tests, arranging un-necessary referrals to other specialities, or refusing to treat high-risk patients. More recently, Bourne *et al*'s (2019) larger survey of 5661 obstetricians and gynaecologists practicing in the UK found that 36% met the 'burnout' criteria, which they also found was associated with increased defensive medical practices as well as worse doctor well-being. Due to the high percentage of clinician's reporting defensive medical practice, both Ortashi *et al* (2013) and Bourne *et al* (2019) speculated that the resultant cost to the NHS budget would be very high and that further (as yet not performed) cost analysis studies are needed.

The identification of these issues in human powered clinical decision-making provides the grounding for considering why AISs might be useful to clinicians. AIS's might be able to help by addressing problems such as clinician bias, managing patient data, and helping to handle the ever evolving evidence on which clinical knowledge is based.

Chapter 3 opens by identifying what an AIS is and discusses its potential use in healthcare. AIS use is noted as additionally problematic when the process that it uses to make its recommendation behaves like a black box and cannot be scrutinised, i.e., the AIS is opaque (Fenech et al. 2018). This is challenging as using an emerging technology raises concerns about how it can be a new source of error (Fenech et al. 2018). Opacity of AIS systems is one of four areas of concern identified in the use of AIS in clinical decision-making; the other three are explained as accountability, responsibility, and liability. If the AIS in use is opaque, the user of the AIS cannot provide a full account for their actions when using the AIS. If the user can only account for their actions by stating that they had followed an opaque AIS generated recommendation, can they reasonably deny taking responsibility for the effects of using that recommendation? This question becomes sharply relevant if it is being asked because a patient has been harmed. The clinician may be able to demonstrate that they had used an AIS appropriately, but the SDC may argue that they are not responsible for any harms resultant from that use as they were not at the bedside and that the clinician had made the final decision to use the AIS's recommendation. If AISs are to be used in clinical-decision-making in the future, the allocation of responsibility needs to be determined and stakeholders will need to understand the ethical and legal allocation of responsibility for any consequences prior to the deployment and use of AIS.

Chapter 4 starts to examine these ethical and legal issues by presenting a narrative literature review. This chapter critically explores the literature to uncover concerns about accountability and the allocation of responsibility and legal liability as applied to the clinician and the SDC. It reveals several problems. Accountability is required by the codes of conduct of the three main clinical professional regulators (GMC, 2020; NMC, 2018; HCPC, 2016) whereas SDCs have no such professional regulation or requirement (Whitby, 2015). It is also confirmed that opacity may interfere with one's ability to 'account' for using an AIS in clinical decision-making. These two factors being in conflict might make it professionally impermissible for a clinician to use an opaque AIS in clinical decision-making. However, as there are international examples of such an AIS being used, there is value in considering how ethical and legal responsibility would in principle be allocated in the context of England and Wales. The review finds the literature somewhat unclear about the allocation of ethical responsibility between stakeholders, and a lack of a body of case law is noted (The Government Office for Science, 2016; House of Lords: Select Committee on Artificial Intelligence, 2018). This demonstrates the need for in depth analysis regarding the allocation of legal and ethical responsibility when AI is used in clinical decision-making.

The following research questions were shaped by the gaps identified in the literature review:

- How, according to current law in England and Wales, will legal liability be allocated between clinicians and SDCs when AISs are used in clinical decision-making?
- How can ethical responsibility for the consequences of the use of AIS in clinical decision-making be determined and allocated?

To answer these two questions, separate legal and ethical analyses were conducted, and presented in the following two chapters.

Chapter 5 provides speculative legal analysis, within the context of the law of England and Wales, focussed on the tort of negligence for the scenario of a patient being harmed due to an AIS being used in clinical decision-making. This chapter demonstrates that the SDC owes a duty of care to the patient, but that that duty is independent of the clinician. However, even if it can be shown that an SDC ought to have known that a defect in their AIS could inflict harm, most legal responsibility for AIS use may fall upon the clinician in a tort claim anyway due to issues of causation. This could happen if the clinician had failed to recognise that an AISs recommendation was inappropriate when they ought to have known it could cause harm to the patient. If a clinician chose to use an AIS recommendation (the *novus actus interveniens*) the SDC's actions would be no longer be considered in a negligence claim. This seems to protect the SDC whilst leaving the clinician vulnerable to negligence claims. Given that both the SDC (via the AIS) and the clinician were both involved in the clinical decision, there is a question of fairness regarding the speculated lack of liability for the SDC.

Chapter 6 explores how this legal position could be challenged ethically. This chapter starts by explaining why ethics and morality matter to the conduct of law, explores some relevant theories of ethics, and then critically explores how responsibility can be allocated to clinicians and SDCs and the technologists it employs from an ethical perspective. It claims that, whilst clinicians can be responsible for their use of an AIS due to their ethical duty of care, there is also scope for SDCs to be allocated their own duty of care and carry some of the responsibility for outcomes. A contractarian approach (*à la* Rawls, 2001) is suggested which would allow stakeholders to discuss and plan a social contract which would allow the fair allocation of the practical legal and ethical burdens of responsibility.

Chapter 7 provides a critical discussion of current, potential, and novel solutions which would facilitate the fair sharing of ethical and legal responsibility between these two stakeholders. The main suggestion is that a shared model of responsibility could be fairer than allocating ethical and legal responsibility to individual actors. Various ideas are discussed but, ultimately, mixed prospective and retrospective moral responsibility approaches are advocated and the practical possibilities of pooling risk between stakeholders (risk pooling) and the regulation of AISs is explored. Suggestions for next steps in adopting a shared model of responsibility are offered and future research considerations noted.

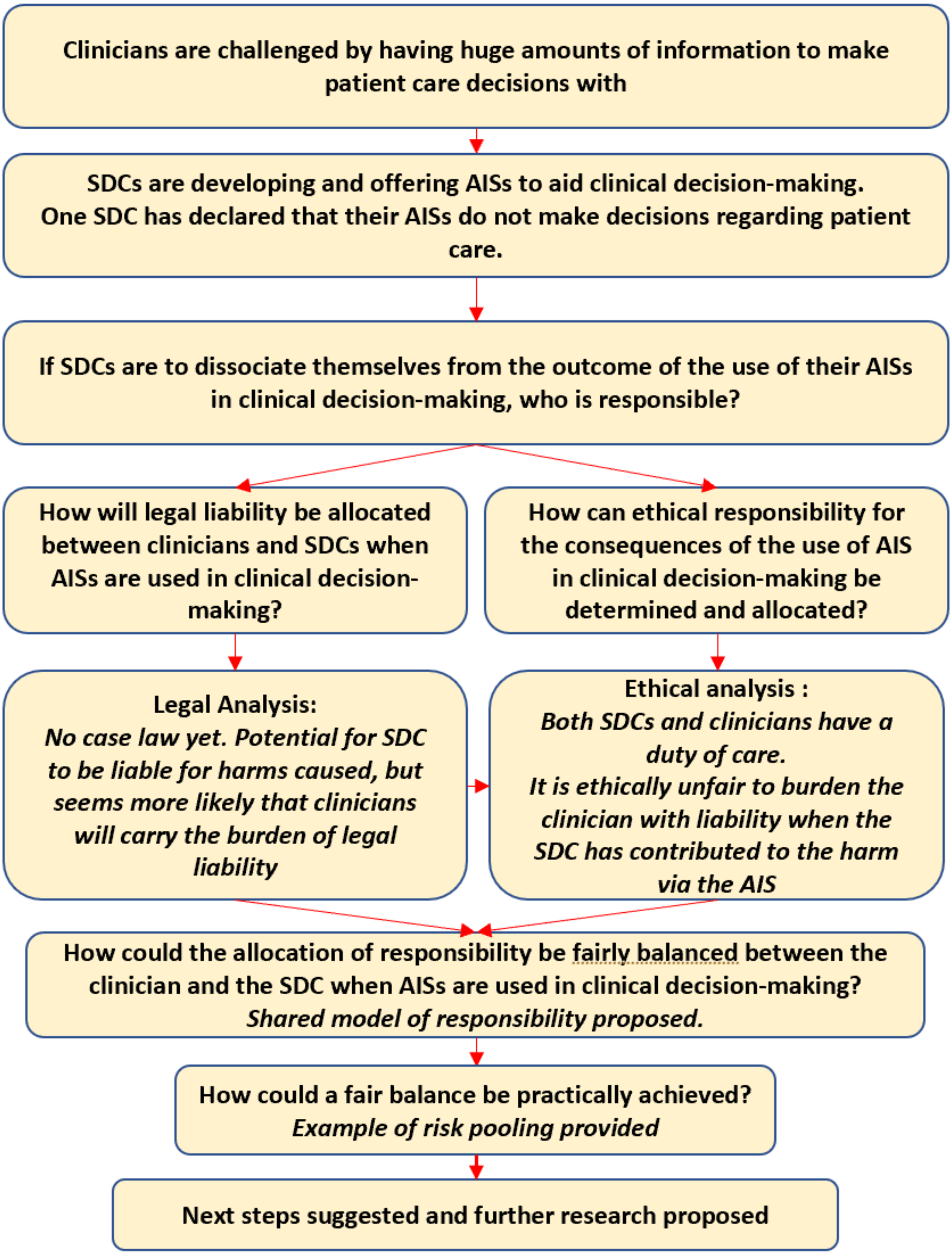
Chapter 8 concludes this thesis by summing up the findings of this work and offering final remarks and avenues for further work.

Conclusion

By means of closure to this introductory chapter, I offer figure 1 below. This shows what Greenhalgh (2019) describes as the ‘red thread’: the flow of problems, questions, and solutions which flows through this thesis. Figure 1 thus allows the reader to orientate themselves with the essence of analysis and discussion which cascades through the chapters.

Having now outlined the reason for this research, the approach and structure of this thesis and an outline of the upcoming chapters, we move to chapter two, where I identify and discuss how clinicians currently make their decisions and why employing AISs might be beneficial to them.

Figure 1: The red thread running through this thesis


Clinicians are challenged by having huge amounts of information to make patient care decisions with
SDCs are developing and offering AISs to aid clinical decision-making. One SDC has declared that their AISs do not make decisions regarding patient care.
If SDCs are to dissociate themselves from the outcome of the use of their AISs in clinical decision-making, who is responsible?
How will legal liability be allocated between clinicians and SDCs when AISs are used in clinical decision-making?
How can ethical responsibility for the consequences of the use of AIS in clinical decision-making be determined and allocated?
Legal Analysis: *No case law yet. Potential for SDC to be liable for harms caused, but seems more likely that clinicians will carry the burden of legal liability*
Ethical analysis : *Both SDCs and clinicians have a duty of care. It is ethically unfair to burden the clinician with liability when the SDC has contributed to the harm via the AIS*
How could the allocation of responsibility be fairly balanced between the clinician and the SDC when AISs are used in clinical decision-making? *Shared model of responsibility proposed.*
How could a fair balance be practically achieved? *Example of risk pooling provided*
Next steps suggested and further research proposed

Chapter 2: How do clinicians currently make decisions?

Professional clinical practice is the main port of call for those who find themselves unwell in the UK. To date, clinical decisions have been made by the clinicians. Unfortunately, human decision-making is flawed, sometimes fatally so; occasionally patients are made worse at the hands of those charged to treat them and those incidences can be fatal. In the UK one in twenty hospital deaths in 2009 had a 50% or higher chance of being preventable, and most of these deaths were due to poor clinical monitoring, diagnostic errors or inadequate drug or fluid management (Hogan et al 2012).

Avoidable adverse drug events from medication processes (prescribing, dispensing, administration, and monitoring of medicines) occur at an estimated rate of 237 million medication errors in the UK annually (Elliott *et al*, 2021). The majority (72.1%) of those errors were reported to have potential to cause minor harm only, but there were a minority (2%) of errors which had the potential to cause severe harm to patients (Elliott *et al*, 2021). The calculated associated financial cost to the NHS amounted to nearly £100 million annually (Elliott *et al*, 2021).

A clinician's recognition of their own flawed decision-making might be unknown to them, and this is exemplified in the final report of the Ockenden review of maternity services at the Shrewsbury and Telford Hospital NHS Trust (Ockenden, 2022). Here it was "found that throughout the review period staff were overly-confident in their ability to manage complex pregnancies and babies diagnosed with fetal abnormalities during pregnancy" (Ockenden, 2022, p.ix). The governance and leadership failing to follow national guidelines combined with delays in escalation and failure to collaborate across disciplines (e.g., not escalating a case to obstetric anaesthetists until the last minute) resulted in poor outcomes for mothers or their babies, including death (Ockenden, 2022).

This is not just an issue in the UK; diagnostic error has been observed at a rate of 5-15% in the USA (Berner and Graber 2008). Makary and Daniel (2016, p.1) have gone so far as to claim that medical error is "the third leading cause of death in the US."

It's not just the patients who are affected; merely being under suspicion of having made an error can have profoundly negative effects upon clinical staff; doctors who have had complaints made against them frequently suffer emotional distress (Bourne et al, 2016). In the UK, the General Medical Council (GMC) has had to recently manage an increase in fitness to practice cases and "said that doctors were only too conscious of the possibility of patients taking legal action against them or complaining to the GMC" (O'Dowd, 2015, p.1). The tremendous effects of such events upon clinicians is illustrated by Horsfall's 2014 report for the GMC which identified 28 cases of UK doctors dying of suicide or suspected suicide whilst under investigation through their fitness to practice procedures from 2005-2013.

As we shall see, technological advances are beginning to offer decision support to clinical areas that adopt it. One hopes that these new endeavours will reduce errors and improve outcomes for both clinicians and their patients and indeed that is a significant, if not the prime, driving force behind their putative adoption. But before we consider how the field of AI can assist in decision-making, we need first to understand how clinicians currently make decisions and how a decision-making process may be weakened due to human factors. Understanding how clinicians make decisions highlights the potential for how well-designed AI systems might help clinicians with their cognitive shortfalls and complement their decision-making. This chapter sets the context of decision-making and outlines the current theories of clinical decision-making.

Models of clinical decision-making

The way that minds manoeuvre to make judgements in clinical practice is "probably that aspect of medical care that we understand the least" (Lighthall and Vazquez-Guillamet, 2015, p.156). The fact that this phenomenon is poorly understood is staggering when one also considers that bedside clinicians have been observed making patient care decisions as frequently as every 30 seconds (Bucknall, 2000) and that "physicians are aware that they are acting and operating within a context of uncertainty, with a high risk of error" (Iannello *et al*, 2015, p.702).

Osler commented that "medicine is a science of uncertainty and an art of probability" (Gupta, 2020). Groopman (2007) elaborates when describing the historically accepted method of diagnosis:

> *"Medical students are taught that the evaluation of a patient should proceed in a discrete, linear way: you first take the patient's history, then perform a physical examination, order tests, and analyze the results. Only after all the data are compiled should you formulate hypotheses about what might be wrong. These hypotheses should be winnowed by assigning statistical probabilities, based on existing databases, to each symptom, physical abnormality, and laboratory test; then you calculate the likely diagnosis. This is Bayesian analysis, a method of decision-making favoured by those who construct algorithms and strictly adhere to evidence-based practice. But, in fact, few if any physicians work with this mathematical paradigm. The physical examination begins with the first visual impression in the waiting room, and with the tactile feedback gained by shaking a person's hand. Hypotheses about the diagnosis come to a doctor's mind even before a word of the medical history is spoken."*
>
> *Groopman, 2007, p.11-12*

Such descriptions give vague accounts of a complex thought process; more precise descriptions of how clinical decisions are made lie in the vast field of cognitive psychology. Various models of decision-making exist in the literature, the key ones of which have been distilled and identified in table 1 below.

Table 1: Key models of clinical decision-making

Models of decision-making	**Description**
Bayesian Probability	The probability of the disease is determined each time a question is answered, a new test is done or the response to a therapy is seen; the calculation of probability changes in light of new information and ultimately a clinician should reach a correct decision (Stigeler and Tung, 2014).
Formalised Pattern-matching	Uses cognitive shortcuts rather than statistics by gathering characteristics into recognisable groups (e.g., the symptoms of hypotension, lactic acidosis and tachycardia would indicate shock) (Latifi, 2016). This approach is taught in medical schools, where a mental library can be built which takes advantage of human's ability to identify by pattern matching (Latifi, 2016, Stigeler and Tung, 2014). This method is limited by what is contained in the individual's mental library and its effectiveness can be affected by bias caused by the clinician's previous experiences and a lack of statistical input.
Heuristics	Strategies which have been developed following exposure to previous comparable events. These create quick and efficient decision-making strategies; for example, the selection of the same list of blood tests for a particular group of patients. These rules of thumb may fail when relied upon in the presence of uncommon factors which has not been accounted for by that strategy (Stigeler and Tung, 2014).
Sensemaking	Whereby a decision is reviewed after the event to "better understand the context from which the action resulted" (Stigeler and Tung, 2014, p. 208). Looking back on past events will allow the reviewer to consider alternate factors and actions which could have been taken which may have meant a different outcome; identification of this information affords the potential for improved decision-making in the future. This model is routinely employed in practice by utilising formalised reflection practices; professional bodies use reflection within the revalidation process which

	ensures their members continue to develop their knowledge to benefit their patients (General Medical Council 2013, Nursing and Midwifery Council 2021).
Dual Processing	Information is processed in two ways; System 1 and System 2 takes from all the above strategies. Bate *et al* (2012, p.615) describes System 1 as a "intuitive, automatic, fast, frugal and effortless process, involving the construction of mental maps and patterns, shortcuts and rules of thumb (heuristics), and mindlines (collectively reinforced, internalized tacit guidelines)." Rapid, automatic processes lead to the final product arriving in the consciousness (Evans, 2003) because of exposure to experience, repetition, formalised learning and observation of others in practice (Bate *et al*, 2012). System 2 is composed of "careful, rational analysis and evaluation of the available information. This is effortful and time consuming" (Bate *et al*, 2012, p.615). System 2 is much slower and makes use of working memory (which can be limited by the individual's cognitive ability) and is where deliberative hypothetical thinking occurs (Evans, 2003 and Evans, 2011). It is utilised when one is learning a new skill and can complete a task adequately with "utmost attention and concentration. The individual is still in System 2 but they are now consciously competent" (Bate *et al*, 2012, p.617). With repetition, the task may be able to move to System 1.

Due to this complexity, human decision makers have been compared to black boxes; patients have never accurately known the contents of their clinician's minds or how they calculate their care decisions (Castelvecchi, 2016; Croskerry, 2016). Whilst, however, a rationale for every single decision made is not always demanded, it is expected that clinicians are be able to justify their actions/outputs in a fathomable manner with a rational evidence base when required.

Croskerry (2002, p.1188) echoes Groopman's (2007) observation in practice that "physicians tend not to be formal Bayesians and instead make judgements based on how well the patient's presentation matches their mental prototype for a particular diagnosis" and that instead they have "developed several decision-making strategies that are part of an informal Bayesian approach" (Croskerry, 2002, p.1185) A clinician could call upon their past experience and use a System 1 heuristic for the issues that they recognise and treat daily; other issues of which they're less sure of would lead them to employing System 2 to check their rationale prior to making and applying their decision. Sense making would aid them in recognising issues that they have seen before; reflection on a prior event would give them the chance to make a better outcome the next time around. There are some hard rules (in the shape of national/local/professional guidance) which govern how clinicians operate and provide packaged Bayesian probability and formalised pattern matching, but there is also the clinician's dual processing which draws on their constellation of education and experience personal to the clinician in question.

It is this spectrum of experience and knowledge which creates some differences in opinion regarding a patient's care. A clinician's approach to a patient's problem and their deduction of a solution may be affected by many issues which will prevent uniform judgements being made by a group of peers presented with the same case. The following examines some of the influences which can affect individual clinicians.

Some problems with human decision-making

Uncertainty

Even with the best planning and organisation of healthcare, the clinician will always have to battle with a stochastic element in their daily working lives. A decision-making environment may be obstructed by any number of issues which impede the gathering of information leading to an incomplete picture for the clinician to analyse. Problems such as investigation results not being ready, patients being unable to fully communicate their symptoms, lack of personal knowledge or access to experienced senior colleagues, and distractions such as irrelevant symptoms all can serve to obfuscate and potentially misdirect clinicians. If the clinician fails to compensate for factors which cause uncertainty there is a risk of misdiagnosis; misdiagnosis may result in a missed opportunity for a patient's effective treatment or the application of an inappropriate clinical intervention (Farnan *et al*, 2008).

Bias

Quality decision-making may be intentionally or inadvertently swayed, regardless of the strategy used, due to bias. Bias is simply "that someone has an inclination to respond in a particular fashion" (Croskerry, 2002, p.1201). It can be either a conscious "systematic preference to exclude certain perspectives on decision possibilities" or the "subconscious influences from life experiences and individual preferences" (Stiegler and Tung, 2014, p. 209). Multiple issues can interfere with good human decision-making even when one is in possession of all the relevant facts. Croskerry (2002, p.1184) calls these issues "cognitive dispositions to respond" (CDRs). These CDRs cause clinicians to make preventable and costly errors (Croskerry, 2002), and therefore one would expect a conscientious clinician to be keen to eliminate any negatively occurring bias which they can identify. It would be challenging for any human to make a decision completely free from undesirable interference, no matter how pure their conscious intentions are. Appendix A contains a collection of biases compiled from a number of authors so that the reader can appreciate how challenging it would be for any human to make a completely unbiased decision.

Just one type of bias will be discussed here for illustrative purposes: implicit bias. Implicit bias was chosen as it is difficult for an individual to identify, recognise and overcome. Staats (2014) defines implicit bias as unconscious stereotypes or attitudes that affect our perceptions, decisions, and actions. A range of negative outcomes can result due to implicit bias and this phenomenon has been identified in various patient groups in a multitude of ways. Chapman *et al* (2013) explain that in the clinical setting this is not helped by information being presented by grouping patients according to their characteristics. This results in clinical training reinforcing stereotypes as it makes decision-making more efficient. Examples of the consequences of this bias include:

- In the UK it was found that males have their pain estimated as being more severe than females, as it is presumed that females have "lower tolerance and greater inclination to express, even to exaggerate, their pain." These presumptions lead to an inequality in prescribing practices whereby females are less likely to be prescribed opiates than males. (Schäfer *et al*, 2016, p.1623)
- Females admitted to hospitals in the USA following cardiac arrests are less likely to undergo therapeutic procedures such as coronary angiography than males. This resulted in women having higher in-hospital mortality than men. Many factors contributed to this discrepancy, but one aspect identified appeared to indicate implicit bias. Women may present with symptoms before their arrest which are atypical to male presentations, such as shortness of breath and experiencing less chest pain. The cause of the subsequent arrest is then not identified as cardiac in origin, which results in less treatments such as coronary angiography being used (Kim *et al*, 2016).
- Physicians in the USA were less willing to treat elderly patients identified with depression and suicidal ideation than a younger employed person as they felt depression and suicidal ideation was rational and normal in elderly patients. This attitude inhibited willingness to use therapeutic strategies due to lack of belief in treatment effectiveness (Ucapher and Areán, 2000).

To combat explicitly negative bias, awareness of its existence should be highlighted and systems created which prevent bias from entering decision-making at all, e.g., the Equality Act 2010 (Government Equalities Office and Equality and Human Rights Commission, 2015). The pessimistic view remains, that there will ever be the presence of prejudice if persons are implicitly biased anyway, despite every effort being made to make changes to the contrary (Staats, 2014).

Croskerry (2002, p.1201) claims that "virtually every cognitive error is judged preventable in hindsight." If implicit bias is difficult for an individual to detect and thus difficult to personally challenge, it might be understandable why an unintentionally incorrect decision could be made by a clinician, even if they had every intention to provide the best care for their patient. They may have felt that, on the balance of the probabilities, and in view of the patient's condition at that time, it was the right decision.

Too much information

Evidence based medicine (EBM) is a movement which has gained momentum since the 1990's. This paradigm shift stressed the "examination of evidence from clinical research" rather than a clinician's decision-making relying on their intuition, experience and reasoning based on their pathophysiological knowledge (Evidence-Based Medicine Working Group, 1992, p.2420). Believing that an action was the best thing to do based only on one's own clinical experience was no longer enough. It is now expected that a clinician identifies the gaps in their knowledge and addresses them by applying critically reviewed, relevant, and adequately powered items of clinical research.

The industry of scientific medical journal publication has grown explosively to feed the clinical professions' EBM appetite for up-to-date evidence to support practices, and access to this body of knowledge aids clinicians to gain, maintain (or change) and improve their evidence-based practices (Garba *et al* 2010). It has been shown, though, that the plethora of published material is far too big for the average clinician to digest. Alper *et al* (2004) calculated that primary care physicians would require 627.5 hours to review the average 7287 relevant articles which would be published in their field each month. With this volume of information jostling to be read, critiqued, and waiting to be accepted or rejected by the healthcare community, it's no surprise that it can take 17 years for a piece of research to be applied to bedside practice (Morris *et al,* 2011).

This enormous cognitive load needs to be simplified so that both patients and clinicians can benefit from the knowledge acquired from research. As a response, EBM has grown from being an endeavour of personal development to a group activity. Members of groups which represent the key specialisms work together to distil the information generated in literature and create guidelines on various matters, which are monitored and reviewed to ensure that they remain up to date. Resuscitation is a strong example: guidelines related to conditions such as cardiac arrest are represented as a simplified algorithm and are reviewed, updated, and republished in light of the evolving evidence base every five years. Each review ensures that practices are in line with current best evidence (Resuscitation Council UK, 2021), but reviews are not so frequent that the clinicians struggle to keep up with changes. This information is presented as an algorithm in the form of a simple flow chart. The Resuscitation Council

UK sets the clinical guidelines for the treatment of conditions such as cardiac arrest in the UK, thus standardising the care of those patients affected. This top-down method of review and implementation of evidence ensures that common issues are managed with consistency throughout the health service. Unfortunately, these organisations can suffer failings which affect the quality of the guidance which they dispense making it impossible for a clinician to adopt their recommendation without the additional labour of critiquing the organisation's conclusions themselves. One such example involves the Cochrane Library. This library "is a collection of databases that contain different types of high-quality, independent evidence to inform healthcare decision-making" (Cochrane, 2020). Roberts *et al* (2015) described incidences of Cochrane delivered reviews which were based on trials which may not have even taken place. Knowledge of this oversight would have dramatically reduced the previously highly held confidence that a reader would have had in any reviews published by Cochrane. Roberts *et al* (2015) addressed this by calling for unregistered trials to be excluded from systematic reviews so that substandard research is not used and that the profile of high-quality research is raised.

Even once the low-powered/poor evidence has been weeded out and the best evidence has been collated and presented in manageable portions, there is still too much information out there. Allen and Harkins's (2005, p.1768) audit demonstrated that in one (quiet) 24-hour admission period they'd admitted 18 patients with 44 diagnoses which equated to 3679 pages of guidelines (122 hours of reading) which the on-call physician would have had to have "read, remembered and applied correctly." This audit had been limited to guidelines generated by the most key organisations only, such as the Royal Colleges and the National Institute for Health and Care Excellence (NICE), but some conditions have had guidelines produced by multiple organisations. Institutions' valiant quests to create good orderly direction out of the overabundance of EBM can easily become cognitively overwhelming for the clinicians they are trying to help.

This problem has been helpfully defined. Bounded rationality was identified by Herbert Simon in terms of there being "so much potentially relevant information available to a decision maker that it is impossible for the human brain to know or process it all" (Bate *et al*. 2012, p.614). Within a practical context, clinicians are constrained by that which they know about a condition, the amount of time that they can spend with each patient to gather information from them as well as the quality of information communicated by the patient, and the time taken to search the literature for unfamiliar issues. Due to the volume of EBM information being completely unmanageable (Bate *et al*, 2012) it is unrealistic to expect clinicians to freshly research the body of knowledge with each episode of patient interaction, and so they work within the constraints of bounded rationality. This leads to clinicians potentially making decisions which are merely satisfactory rather than optimal. Due to this, despite the generation of official guidance, it has been found that physicians tend to use "simple and robust heuristics rather than relying on a powerful memory to remember diagnostic lists" (Ferrira *et al*, 2010, p.5). Bounded rationality is also managed within healthcare by the creation of specialisms; a team of clinicians which has repeated exposure to one grouping of ailments (either by organ system or by symptoms) will practice mostly or exclusively in that field rather than maintaining knowledge and competence about every disease process which could affect the human body. There is also the specialism of being a generalist too; for example, General Practice within the Primary Care setting whereby a physician can treat ailments within their sphere of competence or refer them to a specialist team if required.

Applying the problem of too much information to one of the key models of clinical decision-making demonstrates how limited human decision-making can be, even with the benefit of the sum wealth of EBM. A clinician faced with a patient's problem will either swiftly recognise a diagnosis in the pattern of signs and symptoms (System 1 processing) or will take longer to deliberately think through what they have observed before reaching their conclusion (System 2) (Bate *et al*, 2012). Rapid decisions concerning life and death can be served well by System 1 (for example, recognising cardiac arrest and commencing immediate lifesaving measures within a critical care environment), whereas System 2 decisions need more time to occur (for example identifying a very rare lesion in a dermatology environment) (Bate *et al*, 2012). System 1 is dependent upon the evidence base which may have changed since the last time the practitioner had had to make that particular decision; for this reason, it is important for a clinician to perform a System 2 check of their System 1 knowledge at regular intervals so as to ensure that their practice remains evidence based, relevant and up to date (Bate *et al*, 2012). Without these System 2 checks, errors may creep in resulting in one becoming "unconsciously incompetent" (Bate *et al*, 2012, p.617). But, when the clinical professions are overwhelmed by too much EBM, System 2 checks can be understandably hindered despite the best efforts and intentions of the individual practitioner.

Too much medicine

One strategy clinicians use to guard against making the wrong decisions is to reduce the risk of error as much as possible. By gathering as much information as they can (such as ordering tests, reviewing their results and examining current literature) their decisions are stronger, and they can exclude other possibilities. This approach is known as differential diagnosis (Richardson *et al*, 2000) and it can be overzealously grasped. Defensive practice is a recognised issue in the USA where its litigious culture has encouraged physicians to hedge their behaviour and order tests or consultations from colleagues to avoid being accused of malpractice rather than to directly benefit the patient; this trend has been calculated to increase Medicare costs by over 20% (Reschovsky and Saiontz-Martinez, 2017).

The performance of many types of diagnostic procedures can also lead to harm being caused. To illustrate this point, radiological investigations are incredibly common but do come with their own risks; for example, a CT scan on a child may help the clinician diagnose a problem, but the same scan can almost triple their risk of developing leukaemia and brain cancer (Pearce *et al* 2012). In many instances such an investigation would be thoroughly warranted, but the clinician could also be asking themselves how a test would help or change an outcome (Lenzer, 2016). Brito *et al* (2014) identified that, due to the advent of ultrasonography in endocrinology, there has been a threefold increase in thyroid cancer detection and treatment, but the death rate for this low-risk cancer has not changed as a result. They note that all this extra clinical activity will have created patient harms such as risk of complications from treatment, psychosocial and financial pressures for issues which would have been otherwise benign if left undetected. Being able to strike the balance between too much and too little medicine is one which clinicians shall possibly never cease to battle with. This balance is likely to be made ever harder to strike whenever new technology arrives with the fanfare of making patient care better than ever before.

Individuality

Individuality can create discrepancies in the uniformity of clinical decision-making. The factors which could cause such divergence of practice could be cultural (where the clinician trained or has worked), a personal value set (developed from how they were raised and through their own experiences), an interpersonal relationship style (for example due to the clinician taking into consideration and applying the patient's wishes and values to their decision-making), or due to the clinician's own internal status (have they taken care of themselves as well as caring for others?). The clinician may not be as skilled a communicator as his peer; they might not present a well-chosen diagnosis and treatment to their patient as well as another would: this may affect the uptake of treatments and, subsequently, the possible outcomes of that treatment.

Outside of medicine, individuality has been tackled through an attempt to standardise communication techniques. Ayres (2008) reports that American schools who have adopted the Direct Instruction (DI) technique have the teachers follow a pre-decided script which eliminates teacher discretion on how the curriculum is delivered in the classroom. This top-down organised standardised teaching communication method "outperformed traditional education programs in both reading and math" (Ayres, 2008, p.162). DI obviously cannot be applied to individualised patient care cases, but should technology become available which standardises clinical approaches, there might be scope for a similar improvement in our patient outcomes and clinician wellbeing. But, with such an approach, there is a concern that the element of the humanity of the interaction itself may be lost.

How might AI be employed to help address these issues in clinical decision-making?

Uncertainty can leave clinicians unable to make a safe decision. Bias can sway their decisions in an objectionable manner and the plethora of individual traits can prevent clinicians from executing their care either optimally or in a manner that the patient will accept. EBM was supposed to provide a foundation for improved decision-making but ended up creating the new problem of increasing quantity and variable quality of evidence. Clinicians can maladapt to their shortfalls by practicing too much medicine which brings further risk, discomfort, and expense to the patient.

Artificially intelligent systems (AIS) might be able to assist clinicians in overcoming these issues and support them by negotiating the ever-evolving EBM knowledgebase, helping manage the data collected from patients, structure a diagnosis, and select treatments. As we shall see in the next chapter, bias may not necessarily be eliminated purely on the virtue that the cognitive calculating entity is a computer system, but with increased awareness of this and awareness within the clinicians themselves, bias may be tackled so that everyone can gain equal access to treatments which are proven to optimally treat their condition.

Standardisation of care approaches and optimisation of the delivery of the body of knowledge could possibly be managed by an AIS which could act to supplement the clinician's System 1 thinking. This AIS could cushion the burden of information which needs to be processed to create an excellent System 2 decision by managing and presenting the most relevant evidence base. Rather than presenting information which was ill-defined, an AIS could be designed to guide the clinician to that which was the most up to date and immediately relevant to the patient's condition. With an AIS, there is no System 1 or 2, simply the calculation made with the data which is available at that time. The system would deliver its output freshly based on the information which has been presented to it and update that output each time new evidence was made available.

Advances in computing technology are offering solutions whereby some of this body of knowledge might be presented to clinicians in a manageable format, and it is more than reasonable that clinicians might wish to adopt these solutions. Computerised hazard alerts are already known to reduce medication errors and increase safer prescribing (Schedlbauer *et al*, 2009); this type of intervention helps to correct faulty System 1 decisions where a clinician failed to do a System 2 check, especially on drugs which they may feel they prescribe frequently enough not to need to check. But, in time, one may wonder if the clinician would depend entirely on the computer system to check and calculate doses rather than to continually clarify one's own knowledge.

The use of AISs creates a new decision-making dynamic. Clinical decisions have historically been made by clinicians and, as sole actors, they have taken responsibility for their own actions as laid out by their professional codes of conduct (GMC 2020; NMC 2018; HCPC 2016). If SDCs are developing AISs to directly influence the clinician's thinking, they are siting themselves as new third party in the clinical decision-making space. The SDC's involvement in clinical decision-making raises questions of the allocation of ethical and legal responsibility should something go wrong. For example, if a recommendation were issued by the AIS which was inappropriate for the patient *and* the clinician followed that recommendation *and* harm resulted to a patient due to the following of that recommendation, who would be responsible for that harm? The SDC who provided the AIS for clinical use? Or the clinician who used the AIS? This is important to ask as, without some idea of the answer, stakeholders may unexpectedly find themselves responsible for patient outcomes, both positive and negative. If harm is a potential outcome from the use of AIS in clinical decision-making, it is reasonable to consider routes to how that harm can be rectified. Responsibility can be considered both ethically and legally and this thesis directly explores these ethical and legal concerns.

Conclusion

Human clinical decision-making is imperfect and its flaws multifactorial. Attempts to improve decision-making using EBM has created an overabundance of information which outdates quickly and can be unmanageable for clinicians on a day-to-day basis. Strategies are used to help clinicians to manage this, but there could be benefits to both patients and clinicians if computerised systems would be able to accept and process patient data, manage the evidence base and present unbiased solutions for clinicians and patients. However, if AISs are deployed to aid clinical decision-making, there are questions regarding who would be responsible for the results of the use of AIS recommendations, and how harms to patients as a result of AIS use could be rectified.

Whilst the ethical and legal dimensions are key to this thesis, a little more scene setting is required. Having now explained why clinicians might benefit from computerised aid, the next chapter shall outline what AI is and how AISs could be employed to assist clinicians in their decision-making. This chapter shall also begin to address the pertinent issues which make AI a problem as well as a solution.

Chapter 3: What is AI?

This chapter shall provide a non-technical overview of artificial intelligence (AI) and how it can be applied to clinical decision-making. The definition of AI shall be discussed, along with how artificially intelligent systems (AISs) may be problematic. The issues considered focus on user accountability and responsibility, and the potential for stakeholders to be liable for harms caused when an AIS is used. The use of AISs which are opaque (i.e., designed in such a way that their internal processes are not understood) is noted as especially challenging.

I shall introduce common key terms utilising plain language to clarify terminology relevant to this thesis for the non-computer scientist reader. This thesis will not add to the body of knowledge in any technical capacity but instead shall briefly outline what an AIS is. AISs currently in existence are mentioned within the thesis text only to exemplify relevant concepts; there shall be no attempt to describe the current technical state of computer science in depth in this area. This is intentional so that discussion concentrates on the issue of the use of AISs which have been provided by the SDC to the clinical user. This approach also prevents discussion of specific AIS applications which may become quickly outdated; thus, prolonging the relevance and applicability of this thesis's contribution to the body of knowledge. From the concepts identified and discussion set out in this chapter, one may envision how computerised systems which are powered by AI could be used in healthcare, thus setting the scene for the rest of this thesis.

Defining and Describing AI

Whilst one would expect that this thesis would be able to offer a clear definition of what artificial intelligence is, disappointingly the term 'artificial intelligence' is not widely defined (House of Lords: Select Committee on Artificial Intelligence 2018). There remains a sustained lack of consensus in the definition of AI (Monett et al, 2020) and no universally accepted definition (Rosa, 2020).

Rather than attempting to define AI, Bryson's (2020) suggestion is adopted: to use simple definitions of 'intelligent' as a starting point and building a description from there. The definition of intelligence provided by Bryson and Winfield (2017, p.117) introduces the concept as "the capacity to perceive contexts for action, the capacity to act, and the capacity to associate contexts to actions." As life-forms such as plants can be described as intelligent using this definition, they add that intelligence is conventionally recognised as being cognitive, thus "being able to learn new contexts and actions, and the associations between them" (Bryson and Winfield, 2017, p.117). AI artifacts (usually in digital form) demonstrate such cognitive capacities (Bryson and Winfield, 2017).

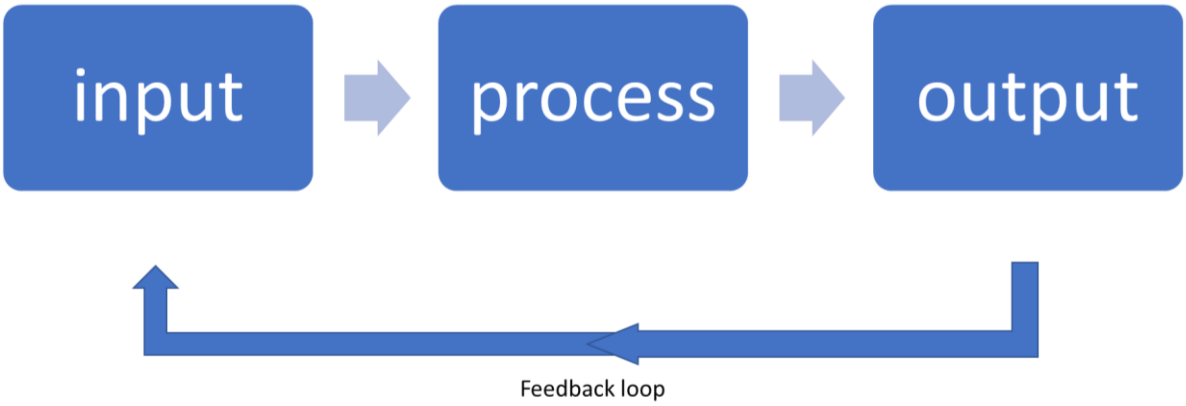
Russell and Norvig (2016) note how the desired properties of AI we wish to harness mirrors the human intelligence we already have. Russell and Norvig (2016, p.1) tell us that human intelligence can "perceive, understand, predict and manipulate a world far larger and more complicated than itself." When applying the concept of intelligence to artificial actors, they describe the noun for a singular artificially intelligent entity as a 'rational agent':

> *"An agent is just something that acts (agent comes from the Latin, agere, to do). Of course, all computer programmes do something, but computer agents are expected to do more; operate autonomously, perceive their environment, persist over a prolonged time period, adapt to change, and create and pursue goals. A rational agent is one that acts so as to achieve the best outcome or, when there is uncertainty, the best expected outcome."*
>
> *Russel and Norvig, 2016, p.4*

The term artificially intelligent system, 'AIS', is used in this thesis over Russell and Norvig's (2016) 'rational agent' to ensure consistent and logical terminology. AIS is akin to the term AI, which is more familiar in common parlance. AI can be used to describe the field of science and engineering which it concerns (Russell and Norvig, 2016), and AI may also be used as a common noun to identify the element of a machine which has that intelligence. However, the use of the term 'AIS' over 'rational agent' does not take away from the desired quality of the AIS in question being able to process the data it receives and provide a useful output. Therefore, in this thesis, the term AIS will be used to discuss the singular agents which will be aiding clinicians with their decision-making.

Rather than an AIS being a technology which rigidly does exactly what it has been constructed and instructed to do (such as a calculator), it can learn and then act on that learning (Bogost, 2017). This artificial learning can result in novel inferences, also known as outputs, which can be designed and positioned to aid human activities.

Figure 2: Extremely simplified AI process flowchart



As shown simply in figure 2, an AIS takes information, processes it, and dispenses an output. The design of an AIS might incorporate a feedback loop which allows it to review the effects of its own previous outputs so it may adjust its future outputs accordingly with other new information inputted. The excitement in the field of AI comes from the potential for an AIS to possess a creative ability to find its own solutions to achieve an outcome; it is this power which challenges people's relationships with the computerised systems in their lives (Government Office for Science, 2016).

Society has previously put much stock in training people to perform specialist tasks and rely on specialists to innovate new approaches to solve the population's problems; but we saw in the previous chapter that humans are far from perfect. The purpose of AI stems from SDCs creating AISs to assimilate representations of our knowledge base to process and improve on it with the aim of helping society to do better. These systems can improve over time and not necessarily with the help of the SDC which created them. Russell and Norvig (2016, p.693) identify benefits of an AIS which can respond and adapt to its observations about the world, thus allowing human cognitive limitations to be overcome. These systems could help us to perform heavy cognitive tasks, such as patient diagnosis, or might use data about our world to make new connections and provide insights previously invisible to human thought, such as identifying new illnesses or suggesting treatment regimes. Depending on how it is designed, an AIS's outputs might eventually have the potential to be more efficient than human thinking and lead to advancements which previous approaches had not achieved. This could potentially benefit society by providing novel cognitive insights which were beyond human capabilities.

As an example, AI development has the potential to develop AISs to assist clinicians in managing data generated by EBM and to cross reference that with what is known about the patient. This could positively affect clinician's delivery of patient care, on both an individual and population level. Once achieved, a step-change defining moment in healthcare AI development will be realised (Watcher, 2015).

However, not all AIS's are the same. They could be designed differently or be working with a different knowledge base to that of others; therefore, different AISs may complete the same tasks differently or reach different conclusions. A change in the information inputted into an AIS, or a change in the process which the AIS subjects that input to, will certainly lead to a change in the output.

AISs with feedback loops have the potential to learn from each task they attempt, thus having the potential to improve the next time it is faced with a similar problem. Here, Russell and Norvig (2016, p.693) explain that learning is desirable because:

1. “designers cannot anticipate all possible situations that the agent might find itself in,
2. designers cannot anticipate all changes over time,
3. sometimes human programmers have no idea how to programme a solution themselves.”

Yet, learning can be unpredictable, and whilst a rule that an AIS has taught itself may be logically correct, it may also be inappropriate and potentially harmful. Exploring the exact natures of any potential flaws that an AIS may acquire (e.g., via inappropriate dataset selection by a human, or by developing algorithmic bias) is far outside of the scope of this thesis, but it might be accepted that, whilst any given AIS might be recognised as imperfect, its outputs might still be useful.[4] However, once the state of the art of AIS development has reached the point where it “is indistinguishable from magic” (Clarke, 1973, p.36), then there is a risk of misplaced user confidence. An AIS might possess an illusion of competence superior to its actual competence, and such an illusion may lead to user trust in the system being misplaced. The next section outlines this problem further.

Opacity

Computer systems with AI properties use algorithms; these are a set of rules which dictate an actor’s behaviour (Weizenbaum, 1976, p.47). It could be argued that using an algorithm in the clinical area is not a novel concept, as instructions which guide actors in clinical practice have long been relied upon. Instructions can be set locally, e.g., standard operating procedures which detail how particular tasks are to be completed within a given institution, or nationally, e.g., instructions which standardise an approach to treatment throughout the nation in which it is applied, such as the last chapter’s description of the UK’s resuscitation guidelines (Resuscitation Council UK, 2021).

As non-computerised algorithms are already in use, it is fair to ask why consideration of the rules which guide an agent’s actions matters. What is the difference between clinicians following rules determined and set by other clinical experts when compared to clinicians following an instruction determined by a computer employing an algorithm housed in an AIS which has been developed by an SDC? The difference is the process by which the algorithm’s instructions have been produced.

A traditional human generated algorithm is created by experts who employ evidence-based medicine to specifically design instructions to be used by clinicians for a target patient group. The algorithm’s authors can explain the reasoning for the instructions in their algorithm, and then manually update the algorithm in light of new information relevant to the algorithm’s area of use. The human generated algorithm’s creators can explain and demonstrate the reasons for their instructions in an understandable format, and these instructions are visibly underpinned by research that can be

[4] Adapted from Box’s (1976) statistical aphorism, ‘all models are wrong, but some are useful’

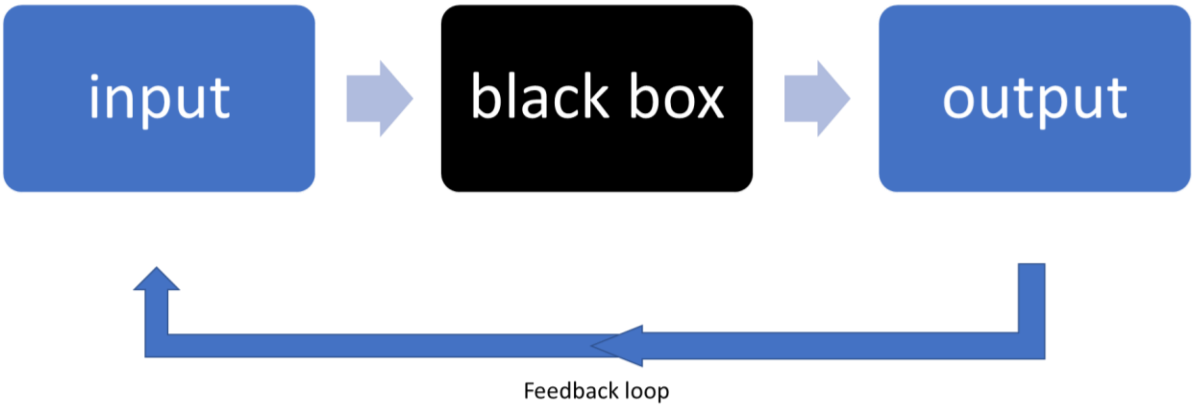
understood by those using the algorithm. Again, the UK's resuscitation guidelines are a good example of this; they are developed and updated periodically by resuscitation specialists (Resuscitation Council UK, 2021). Users recognise these guidelines as the national standard, and that the resuscitation guideline's algorithms are responsibly and expertly developed and safe to use. These actions ensure that the algorithms provide information which, if followed, allow a user to give a patient in cardiac arrest the best chance at a good outcome.

Human generated algorithms are usually generalisable to a patient group sharing similar relevant characteristics rather than to an individual patient. Whilst a broad one-size-fits-all approach may work well in the case of cardiac arrest, it might be less acceptable for patients with other conditions. For example, a human generated algorithm to titrate therapies for a diabetic's blood sugar levels might result in good control for some individuals but be unsuitable for others. At this point, professional clinical knowledge and experience is required to address the gap existing between the algorithm and the patient's problem. Yet, whilst professional clinical knowledge may address such a gap, as noted in chapter 2, human decision-making is flawed. Thus, even experienced clinicians may lack an answer or dispense an incorrect answer to the patient's problem.

The generalisable nature of human generated algorithms may contrast to some AI algorithms which make up AISs. AISs could conceivably take information from the evidence base applicable to the patient's condition as well as specifically employing information about the patient themselves; the AIS's outputs might then be more patient specific rather than generalisable to a population. However, the algorithm required to create an AIS which can deliver such outputs would need to be complex.

Depending on how the AIS is developed, it is possible that the process by which its output is determined is not always known or understood, due to how its algorithms have been designed (as shown in figure 3). The complexity of the design of an AIS could obscure the process by which its outputs are determined. Thus, the reasoning for the AI's output cannot be meaningfully scrutinised. This makes the procedure by which an AI makes its outputs like a black box; the process is 'opaque' (Fenech et al. 2018).

Figure 3: The AI process as a black box



Opacity is a relative concept rather than absolute; for example, a process used by an AIS may be so complex that it is effectively obscured to a non-technically trained clinical user, whilst remaining simple to understand to a technologist who is proficient in that area of computer science. A clinician may be additionally skilled in the design and use of AIS in the clinical environment, but this is currently not a required professional standard. Ideally, a user should be able to clearly see and fully comprehend how one's tool works; that way, the user may ensure that it is functioning correctly, in the way that its creator had intended, and how its user expects.

Early AI projects were able to do this. They included Stanford's MYCIN project (Buchanan and Shortliffe, 1984). MYCIN was able to give expert solutions to complex problems and was described as an expert system. Its creators had surmised that clinicians used rules, empirical associations, and physiological facts to reason when considering illnesses. MYCIN operated by asking the clinician questions and used around 450 rules to diagnose infections (Russel and Norvig, 2016). This AIS was able to show the steps it took to make its output, and thus make visible some form of explanation for the antibiotics which it would recommend (Holzinger *et al*, 2019).

Such transparency in all algorithmic design has not persisted. Opacity in AISs has resulted from increasing complexity, as developments in this area of computer science have evolved over time. Complex answers are now given to us by AISs, simply because the real world is complex (Castelvecchi, 2016); as a result, projects for clinical application have increased in complexity since MYCIN. IBM's Watson for Oncology is a famed example which utilises algorithms employing a technique called machine learning (Keikes *et al*, 2017). Machine learning is a field of AI where a system learns for itself having been given a large amount of data (Marr, 2016). This technique allows the AIS "to adapt to new circumstances and to detect and extrapolate patterns" (Russell and Norvig, 2016, p.2). As well as using machine learning, Watson is also trained by clinical specialists at New York's Memorial Sloan Kettering

Cancer Center (Keikes *et al*, 2017). Thus, IBM Watson for Oncology's development was guided with clinical specialists who identified the information needed to treat patients with specific characteristics (Ross and Swetlitz, 2017). Such training means that the AIS's learning is supervised, i.e., a human supervisor tells the AIS initially how to act by pairing inputs with outputs, and correcting mistakes (Russell and Norvig, 2016). For machine learning to produce its outputs, it makes associations in the data which has been given to it, but the reasoning for those associations can be unclear and the outputs resulting from those associations can be nonsensical, even in the presence of human supervision (Khan *et al*, 2017). Supervision of an AIS's learning does not prevent the AIS from being opaque; it may accept human generated rules, but if it has the capacity to learn and make its own associations, there is still the potential for the associations it makes to use erroneous reasoning. Given these issues, it is conceivable that even a supervised AIS might be designed to improve clinical accuracy and enhance patient care, yet still possess the risk of creating entirely new AIS generated errors.

An AIS which uses machine learning and feedback loops will give outputs which are subject to change as its learning evolves over time. Each time it learns from new experiences, it will add those to its previous experiences and will adjust its behaviour accordingly. Depending on feedback from previous events it may reach a different conclusion each time it is confronted with similar problems. Because an AIS is learning and changing its outputs over time, this output might be optimal, but might not necessarily be predictable to a human holding the same information. Away from healthcare, this point is exemplified in AIS's using machine learning for board games. Early in a game of Go where DeepMind's AlphaGo programme beat Grandmaster Hui in 2016, AlphaGo made a move which was considered initially to be a mistake (Metz, 2016a). The AIS's position of the stones were described as surprising and not human (Metz, 2016b). As the aim of the game of Go is to defeat one's opponent, this rational but unconventional move helped AlphaGo to achieve its goal. However, even though this highly effective yet unpredictable behaviour may be entertaining within a gaming context, it is undesirable in others, such as application to healthcare.

If an AIS's processes are opaque and the AIS is permitted to learn in the clinical environment, it is conceivable that an AIS could calculate and deliver an innovative output (a diagnosis or treatment strategy for example) for an individual patient. An opaque AIS in this scenario would raise issues of trustworthiness when an unexpected output is produced. A surprising yet correct output might be generated by the AIS; here, similarly to AlphaGo, clinicians might view that treatment recommendation from the AIS as a mistake and chose to disregard it - purely because that output was novel and not an approach that a human would have thought of to do.

If a clinical user is faced with using an opaque AIS under these circumstances, they might not be able to determine the appropriateness of the AIS's outputs to the situation they are facing. It has already been described that a clinician may reject an AIS output as an error, but they might also accept an error as correct. Should an AIS output an inappropriate instruction which the clinician does not recognise as harmful and then utilises (e.g., by giving or withholding of a critical drug), there is a risk of patient harm eventuating. This illustrates AIS use as potentially problematic; if the clinical user does not fully understand the AIS's offering, the clinician may not be able to determine when the system's outputs are safe or inappropriate for a patient. A clinician employing their professional training to identify the most appropriate course of action of a patient is one thing, but a computer system wrongly recommending and convincing the clinician of a different and harmful course of action is another. It is principally the patient who would lose out in these scenarios, either by losing an opportunity to benefit from a AIS calculated novel approach, or by harm eventuating due to the clinician using the AIS's erroneous recommendation.

Yet, even if non-opaque AISs are used, there is potential for misinterpretation of the data provided by the system and subsequently its human users; this misinterpretation could also cause harm. For example, Caruana *et al* (2015) found that their system indicated that the risk of dying from pneumonia was lower in those with asthma than the general population, but this was because the data outputted by the algorithm had arisen from the data fed to it: asthmatics received aggressive treatment in critical care departments, thus their mortality rates were lower due to this care. Had this system's outputs been used without this accompanying contextual knowledge, an incorrect presumption could be made that having asthma lowers one's risk of dying of pneumonia.

In summary, using an emerging technology raises concerns about how it can be a new source of error (Fenech et al. 2018). Mistakes in the high-risk area of healthcare might lead to significant consequences for the patient affected (Harwich and Laycock 2018). This is important to consider as patients encounter clinicians at times in their lives when they are potentially at their most vulnerable (Nursing and Midwifery Council, 2018). Conscientious clinicians are aware of how clinical errors can increase the potential for an increase in morbidity and mortality (Makary and Daniel 2016), and it is reasonable that clinicians should oversee all applications of AIS use. Concern for AISs being a new source of error and the nature of opacity creates novel problems when applied to the clinical environment.

Whilst attempts to address AIS opacity are underway (Castelvecchi, 2016), relief of this problem today will not prevent AISs from becoming even more complicated, thus becoming opaque again in the future. But reducing the opacity of AISs will not solve the problem of understanding outputs and detecting errors; it is the clinical user's interpretation, use, and potential reliance on the AIS's outputs which also creates a potential risk to patients.

What Are the Ethical/Legal/Professional Challenges of Using AIS in Clinical Decision-Making?

Opacity

Due to the trend of increasing technological complexity, there will potentially come a time where an AIS will have generated an output that no human can fathom. The output may be correct, but inscrutable. Mukherjee (2017) outlines the scenario of a clinician having no idea how an opaque AIS's answer is created when they asked it a question and, as identified by Char et al. (2018), those who create the AISs for use are unlikely to be at the patient's bedside with the clinician. As Masnick notes:

> *"The more machine learning "learns" the less possible it is for people to directly understand why it's making those decisions. And while that may be scary to some, it's also how the technology advances"*
>
> *Masnick, 2016*

This is true not only with machine learning, but also with methods such as decision tree systems. Decision tree systems ought not be opaque as their structure can be scrutinised, but they may become ever larger and more complex; and that size and complexity prevents someone from seeing how that system's outputs was calculated. For this reason, if society demanded that AIS's were restricted to using techniques which were interpretable it could choke the innovation and progress which could be made for society's benefit.

It is because of these issues that this thesis has adopted the terminology of opacity rather than transparency. Transparency indicates that if the algorithm which an AIS uses to operate could be shown to a stakeholder, then the stakeholder would be able to understand why the AIS functions as it does. But, even with the algorithmic code in full view, the computations can be veiled through the sheer complexity of the AIS in use. Opacity can fluctuate between each stakeholder group; an algorithm which is opaque to one person will not be so to another. By way of comparison, a car engine can be opaque to many drivers as it's too complicated for them to understand, but many would be able to recognise when it was not working correctly; a vehicular engineer would be able to specifically identify a problem in the vehicle and know how to repair it.

As a hypothetical example, an output from an opaque AIS tells its user that a patient's mammogram image is showing cancerous signs which require significant interventions such as mastectomy so that the patient's life can be saved, but cannot explain to us what these cancerous signs are. It puts both the patient and the clinician in a difficult position if they know that the AIS has historically been accurate, but they do not understand how the AIS has reached its recommendation. If users cannot see how the conclusion was reached, they cannot check that the AIS's algorithm had functioned in the desired manner. Accepting an output generated by an opaque AIS, even if it were entirely accurate, would require a complete step change in clinical practice as it would require the clinician and the patient to accept that an AIS's output in this instance was trustworthy.

This is made additionally difficult if the user is unaware of the information which the AIS is basing its decisions on. For example, if the patient were male and AISs made recommendations using data which was sampled only from mammograms performed on females, the results could be affected by gender bias which may result in AIS recommendations which were inappropriate for a male patient's anatomy.

Whilst AISs can be problematic, there is the potential for it to be beneficial to clinical decision-making. An opaque AIS could potentially improve clinical decisions if it had been designed to remove issues such as human originated bias, improve safety by identifying, highlighting, and therefore hopefully preventing human errors. Maybe every clinician could give better care if all treatments could be recommended by an AIS and were completely evidence based on the best knowledge of the hour. Conversely, if the outputs of the machine agent are not explainable and the clinical user cannot critically analyse those outputs, then there is the risk of a clinician not spotting that an erroneous AIS output had been given, and unwittingly causing harmful consequences to their patients by using incorrect/inappropriate outputs originating from these opaque AIS. If the choice were made to deploy AIS's to aid clinical decision-making whilst accepting that the AIS could dispense erroneous outputs which could lead to harm, the consequences of that choice ought to be carefully considered.

Consequences of AIS Opacity: Accountability, Responsibility, and Liability

Because there has been no reported incidence yet of an AIS being used in clinical decision-making which has resulted in harm, there are no real-world examples to scrutinise in this thesis. Instead, there is space to speculate issues which could arise; speculation and consideration of consequences of potential issues allows for pre-emptive preparation (and maybe evasive action) for problems before they arise.

Consider the following scenario. An opaque AIS was used to aid the decision-making of a clinician. The AIS system's output was inappropriate for the patient, e.g., the wrong drug or drug dosage, but the clinician still used the AIS's output. As it was an opaque AIS, the clinician did not understand the rationale for the AIS's output. Maybe the clinician was working with a patient who had conditions which were not in their area of expertise, but the AI had previously reliably dispensed accurate outputs for that particular patient group, so they felt their trust in using it was well placed. As per Holm *et al*

(2021), whilst clinicians should think carefully about the outputs that they use from AISs and reject that which they detect as flawed for their patients, it is certainly possible for a conscientious clinician to use an inappropriate output.

This scenario is concerned with the two key actors which this thesis is concerned with, the clinician and the SDC. It is the SDC who creates and deploys the AIS in question, and it is the clinician who uses it at the point of patient care. The SDC's contribution is designed to infiltrate the clinical area and to influence the clinician's decision-making; so, although the SDC might not be at the bedside when their system is consulted, their system's contribution is nonetheless the influence for an inappropriate clinical decision which might have otherwise not happened. Should an AIS output be used by the clinician and a patient be harmed as a result, questions relating to both ethics and law may be asked:

- how could anyone account for their decision-making when the outputs of an opaque AIS were used and lead to that harm?
- how, according to current law in England and Wales, will legal liability be allocated between clinicians and SDCs when AISs are used in clinical decision-making?
- how can ethical responsibility for the consequences of the use of AIS in clinical decision-making be determined and allocated?

In considering these questions, we may also consider whether AIS opacity creates a fundamental conflict which prevents a clinical professional from using AI in their clinical decision-making.

Accountability is defined as a person's explanation and justification for their intentions and beliefs about their behaviour (Dignum 2019; Oshana 2004). If a clinician acted on the conclusions given by an opaque AIS, even if its outputs were consistently correct, there would be a loss of accountability as one would be acting on information which had not been determined and judged as appropriate by a clinician themselves. Accountability differs from responsibility. A person's account of their actions is linked to responsible behaviour, which is characterised by the "common norms which govern conduct" (Oshana 2004, p.257). Additionally, personal moral responsibility is the individual's obligation or duty to ensure that something is acted or obtained; this individual's burden is attached to them due to the role that they fill within the context being discussed (Zimmerman, 1992, p.1089). If one cannot rationally account for their behaviour in accordance with the accepted norms, any claim that their actions are responsible could be open to challenge; thus, one of the ways that responsible actors demonstrate carrying the responsibility for their actions is by accounting (being accountable) for their actions. The user of an opaque AIS would not be able to provide a full account for basis of the decision which they had made if their decision was based on the opaque AIS's recommendation. They may be able to explain why they felt comfortable following the AIS's recommendation, e.g., the AIS's recommendation made sense in the context it was given, but if the system was truly opaque they would not reasonably know or be able to justify the course of action which the AIS had recommended. A clinician cannot reasonably account for their actions if, when asked why they prescribed or administered a particular dose of a particular drug, their answer is "the black box made me do it" (Castelvecchi, 2016); such a statement could easily invoke the sarcastic response of "if it asked you to jump off a cliff, would you have done that too?" But if a clinician is unable to account for their actions does that mean that they may avoid responsibility for the negative consequences resulting from those actions?

If a patient is harmed due to a clinician using an AIS, the clinician might consider attempting to deflect ethical and legal responsibility for using the AIS 's outputs towards the SDC who had deliberately and intentionally created and deployed the AIS, as it was their AIS which had influenced their decision. If the AIS is opaque, the SDC's AIS might be unable to account for its outputs at the point of use, but the SDC may attempt to argue that they are not the ones making the final decision to use the system on the patient.[5] An SDC might also attempt to argue that if an AIS was still learning when it was deployed to the clinical area that the AIS has moved beyond its original programming, therefore the SDC cannot account for their system's actions, thus cannot be responsible for its outputs.

Conclusion

Creeping erosion of clinical autonomous decision-making due to application and trusting of opaque technology is a conceivable outcome which will affect accountability and responsibility in clinical practice. Stakeholders will need to understand the ethical and legal allocation of responsibility for the consequences of the use of AIS; if they do not, they risk sleepwalking into accepting the presence and adoption of AIS in the clinical environment without fully considering the implications of its use upon their patients and themselves.

Whilst it is desirable that all AIS's deployed in the clinical environment would be perfectly accurate and that every clinician would utilise them correctly, there is a theoretical risk of harm eventuating to patients from that use. It is reasonable to say that, without clear determination of ethical and legal responsibility and some kind of roadmap for how to respond if harms did eventuate due to the use of AISs, those who would use the AISs (i.e., clinicians) and those who would be affected by AIS use in clinical decision-making (i.e., patients) would be justified in declining, or at the very least heavily questioning, its use in their care.

[5] I mentioned this in the introduction and shall discuss in the literature review how an IBM representative stated exactly this.

As discussion of AI ethics has gained in popularity in recent years, there will be great value in identifying what has been determined about opacity in AISs and the accountability of their use, as well as how ethical and legal responsibility is allocated in the context of using AISs in clinical decision-making in England and Wales. The next chapter starts to address these issues by performing a literature review considering opacity, accountability, responsibility, and liability regarding the use of AISs in clinical decision-making.

Chapter 4: Literature Review

Aims

The aim of this literature review was to explore concerns about the use of opaque AIS in clinical decision-making. The issues of accountability, and the allocation of responsibility and legal liability as applied to the clinician and the SDC are examined. This review employs a narrative review supported by a systematic approach.

The searches for this review were performed in February 2018; much had been published between then and the submission of this thesis in January 2022. To bring this literature review up to date for 2022, an additional section, "Updates to this literature review", was drafted and sited towards the end of this chapter.

Methods

Employing a systematically inspired strategy to select and review the literature aids data capture (Khan et al. 2003). The expectation was to find non-homogenous materials in the literature searches, thus careful selection of the type of review process was needed which would accommodate this.

No single theoretical framework proved ideal; Braun and Clarke (2006) identify this as an issue in the selection of research methodology and recommend that "the theoretical framework and methods match what the researcher wants to know". Thus, I have adopted Strech and Sofaer's (2012) four-step model of systematic reviews and adapted it to incorporate the concept of Braun and Clarke's (2006) use of themes in steps three and four for the purposes of this review, as outlined below. The following outlines what was undertaken in each of Strech and Sofaer's steps.

Step 1: Formulate the review question

The literature review aimed to answer the questions "Is it considered permissible for a clinician to use an opaque AI system in clinical decision-making?" and "What concerns are there about opacity, accountability, responsibility and legal liability when considering the stakeholders of SDCs and clinicians in the creation and use of AI systems in clinical decision-making?".

To aid the selection of items to include in this review, inclusion/exclusion criteria specific to the literature review's aims were used to determine the eligibility of materials to be considered for review. This helped me to identify items with relevant arguments and argument themes whilst checking for flaws, credibility, contribution, relevance, and coherence in each item selected for inclusion to this literature review. Each item selected from search results was formally reviewed to ensure quality. The applied inclusion/exclusion criteria were as follows:

- Included:
    - Items must have content pertaining to ethical and legal issues in applications of AIS in clinical decision-making as it relates to opacity, accountability, responsibility, and liability in healthcare.
    - Items must be published in the past ten years. However, if an item is older and its contents offered a significant contribution of value to this review, it was not excluded purely on its age.
    - Items describing the use of AISs in all areas of clinical practice (i.e., inclusive of all fields of medicine, surgery, paediatrics, adult, mental health etc.).
    - Literature must be presented in English.
    - Items which discuss legal theory must be limited specifically to the context of the law of England and Wales (else this review would have become unwieldy with international comparative examples).
- Excluded:
    - Items that did not meet the inclusion criteria.

The literature search found diverse materials in a multitude of formats such as journal articles, books, opinion pieces, reports, editorials, items of discussion and analysis.

Step 2: Identify all the literature that meets the eligibility criteria

Searches were performed across nine relevant databases (see figure 4 below) in February 2018, using the following search terms:

"Artificial intelligence" AND (liability OR responsibility OR accountability OR transparency OR opacity) AND (ethic OR law) AND (healthcare OR clinical OR medical)*

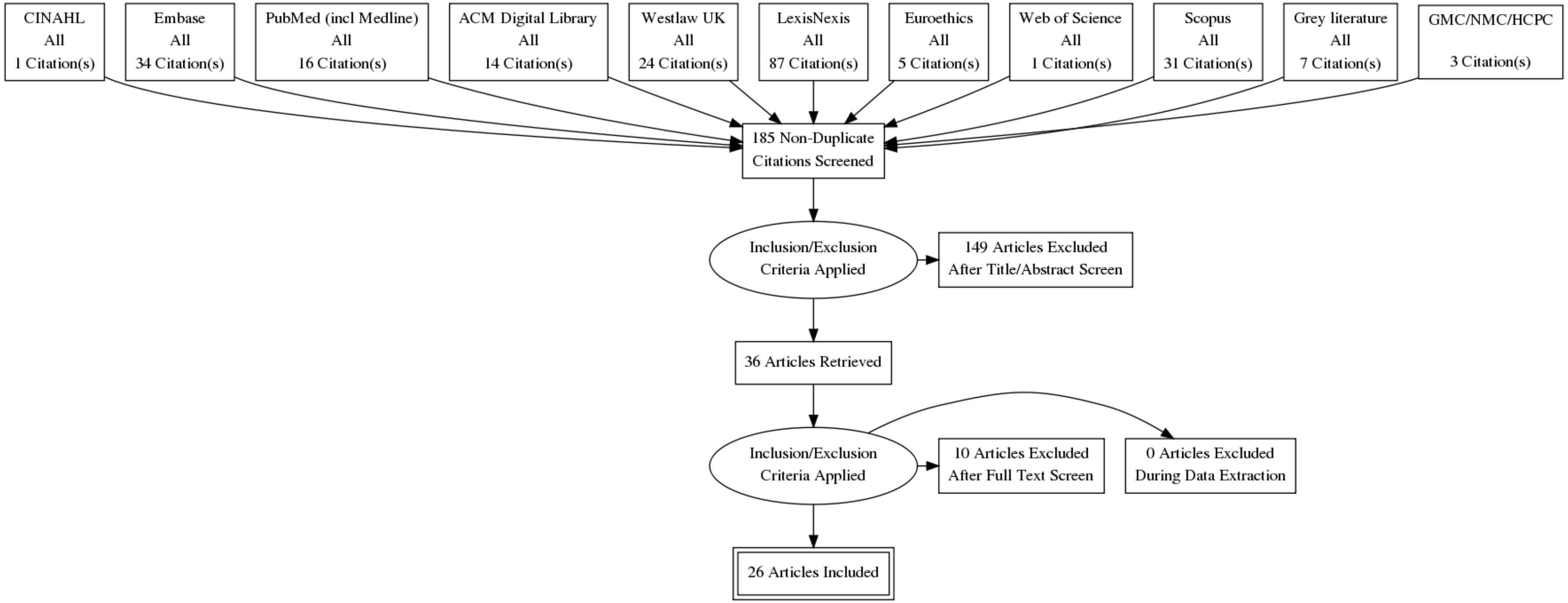
The databases chosen were from a spread of disciplines, not just limited to healthcare, law, and ethics but also to computing and general scientific sources.

Hand searches were additionally performed on the websites of organisations who collectively regulate clinical professionals: The General Medical Council, the Nursing and Midwifery Council, and the Health and Care Professions Council. Relevant grey literature originating from governmental and non-governmental organisations which had been found outside of the searches and had come to the attention of the author during the period of composing the review were included for consideration alongside the formal search results.

This approach generated 185 non-duplicate citations. The author screened each title with the inclusion/exclusion criteria to decide the relevance of each item. Items which passed title screening proceeded to abstract screening and were again subjected to the inclusion/exclusion criteria. In total, 36 items passed title and abstract screening; these 36 items were then subjected to full-text screening by being read fully and the inclusion/exclusion criteria applied again. Nine items were excluded after full-text screening, which left 26 articles.

Figure 4's PRISMA diagram shows the databases used, the number of items identified by each database search and the number of items excluded from the final collection of literature for synthesis and analysis.

Figure 4: PRISMA diagram



The EndNote Online reference management system was used to capture the citations from the searches from each database. From the EndNote hosted catalogue, results were screened, and irrelevant items removed as per the inclusion/exclusion criteria. An independent second review of 10% of the search results was performed by an academic colleague who was external to this project to facilitate robustness and reliability of the selection process (as exemplified by Kyte *et al.,* 2013). The independent reviewer agreed with how the literature had been included or excluded within that sample as per the inclusion/exclusion criteria. This process yielded 26 items of literature that are included in this literature review. Please see Appendix B for the list of 26 items of literature.

Step 3: Extract and synthesise data by the allocation of pertinent points to theme headings

Data extraction was performed on these 26 items using Strech and Sofaer's (2012) method of extraction and coding of data. This technique's strength lies in its promotion of the identification of ethical analysis and argument within an item's content. Relevant arguments and argument themes were identified whilst checking for flaws, credibility, contribution, relevance, and coherence in each item selected for inclusion in this literature review.

The concept of Braun and Clarke's (2006) 'themes' were adopted so that any additional themes found in the literature during the data extraction process could be flexibly considered for addition in the review's findings. Using themes allowed the additional sub-themes of 'why accountability was important' as well as 'why opacity interferes with accountability' to be recognised and explored under the initial core topics of opacity, accountability, responsibility, and liability.

Step 4: Derive and present results organised by themes

The following findings resulted from the above careful searches and selection process. The findings have been structured as per the themes identified in the research question. Corralling the findings into themes enabled stratification of information, thus aiding the analysis and critique of the literature when identifying concerns of AI use in clinical decision-making.

Findings

As per step 4, the 26 items selected for this review were examined for concerns related to the research question's key themes of AIS opacity, accountability, responsibility, and legal liability regarding the clinical use of AIS in decision-making. The following few paragraphs contain the key findings and the high-level literature synthesis generated from that identified in the review.

Regarding accountability, clinicians have a regulatorily enforced professional requirement to be able to account for their actions, whereas SDCs and the technologists which they employ do not; instead, ethical codes of practice are employed in this sector. This comparison raises the question asking if

SDCs/technologists should also be regulated if their AIS is to be deployed in the clinical environment and directly affect patients.

Regarding opacity, clinicians will be challenged by issues of safety and accountability when using AIS's which do not explain their outputs. If a clinician cannot account for the output of the AIS they are using, they cannot fully account for their actions if they choose to use that output. This lack of accountability raises the potential safety issue of using unverified or unvalidated AISs in the clinical environment.

Opacity is not a problem limited only to clinicians; it can also affect SDCs. To recognise this, scenarios which encompass how opacity can affect each stakeholder are detailed in the discussion later in this chapter.

Regarding responsibility, there is a lack of formal clarification regarding who is responsible for the outcomes of AIS use. There is an agreement that one should take responsibility for one's actions when choosing to use an AIS; this includes evaluating the AIS's outputs before using them in the clinical context. SDCs are often considered to be responsible for the accuracy of their systems, but the literature generally pushes back against the idea of SDCs holding any responsibility for the effect that their AIS would have in the clinical environment; often justified in terms of the AIS assisting the clinician rather than replacing the clinician. Shared responsibility is discussed, but there is broad consensus that clinicians will carry the burden of responsibility for AIS use.

The potential for responsibility to be shared in the future is mentioned, and a possible retrospective approach is mooted to determine the allocation of responsibility to shareholders through an analysis of each given incident. The literature agrees that human actors should be responsible for an AIS and that an AIS should not be responsible for itself, however it is permissible for it to carry out tasks if appropriately supervised.

Regarding liability, the literature has not predicted the outcomes of negligence and liability in this area due to no body of case law in this area.

Each of these themes is discussed in depth below, via narrative presentation of key ideas and position found in the literature reviewed.

Why is accountability important?

Professionalism is the vehicle which formalises the notion of trust within organisational structures that gather those with similar skill sets together. By cohorting these skilled persons, standardisation of desirable behaviours can be achieved which serve to promote trust within that professional group (NMC 2018).

Codes of conduct are created by the statutory bodies who oversee their respective healthcare professional groups. In the UK, the General Medical Council (GMC), the Nursing and Midwifery Council (NMC) and the Health and Care Professional Council (HCPC) cover a significant number of practicing clinical professionals; these shall be the three bodies I call upon to exemplify codes of conduct and professional issues.

Accountability from clinicians is required by the GMC (2020),[6] NMC (2018), and HCPC (2016) codes of conduct. GMC (2020) and HCPC (2016) codes of conduct specifically require that the clinician must be able to justify their own decisions, and the NMC (2018) stipulates that a Registered Nurse should be able to fully explain all aspects of a patient's care. The existence and enforcement of these codes result in the clinician's requirement to provide good care with an emphasis on safety. Breach of these codes of conduct would lead to the clinician being exposed to sanctions from their professional regulator: for example, the clinician being prevented from practicing.

Interestingly, despite it being a requirement in professional clinical practice, the literature searches failed to yield a unified definition of accountability and, therefore, the literature is reviewed with the definitions provided in chapter 3 in mind. This encompasses the spirit of that aimed for by the governing bodies; that accountability is when an individual is obliged to explain (account) to those who are entitled to ask (e.g., regulators, a patient) for the decision-making process which guided their actions or omissions.

Hengstler *et al* (2016, p.106) identify trust as "the willingness to be vulnerable to the actions of another person". Given that the patient is already vulnerable due to the nature of their ailment and that a clinician may have to do harm to create the conditions whereby the patient may heal (e.g., a surgical incision whilst under general anaesthesia), trust is, logically, both a relevant and necessary quality which the patient will need if they are to be comfortable to approach a clinician for help and for them to tolerate the treatment pathway under that clinician's care. Armstrong (2018) describes how even when a clinician may be uncertain about their decision-making, the act of communicating and expressing that uncertainty can lead to increased trust from their patient rather than the loss of their confidence. It is reasonable to deduce that if a clinician communicates their uncertainty to their patients and makes clear their thinking process, they are acting in an accountable manner; thus, accountability and patient trust are linked.

[6] The GMC's code of conduct was updated in 2014, and then again in 2019 and 2020. The 2013 version was consulted for the published version of this literature review. The further 2014, 2019, and 2020 revisions also do not consider their registrants' use of AISs.

The historical background of professional cultural carefulness in the clinical professions does not appear to be shared in the field of computer science (Whitby 2015). This was exemplified by Lanfear (evidence to House of Lords: Select Committee on Artificial Intelligence 2018, p.122) who was unable to describe how his artificial intelligence company, Nvidia, was ensuring compliance of their own corporate ethical principles, stating that "as a technologist it is not my core thinking".

Whitby (2015, p.227) notes a lack of compulsory professional standards or formal qualifications for technologists, and that the information technology (IT) industry is "barely regulated"; noting further that whilst medicine is highly regulated "the IT industry is barely regulated at all." Ethical codes of practice do exist for technologists; for example, there is a Code of Ethics and Professional Conduct published by the world's largest computing society, the Association for Computing Machinery (ACM 2018). The ACM code recommends that decision-making "is accountable to and transparent to all stakeholders" and stipulates qualities that technologists should possess, such as avoiding harm and acting honestly. In practice, the ACM does not have the power to enforce rules upon individuals beyond low impact punitive measures, such as termination of membership of the ACM. Termination would only demonstrate disproval from the ACM body, and it would not prevent a technologist from continuing their practice (ethical or not), but it is conceivable that ACM membership termination potentially might affect their access to activities such as future collaboration or funding opportunities. There is an enforced requirement for clinicians to be personally professionally accountable, via their professional codes of conduct, but no similarly enforced requirement for technologists to be personally professionally accountable. Technologists and the SDCs which employ them do not have an obvious direct relationship with patients, but they are designing AISs which aim to contribute to clinical decision-making with the clinician at the patient's bedside. This raises the question of whether there ought to be a requirement for technologists and/or the SDCs which employ them to be regulated in a similar fashion to the clinicians, or, whether regulation of technologists is necessary if they are not directly interacting with the patients.

Having outlined how accountability is an enforced professional requirement for clinicians and not for technologists or the SDCs which employ them, the next section shall explore how the use of an opaque AIS interferes with accountability.

How does opacity interfere with accountability?

The inner workings of computerised systems are not always made visible. 'Opacity' is when the process by which an output from an AIS is made is either too complex to be understood by one, many, or all stakeholders or that the decision-making process has been withheld completely from the stakeholder. Opacity is not the sole term used to describe this problem, for example, when an AIS's decision-making process is obscured, it can be described as a “black box” (Mukherjee 2017), or as not being transparent (Hengstler et al. 2016).[7]

Mukherjee’s (2017) commentary identifies that AISs are being developed in such a way that the process by which an AIS’s outputs are calculated can be opaque, and some of these systems are being designed for use in healthcare contexts with the goal being to help clinicians to improve patient outcomes. The problem here is that a clinician may ask an opaque AIS a question and they may have no idea how the answer outputted to them was created (Mukherjee 2017).

This is additionally complicated by the fact that using AIS outputs that are delivered without verification risks the use of unpredictable or unwanted outputs (House of Commons: Science and Technology Committee 2016).

Hengstler et al (2016) identified that trust is key to ensuring perceived risk reduction and that trust will be reinforced if the trustor is given algorithms that are transparent.[8] Thus, it is reasonable to say that trust will be hard to win from the clinician if they are faced with an opaque AIS to use. As a solution, verification and validation of AISs are recommended by the Association for the Advancement of Artificial Intelligence (House of Commons: Science and Technology Committee 2016, p.16), which states “it is critical that one should be able to prove, test, measure and validate the reliability, performance, safety and ethical compliance—both logically and statistically/probabilistically—of such robotics and artificial intelligence systems before they are deployed.”

Verification and validation might assist the clinician to reasonably account for their actions if they chose to use an AIS. This is important because, as noted earlier, the clinical codes of professional conduct do not permit practice which is not accountable (GMC 2020; NMC 2018; HCPC 2016). There is no mention in the literature reviewed of how a clinical user would know if the verification and validation of an AIS was appropriate, or how sufficient levels of safety from an AIS could be determined.

It has been argued, though, that it is not only AISs which can be opaque; clinicians are also opaque. When interrogated, clinicians are not always able to explain exactly how they may come to a decision for an individual patient because their clinical judgement would be drawing from their experience as well as accepted rules which guide clinical care (Miles 2007). But this does not seem to be considered problematic in the literature reviewed. Thrum (interviewed by Sukel 2017a) exemplifies that if a

[7] As noted in chapter 3, I have identified that opacity and transparency are not the same thing, but I report in this review the terminology as expressed by the author.

[8] See above footnote.

clinician advises their patient that they have a melanoma, the patient does not interrogate the clinician's decision; instead, they accept the biopsy and the subsequent treatment suggested.

Thrum described how patients have traditionally accepted the opacity of medical decision-making and that diagnostic procedures and treatments are usually embraced without interrogating the practitioner's method of determination. One could say that it seems that it is acceptable for people to be opaque, but not the AIS that they are using, but the patient can take advantage of their clinician being professionally bound to be accountable for their practice (GMC 2020; NMC 2018; HCPC 2016); something which neither an AIS nor its creator currently is bound to.

It is difficult to picture the issues generated by the use of opaque AISs in the clinical environment as its use is currently highly limited. The literature did provide some scenarios, both speculative and factual, and these are useful to explore.

Examples of opaque AI scenarios

This review identified three main scenarios in the literature reviewed which illustrated the potential clinical use of opaque AISs and identified opacity as a source for concern regarding accountability of clinical decision-making:

1) The AIS is understandable to one or more stakeholders but not all. Thus, the AIS is not opaque to the SDC who builds it but is opaque to the end user: the clinician (Hartman 1986). The clinician would be experienced in their field, but might argue that they cannot use an opaque AIS as they would not be able to account for the determination of its outputs, and thus be working against their code of professional conduct (as stipulated by the GMC 2020, NMC 2018, and HCPC, 2016). Given that technologists and the SDCs which employ them are not regulated, and arguing that the public would not tolerate clinicians without qualifications to practice, Whitby (2015), p.227 finds it remarkable, "if not downright alarming", that clinicians would base their decision-making on AIS created by "gifted amateurs".

2) Scenario 2 is as per scenario 1, but here the clinician does not hold specialist knowledge of the area which the AIS is advising them on (Ross and Swetlitz 2017). The use of IBM Watson for Oncology in UB Songdo Hospital, Mongolia, is an example of this and was investigated by Ross and Swetlitz (2017). They reported that this AIS is being used to advise generalist doctors who have either little or no training in cancer care. They describe that Watson works by looking at a patient's medical record, choosing what it calculates as the patient's options from a list of treatments, scoring those treatments as a percentage based upon how appropriate they are for the patient, and then presenting these options as recommendations for the clinician to consider. The options are presented to the clinician

as a list ranked ordered by a score from highest to lowest. The Watson AIS is opaque to the clinician as it is unable to explain why it gives treatments their scores (Hogan and Swetlitz video embedded in Ross and Swetlitz 2017). Suggestions from Watson are reportedly followed at UB Songdo Hospital almost at a rate of 100% despite the programme not explaining how its output was generated. Ross and Swetlitz (2017) demonstrate why this is concerning by describing the experience of an oncologist, Dr Kang, using the same Watson AIS in a South Korean hospital.

> *"Sometimes, he will ask Watson for advice on a patient whose cancer has not spread to the lymph nodes, and Watson will recommend a type of chemotherapy drug called taxane. But, he said, that therapy is normally used only if cancer has spread to the lymph nodes. And, to support the recommendation, Watson will show a study demonstrating the effectiveness of the taxane for patients whose cancer did spread to their lymph nodes. Kang is left confused as to why Watson recommended a drug that he does not normally use for patients like the one in front of him. And Watson cannot tell him why."*
>
> *Ross and Swetlitz, 2017*

Watson may arguably be safe in the hands of someone such as Dr Kang who knows the subtle differences in the appropriate use of each of the treatments that the AIS recommends, but when the same technology is deployed in areas where that experience is lacking, the patient is at risk of receiving inappropriate treatments due to a lack of clinical safeguarding. In Mongolia, the specialised clinical knowledge base does not appear to have ever been present. A clinician may look to the SDC to provide reassurance that the AIS can be trusted, but that reassurance is lacking in this scenario. It would have been reassuring to know that Watson had been exposed to critical review by third parties outside IBM, but Ross and Swetlitz (2017) assert that this did not happen. It also appears that the company has also distanced itself from Watson's own outputs when applied in clinical practice; an IBM executive has been quoted by Hengstler et al (2016), p.115 saying that "Watson does not make decisions on what a doctor should do. It makes recommendations based on hypothesis and evidence based [sic.]".

3) The AIS is opaque to both the clinician and the SDC; its processes cannot be understood, resulting in outputs which may lack context (Mukherjee 2017). This risks a AIS's outputs being misunderstood, for example, the AIS being used in a context which does not match its intended use resulting in its outputs being misapplied (Doroszewski 1988). Here it is arguable that accountability is unachievable by anyone prior to clinical use.

From the literature reviewed so far, it may be said that opacity may interfere with one's ability to account for using an AIS in clinical decision-making, and that there are multiple scenarios where using an opaque AIS in clinical decision-making could raise issues of safety. It is suggested that there is merit in an opaque AIS being subjected to a process of validation prior to use, but that such validation needs to be understood by the clinical user as being appropriate and sufficient. Given that the clinical professional bodies require their members to be accountable and to ensure patient safety, in the absence of an appropriate process of validation, the answer to this review's first question of "is it considered permissible for a clinician to use an opaque AI system in clinical decision-making?" is currently 'no'.

Responsibility

As with accountability, no unified definition of responsibility in this context was yielded by the literature searches. Again, for the purposes of clarification within this review, the literature is reviewed considering the definitions provided in chapter 3.[9] On a practical level, this can mean that agent/s may be ascribed the blame or praise for the outcomes of their acts/omissions. Allocation of the responsibility for the consequences of AIS use may one day become needed if there are unintended consequences of AIS use, and the following outlines those concerns as they appeared in the literature reviewed.

Understanding of the allocation of responsibility is illustrated by Whitby (2015); he is concerned that lack of clarification regarding who holds responsibility for actions involving AIS use could result in detriment to patient welfare (e.g., stakeholders blaming the AIS or each other rather than proactively ensuring that the AIS is functioning and being applied correctly). Can AISs be responsible for themselves? Luxton (2014) is concerned that systems do not share the human suffering of moral consequences. Van Wynsberghe (2014) agrees, if a AIS cannot be punished, it cannot assume responsibility for roles incorporating the care of humans. Whitby (2015) warns that managers of AIS users should be explicit that clinicians cannot blame the AIS to avoid responsibility.

The literature seemed to agree that clinicians should take responsibility for opaque AIS's that they chose to use in clinical decision-making. For example, Van Wynsberghe (2014) holds that AISs can be delegated small roles where no harm to the patient can be caused; but may only carry out these roles when supervised by clinicians who hold responsibility for the patient. Here the clinician is the one who ensures that the AIS works as intended when deployed. Delvaux (2017) asserts that an AIS should

[9] Thus, responsible behaviour is characterised by the "common norms which govern conduct" (Oshana 2004, p.257). Additionally, personal moral responsibility is the individual's obligation or duty to ensure that something is acted or obtained; this individual's burden is attached to them due to the role that they fill within the context being discussed (Zimmerman, 1992, p.1089).

assist the clinician; that the planning and final decision for the execution of a treatment must be made by a clinician.

Pouloudi and Magoulas (2000) warn that an AIS's user is responsible for evaluating its outputs before using them. Whitby (2015) insists that clinicians should maintain responsibility for outcomes when they use AISs and that clinicians ought not be allowed to escape that responsibility by blaming the AIS should negative outcomes arise. Kohane (interviewed in Sukel 2017b) explains that if there is a human clinician in the decision-making loop, the responsibility remains with them. The human would undertake to ensure that that which is advised by the AIS is safe and appropriate for the patient, and the responsibility for the patient outcome of using that AIS output lies with them. When discussing AISs which make diagnoses, Kohane (interviewed in Sukel 2017b) also states that if there is a decision-making disagreement between an AIS and the clinician using it, human third parties could "break the tie".

The literature was less clear regarding allocating responsibility to technologists or the SDCs which employ them. The ACM code states that "public good is always the primary consideration" and that its members should minimise the negative effects of their work such as threats to health and safety (ACM, 2018). But, beyond the ACM Code, the literature is divided, and it all seems to depend on what it is that one is asking SDCs to be responsible for.

Delvaux's report (2017, point 56), Doroszewski's essay (1988) and Vallverdú and Casacuberta's discussion (2015) place responsibility for an AIS's accuracy at the door of the person who trained that system. Doroszewski (1988) stresses the importance of this responsibility upon the SDC as the consequences of misrepresenting information in an AIS to be used in healthcare can be dire. Doroszewski (1988) demonstrates that allocation of responsibility to an individual may not be easy, though. In the case of multiple authors making additions to an AIS, it might not be obvious who will take responsibility for the accuracy of the AIS which is ultimately created.

Some SDCs are pushing back against this idea of responsibility and refer to how their AISs are designed to defend against being assigned responsibility for the use of their creation's outputs. Fenech et al.'s interviews (2018) identified the opinion that SDCs should not hold responsibility for a system when it was designed to assist clinical decision-making rather than replacing it (e.g., the DeepMind system); that in this case, the responsibility remained with the clinician using it. This opinion was echoed by an IBM spokesperson interviewed in Hengstler et al. (2016), p.115), who said that Watson makes recommendations for a clinician and does not make the ultimate decision for patient treatment. Inthorn *et al.*'s discussion (2015) holds that doctors should retain the authority of decision-making as justifying and explaining the treatment to patients is their role, not the SDC's. The only exception to this rule is when a AIS is designed to work without clinical supervision; here, the SDC should be held responsible for AIS outcomes (Fenech et al, 2018; Kellmeyer et al. 2016).

Should multiple stakeholders hold responsibility for a system's use rather than an individual?

When a SDC releases a system, multiple stakeholders can hold responsibility throughout the process of making, approving, and using a system (Pouluoudi and Magoulas, 2000; Kellmeyer *et al*, 2016). Nissenbaum (Nissenbaum, 1996) calls this the 'problem of many hands'. A system malfunction may originate from the SDC who designed it (Vallverdú and Casacuberta, 2015), but the regulator may fail to recognise or act on a discovered flaw (Kellmeyer *et al*, 2016), or the clinician or patient may use the system incorrectly or use it without employing care and attention having been warned that it is not infallible (Whitby, 2015). Despite multiple stakeholders interacting with the system before it is ultimately used, whomever decides to use the system's outputs at the point of use will shoulder the responsibility of a negative outcome from using the system's outputs (Whitby, 2015; Fenech et al, 2018; Kellmeyer *et al* 2016). This seems unfair that the end user bears this responsibility when so many other parties have also been involved in creating the system.

To mitigate this imbalance, Pouloudi and Magoulas (2000) propose defined obligations of interdisciplinary working between the SDC and the clinicians. They suggest that all stakeholders act professionally and that technologists and the SDCs which employ them should be disciplined should they fail to act as professionally as clinicians are obliged to; again though, I note that there is no professional regulatory infrastructure to enforce this.

It's not just about the clinicians and the SDCs though. A societal perspective is suggested by Doyle (in his evidence to House of Lords Robotics and Artificial Intelligence Report, 2016), who argued this is needed in system development rather than a solely technological viewpoint. By stakeholders collaborating and communicating their own limitations to each other and disclosing limitations when they are found in the system, the goal of successful system development is more likely to be reached (Pouloudi and Magoulas, 2000). Pouloudi and Magoulas (2000) stipulate that system limitations must be communicated to all users (including patients, even if the system is being deployed by a clinician) before the system is put into use. This goal could be achieved by employing Whitby's (2015) advice that training of any system which is to go into service should demonstrate any system limitations and remove the myth that the system is infallible.

Despite saying that responsibility is contextual and should be shared out between stakeholders, Pouloudi and Magoulas (2000, p.461) somewhat contradict themselves by warning that this involvement does not change the fact that the user is responsible (rather than anyone else) for evaluating the system's outputs before using them, because the user should compensate for a system's deficiencies. They advise that SDCs should continue to monitor the dynamic system after it has been released for use to clinicians to ensure that its learning strategies are appropriate, and its outputs correct (Pouloudi and Magoulas, 2000 p.464), but, to summarise their position, although this monitoring must take place, the clinician user is still responsible.

Whitby (2015) argues that SDCs must share responsibility for consequences with clinicians when inappropriate advice is given by an AIS and used by the clinician. Whitby's (2015) position is overall more collaborative than other authors, arguing that responsibility should be shared between clinicians using a system and the SDCs who design it rather than by the clinician alone. He suggests (p.232) that in the event of a negative outcome from using a system's outputs, interdisciplinary investigations should include all stakeholders and that blame should not be allocated, rather that the aim should be to prevent future harms. Yet, this seems somewhat naïve as without prior agreement of shared responsibility, clinicians could still be held *professionally* accountable for using the AISs (via their profession bodies) whereas SDCs/technologists will not; it appears that clinicians may carry the burden of responsibility after all.

The question of the allocation of responsibility if harm to a patient is caused due to a clinician using an AIS does not appear to have been fully resolved in the literature reviewed. There seems to be agreement that clinicians should be responsible for their action of choosing to use the AIS and the outcomes of that choice, but there is no recognition that technologists or the SDCs which employ them should be allocated responsibility for the consequences of AIS use. This seems strange as the AIS has been designed to affect patient care and deployed to the clinical area to specifically influence the decision-making of the clinical user. Whilst Whitby (2015) suggests that responsibility could be shared, no authors specify how this responsibility could or should be allocated between the two key stakeholders. Instead of making definitive statements about who should be responsible for what, Whitby's (2015) 'investigate don't blame' idea takes a softer approach, yet this proposal lacks a detailed plan of action. The current overall lack of clarification regarding the allocation of responsibility between stakeholders should be concerning for those affected, as they would be unable to predict or plan for the consequences of deploying or using AIS in the clinical environment. Actions which entail harmful consequences must be answered for, and the route that answerability takes is often legal. For this reason, this review turns to consider liability next.

Liability

The Government Office for Science (2016) and Clarke (in House of Lords: Select Committee on Artificial Intelligence 2018) confirm that there is no body of case law yet to guide negligence and liability in this area. Little (in House of Lords: Select Committee on Artificial Intelligence 2018) advises that if civil and criminal liabilities and responsibilities are not considered before individual cases are brought, the resolutions resulting from existing legal frameworks may not be desirable. Yeung (in House of Lords: Select Committee on Artificial Intelligence 2018) points out that if the courts have to find a solution for responsibility and liability then someone would have been harmed already. The Law Society (House of Commons: Science and Technology Committee 2016) explains that the downfall of relying on common law is that legal principles are developed after an untoward event, which is both expensive and stressful to stakeholders. Additionally, it can be said that reliance on common law makes forward planning difficult to carry out as well as making the planning and acquisition of appropriate insurance problematic.

Due to the lack of clarity, the Law Commission was asked to investigate if current legislation is adequate to address liability and to make recommendations on this area (House of Lords: Select Committee on Artificial Intelligence 2018). There has been no word of the Law Commission starting this requested work, and no one has speculated what the content of this review could be. Bainbridge (1991) discussed how the areas of negligence and contract could be applied in English law when AISs are used in the clinical context, but this work is of limited value nearly 30 years later as negligence law has moved on.

In summary

As already noted, this review's first question of "is it considered permissible for a clinician to use an opaque AI system in clinical decision-making?" is currently 'no'. This review's second question asked, "what concerns are there about opacity, accountability, responsibility and liability when considering the stakeholders of SDCs and clinicians in the creation and use of AI systems in clinical decision-making?". This literature review has suggested that there are multiple multifaceted concerns that:

1. it is not possible to account for an opaque AIS's outputs; thus, if one cannot account for the outputs, one cannot give a reasonable account for choosing to use those outputs.
2. if SDCs provide opaque AISs to aid clinical decision-making, they may risk clinicians choosing not to use them as it would affect their ability to be accountable practitioners.
3. the formulation whereby responsibility is allocated is not concrete. There seems to be a consensus that clinicians should hold responsibility for choosing to use an opaque AIS, but there is no such accord for technologists or the SDCs which employ them joining them in holding that responsibility, even though some authors indicate that responsibility could be shared.
4. there is no case law or legislation in the law of England and Wales which is specific to negligence and liability cases in the use of AISs in clinical decision-making; this lack of clarity

might prevent stakeholders from confidently planning for the undesirable scenario of patient harm resulting from the use of an AIS.

5. waiting for the courts to find a solution to the allocation of responsibility and liability would require that someone came to harm first.

It is reasonable to say that there is a current opportunity to proactively address these issues before harm takes place, rather than allowing harm to take place and retrospectively allocating ethical and legal responsibility. If this opportunity is taken, avoidable harm could be prevented.

To summarise: this literature review suggests that there are multiple concerns about opacity, accountability, responsibility, and liability when considering the key stakeholders of technologists/SDCs and clinicians in the creation and use of AIS in clinical decision-making. Accountability is challenged when the AIS in use is opaque, and allocation of responsibility is somewhat unclear. Both ethical and legal analysis might help stakeholders to understand their obligations and prepare, should an undesirable scenario of patient harm eventuate when AISs are used.

Limitations

There are limitations to this review.

This review found a lack of consistency in the language used when considering opacity as well as an enormous variety of subgroups of AIS systems in use. These two factors challenged me to appropriately and inclusively recognise the multitudes of terms and programming language in existence which populate the literature discussing material pertinent to this review.

Regarding the subgroups of algorithm types in AI, this review intentionally did not identify a particular group (such as machine learning) lest the discussion become side-tracked by specifically *which* AIS's are being used rather than consideration of *how* the AIS are being used. Currently, machine learning is well-represented in the current debate, but this has not always been the case and another subgroup may prove to be more popular in the future (House of Lords: Select Committee on Artificial Intelligence 2018). The lack of consistent terminology made the literature searches challenging; for example, AI opacity could also be described as the AIS being a black box, or that there was a lack of AIS transparency. Increasing the number of search terms to attempt to capture the variety of terms did not improve the number of search results returned, nor the relevance of those search results. Ultimately, 'opacity' was adopted as the primary descriptor and employed as the search term as it yielded the highest volume of relevant results in the literature searches.

I suspect that there has likely been much relevant material from a variety of worthy sources which has been lost to this review due to the changing nature of how information (especially regarding technology) has been communicated in recent years. Relevant and worthy ideas, concepts and opinions are no longer routinely published in the traditional way, i.e., via peer-reviewed journals, thus are not admitted to academic database searches which are the main pathway for discovering material for a systematic review such as this. For this reason, media items from outside of the realms of the traditional academic sources were selected when they were determined as pertinent to this review. For example, Ross and Swetlitz's useful and demonstrative report of the use of IBM Watson would have been lost to this review had media been excluded.

Updates to this literature review

This literature review was undertaken as a closed piece of work to provide a foundation for this thesis in 2018 and was published in 2021. Since then, there have been numerous publications relevant to the scope of this review. This section presents the most pertinent findings to bring this chapter up to date and to ensure its finding's ongoing relevance in this thesis.

There has been a lot of work discussing AI applications in general, rather than limited to healthcare. The generalisability of this work makes it pertinent to note when considering the use of AIS in healthcare; thus, these works shall be noted first.

There have been multiple ethical codes and standards released by various organisations and authors which pass comment on the allocation of ethical and/or legal responsibility when AI is used.

The Engineering and Physical Sciences Research Council's Principles of Robotics (Boden *et al*, 2011) were one of the earliest ethical frameworks proposed. Multiple others have followed, ranging from non-profit initiatives, for example the Future of Life Institute (2017), to high level international organisations such as the European Commission's High-Level Expert Group on Artificial Intelligence (2019). Jobin *et al*'s (2019) recent comprehensive and acclaimed review of global AI ethics guidelines noted 84 ethical guidelines. These were found to have five converging ethical principles (transparency, justice and fairness, non-maleficence, responsibility, and privacy) but with diverging approaches regarding areas such as interpretation and importance.

Some organisations have offered practical council in the form of guidance and standards. For example: the Institute of Electrical and Electronics Engineers which produced the Ethically Aligned Design series (IEEE, 2017) and the IEEE 7000 series which offers standards including guidance for stakeholders to address ethical concerns during system design (IEEE, 2021). The OECD's (2018) AI Principles include those of transparency, explainability, robustness, security, safety, and accountability. These principles will in principle address the problem of opacity by encouraging SDCs to develop AISs which are understandable to those who use and are affected by them (e.g., clinicians and patients). Where the OECD's (2018) principles advocate for accountability, they require actors to describe their actions and decision-making processes throughout an AISs lifecycle; this will help stakeholders to fully understand scenarios and perform investigations where responsibility and liability need to be determined. Yet, despite all of these above works, an overall accord of *how* responsibility ought to be allocated has not been attained and might likely too simplistic a goal.

IEEE (2021) briefly notes that "it is in the public interest to know who is responsible under the law" (p.67) but neither explores nor indicates where that responsibility may lie. The UK Statistics Authority (2021) asks SDCs if "a chain of human responsibility [has] been established, with each stage of the project's lifecycle being documented to show the human oversight" but does not offer how that responsibility should be allocated. Neither does UNESCO (2021), who state that "The ethical responsibility and liability for the decisions and actions based in any way on an AI system should always ultimately be attributable to AI actors corresponding to their role in the life cycle of the AI system" (UNESCO, 2021, p.11), but without elaboration of how responsibility or liability would be determined. The Panel for the Future of Science and Technology's (2020, p.26) study on AI ethics admitted that "different applications of AI may require different frameworks"; thus, a framework which may be, for example, appropriate for a military application of AISs will not be appropriate for the clinical applications which this thesis is exploring. This somewhat explains the lack of consensus surrounding the allocation of responsibility and makes applying a substantial amount of the new literature challenging. Furthermore, these new additions continue to be wholly voluntary codes and standards. They make good recommendations – for example, IEEE (2021) advocate that SDCs develop core values such as accountability and transparency in their working practices, and UNESCO (2021) states that there should be commitment by AIS actors[10] to ensuring that AISs are meaningfully explainable thus users may account for their use - but no regulatory force underpins their content. When considering ethical stewardship, UNESCO (2021) proposes that, to deal with the lack of legal frameworks related to AIS use, stakeholder involvement should be encouraged to develop 'norms'[11] into mature best practices, laws and regulations – this approach does not proactively inform those who wish to develop and use AISs responsibly whilst those norms are being identified.

So, many major organisations have noted now that ethical responsibility and legal liability need to be considered, but none have attempted to do so in depth. The aforementioned review of global AI ethics guidelines performed by Jobin *et al* (2019) had comparable findings to that which this review

[10] For the purposes of this thesis, SDCs.

[11] They do not elaborate which norms, presumably they mean ethical and practical.

presented above. They noted that recommendations regarding legal liability and attribution of responsibility ought to be clarified prior to the use for AIS, yet they found no consensus on *who* ought to be responsible for the consequences of AIS use.

At the level of government, the "Algorithms in decision-making" (House of Commons Science and Technology Committee, 2018) report identified that someone needs to be responsible for an algorithm, but that the allocation of this responsibility is uncertain; therefore there is sense in their advising that responsibility needs to be considered when designing a system so that incorrect assumptions around it can be avoided (but they do not specify how to allocate that responsibility). They state that transparency (i.e. relieving AIS opacity) is key to ensure accountability, and that to create a system of accountability, standards are needed. However, they do equate 'transparency' with 'lack of opacity' and, as argued above, making processes transparent (i.e. visible) does not necessarily make them any less opaque.

Some governmental guidance documents have been released since the Algorithms in decision-making report. They refer to the development of AI applications of non-specific use, but the consideration of accountability and responsibility for the use of AI is spread out in these documents in a way which is challenging for readers to usefully assimilate. For example, there are three distinct guidance documents from the Government's Digital Service. The first document, the 'Data Ethics Framework' (2020), has the AI adopter consider accountability by asking how they ensure that they can demonstrate that their AIS has achieved the correct output. The second document, 'Understanding Artificial Intelligence Ethics and Safety' (2019), stresses the need to build cultures of responsible innovation (e.g., that an AI project ought not be discriminatory), but without discussing responsibility for the consequences of AIS use. The third document, 'Assessing if artificial intelligence is the right solution' (2019), instructs the use of a responsibility record to show who was responsible for what when constructing an AI project. Accountability is aided here by helping developers and adopters to ensure that they can ascertain what has gone wrong if unwanted consequences eventuate, but this guidance would not be able to help allocate responsibility either legally or ethically beyond encouraging AIS developers and adopters to create a record of events showing what happened and why. Such a record might, presumably, be useful for the determination of responsibility should unwanted consequences arise in the future, but that would be a separate process informed by the record.

Instead of becoming overwhelmed by the vast and non-specific global conversation which has erupted around AI ethics since 2018, the focus of this thesis remains firmly in its niche. The reader will note that, where relevant, relevant publications (found through an unstructured monitoring of news and current public and academic debates rather than via a formal review process) have been woven into the matter which constitutes the remainder of this thesis. But there have been several developments, at both global and national levels, concerning the development and potential use of AIS in the clinical NHS environment of England and Wales; these must be acknowledged before we proceed to the next chapter.

The WHO's (2021) guidance on Ethics and Governance of Artificial Intelligence for Health wants everyone to benefit from the use of AISs, and says that all involved should work to reduce the risk of AIS harms. WHO (2021, p.26) promotes the development of AISs which allow the user to account for their actions by specifying that AISs should be "intelligible or understandable to developers, users and regulators", thus allowing meaningful public consultation regarding AIS development and use. They discussed the problem of the allocation of responsibility when persons have been harmed because of AIS. The way that the WHO says responsibility should be allocated, though, is somewhat confused. They note the problem of many hands (which they call 'diffusion of responsibility') and suggest that a collective responsibility model could be employed to prevent an outcome of "everybody's problem becomes nobody's responsibility" (p.28). However, even though they suggest a collective model, they then contradict themselves by discussing how responsibility can be applied to individuals whilst failing to offer a clear method for how responsibility should be allocated. They do this by firstly saying that there are "reasons for not holding clinicians solely accountable for decisions made by AI technologies" (p.44) and recognise that "if there is an error in the algorithm or the data used to train the AI technology, for example, accountability might be better placed with those who developed or tested the AI technology rather than requiring the clinician to judge whether the AI technology is providing useful guidance" (p.44). They then note that a system is a not an isolated tool available to clinicians and that they ought to employ their clinical judgement. Yet still, the WHO asks, "If the physician makes the wrong choice [in following an AISs outputs], what should the criteria be for holding the physician accountable?" (p.44). It seems unfair to look specifically and continually to hold clinicians responsible for outcomes when they would clearly be partners with SDCs in this scenario, and there was a missed opportunity to explore the interesting notion of collective responsibility. Interestingly, the WHO (2021) recognises a 'responsibility gap' where clinical users are burdened by the outcomes of using an AIS when they were not involved in its development or design.[12] They note that "assigning responsibility to the developer might provide an incentive to take all possible steps to minimize harm to the patient" (p.42). The WHO's (2021) guidance also notes that liability rules should be modified

[12] This differs from Matthias's (2004) description of a responsibility gap. Here, no one is responsible for a machine because its actions are not fixed because the machine is learning. For this reason, no one has enough control over the machine's actions to be responsible for it. This paper is highly cited, and I nod to it here, but it is not quite relevant to this thesis as it addressed autonomous systems acting without human supervision.

for the assessment and assignment of liability and that that there should be a process for those which have suffered harm due to the use of AISs. They make limited suggestions that redress should include "compensation, rehabilitation, restitution, sanctions where necessary and a guarantee of non-repetition" (p.28), but the fine details of the specifics of these suggestions are not offered and there is a lack of directly applicable suggestions relevant to the context of the jurisdiction of England and Wales.

The Department of Health and Social Care does cover England and Wales, and their flagship publication for AI is the Guide to Good Practice for Digital and Data-driven Health Technologies (2021b). This guidance does speak to opacity, accountability, responsibility, and liability. Regarding accountability and opacity, it asks those who develop, deploy, and use data-driven technology in the NHS to use best practice to explain algorithms to those who are using them. Regarding liability and responsibility, it says that the implications of responsibility and liability when introducing AI should be considered. Once again, the guidance does not specify how stakeholders ought to consider the implications of responsibility and liability, only that they should be considered.

There have been some initiatives to bring clinicians into the digital arena, such as NHS England's (undated b) Digital Academy, which come highly advised by the recent Topol Review (2019) to safeguard patients against problems such as harm and health inequalities. Yet, despite the large volume of activity outputted by organisations such as NHSX, disappointingly, there is still no guidance crafted for the clinical professionals who may be confronted with using AISs in their practice. The actual production of guidance is outside the scope of this thesis, but if society expects clinicians to use such tools, it is essential that they are provided with comprehensive practical high level guidance informing them how to engage with AISs. Such guidance could consider wider aspects of clinical practice such as critical appraisal of the AISs that they are faced with using as well as concerns about ethical and legal responsibility for its use. No clinician is an island; as with many issues which a clinician may face within the course of their career "authoritative national ethical guidance should help to bring clarity, consistency and fairness to decision-making" (Huxtable, 2020). Without clear practical or ethical guidance, individuals, or the organisations which they work for, will be left to determine their own interpretation of the standards of clinical practices when AISs are used. This will lead to variations in the approaches to, and the quality of, care which patients receive when AISs are employed at the bedside rather than an agreed and unified standard of practice. Specific ethical guidance will affect the interpretation of more general guidance, especially when dilemmas arise that need to be resolved by the clinician. Thus, specific ethical guidance will have scope to develop once general guidance regarding AIS use has been issued. Creating guidance will be challenging, though, as the healthcare workforce will need to be prepared "for jobs that have not yet been created, technologies that have not yet been invented and problems that we don't yet know will arise" (Topol Review, 2019, p.21).

Whilst this chapter has raised several challenges, there have been two particularly positive developments which show that these challenges are to be addressed. Firstly, guidance has been issued which recommends that users are given information about opaque systems which is meaningful; for example: the MHRA's (2021a) principles in Good Machine Learning Practice for Medical Device Development noted that users be given clear information about the system they are presented with. That system limitations and the "basis for decision-making" should be available for users and patients, as well as "a means to communicate product concerns to the developer" (MRHA, 2021a). Other documents such as the guide to good practice for digital and data-driven health technologies from the Department of Health and Social Care in 2021(b) and the Data Ethics Framework from the Government Digital Service in 2018 recommend that those implementing AIS projects can *show* that their AISs can reach correct outputs, and that the Information Commissioner's Office's (2020) best practice on explaining AISs should be employed. On a practical level, projects are being undertaken to relieve opacity. For example, the Project ExplAIn (Information Commissioner's Office and the Alan Turing Institute, 2020) supports SDCs in making their AIS's meaningfully explainable through being explanation-aware throughout the cycle of AIS development, thus tackling the problem of opacity. I interpret this collection of efforts as recognition that users will struggle to understand (thus struggle to account for their use of) opaque systems and, in response to this, that organisations are attempting to address this issue via the publication of guidance which pushes SDCs to supply AISs which users can meaningfully understand and thus user accountability is retained.

The second positive development, which has the potential to help guide technologists' practice to a decreed benchmark, is the announcement from the Royal Statistical Society (in collaboration with other influential bodies) who are developing accreditation and preparing industry-wide professional standards for data science (Royal Statistical Society, 2020). This framework would envelope technologists who develop AISs. This project is only in its very early stages, so cannot be discussed further in this thesis, but those involved recognise that standards of practice in this field are not on par with others, such as those of their clinical counterparts (Royal Statistical Society, 2020).

This closes the roundup of the latest key documents to be considered by this literature review; the following will encapsulate the essence of the key observations found in those documents and how it shall affect the rest of this thesis.

Conclusion

Through this review, I have come to realise that accountability and opacity are factors in the ethical discussion rather than discrete ethical issues themselves. I note that accountability, transparency, and explainability are all frequently mentioned in the government documents discussed above. The issue of accountability is very much rooted in the technical problem of opacity; opacity needs to be addressed first by means of providing understandable and meaningful interpretations of an AIS's processes for its users. If projects such as ExplAIn are successful, and agents become able to understand and account for the AIS's which aim to inform a clinical decision, it will allow clinical users to fully account for their decision to (not) use an AIS's output. That development would mostly resolve the problems raised by opacity and the need for accountability detailed above. Given that this work is being undertaken by others, the remainder of this thesis will focus on liability and responsibility. Opacity and accountability will continue to be touched upon, but only insofar as they inform discussion surrounding the legal and ethical allocation of responsibility.

Whilst work is being undertaken to solve the problem of accountability, this literature review noted that issues around responsibility and liability have not been adequately addressed. Regardless of whether an AIS is opaque or not to its user, the process and criteria for allocating responsibility is not clear and there is a lack of legal clarity specific to negligence and liability cases in the use of AISs in clinical decision-making. One could have hoped that the more recent items identified in this section would have begun to address these issues, but instead they consider how to ethically and responsibly develop and use AISs; none address how to deal with harmful consequences of that use, save a single non-specific reference in the Guide to Good Practice for Digital and Data-driven Health Technologies (Department of Health and Social Care, 2021b), which nonetheless fails to outline how liability and responsibility ought to be allocated between stakeholders. Whilst planning for successful development, deployment, and adoption of AIS in the clinical environment is laudable, the consequences of AIS use must be planned for if there is a risk of patient harm; to not do so indicates a lack of insight to that risk and an abundance of optimism with a lack foresight about what could go wrong.

Rather than addressing this gap, the organisations who are addressing the development and adoption of AIS in the clinical environment in the NHS of England and Wales still have not achieved a level of specificity which would aid the clinical user or the SDC in fully understanding where ethical or legal responsibility lies should a patient come to harm due to the use of AIS in clinical decision-making.

In summary, actors need this information to be able to judge for themselves if the development and use of AIS is an activity which they wish to engage with, or if the burden of ethical or legal responsibility is unacceptably high.

This literature review's findings confirm that the two key questions asked in chapter 3 of this thesis remain comprehensively unanswered:

- How, according to current law in England and Wales, will legal liability be allocated between clinicians and SDCs when AISs are used in clinical decision-making?
- How can ethical responsibility for the consequences of the use of AIS in clinical decision-making be determined and allocated?

The remainder of this thesis seeks to answer these questions. Stakeholders need to be informed of specific consequences of the implementation of AISs in healthcare: specifically, the possible harmful consequences of AIS use and the subsequent ethical and legal repercussions which could impact on patients, SDCs and the clinical users. If this is considered prior to the adoption of AISs to aid clinical decision-making, then informed and aware stakeholders will be better equipped to think through the risks before choosing to deploy or use such technologies to inform patient care.

The next chapters of this thesis will analyse and specify how legal liability can be allocated in the context of negligence, and then will provide ethical analysis of how responsibility can be ethically allocated in light of the legal analysis performed.

Chapter 5: Concerning the tort of negligence

As noted in chapter 4's literature review, there is currently a lack of clarity surrounding the sufficiency of legal mechanisms for liability when applied to the use of AISs. In 2018, the House of Lords Select Committee recommended that the Law Commission investigate whether current legal principles were adequate to address liability issues when using AI and to make recommendations in this area, but a formal reference has not yet been made by the Government. This chapter contains a legal analysis of some of these issues within the jurisdiction of England and Wales. Specifically, it asks, based on current law, how will legal liability be allocated between clinicians and SDCs when AISs are used in clinical decision-making?

The fields of law which may be considered in the use of AISs are vast. Whilst areas in liability such as product liability, including medical devices regulation would be highly worthy of discussion regarding the use of AIS in the clinical environment, it would be impossible to cover all such areas within the confines of a single thesis. This work is limited, and the literature review indicates a need to consider how responsibility may be allocated to stakeholders. As a clinician myself, I am driven to understand where and why a clinical practitioner's actions when using an AIS may be considered negligent. As an SDC would provide an AIS to influence a clinician's decision-making, I also wish to understand how the actions of an SDC might be either considered separately or intertwine with a clinician's if a liability claim were ever brought. For these reasons, the following analysis, and the overall theme of the thesis itself, is focussed on 'fault' of individual actors rather than 'strict' liability as related to the AIS. However, it is worth briefly exploring now the reasons why a negligence claim would be a real possibility, and could be a preferable route for a claimant than product liability.

Outside of - but related to - this thesis, I performed and published a legal analysis with Kit Fotheringham regarding the use of AIS products liability (Smith and Fotheringham, 2022). In this paper we noted that:

> 1) The Consumer Protection Act contains the criteria for product liability in part 1, however, there are limits to the remedy which is offered by the Consumer Protection Act 1987; e.g., it is unclear if the AIS would be classed as a product if it is not embodied as a component of a physical product.[13]

[13] Given that much software access is via clouded services, it is not at all unlikely that AISs deployed to the bedside will be retrieved via third party devices - much like accessing email via a smartphone, the email service is accessed via the phone but the smartphone is separate unit to the email service.

2) That the court might well place some of the burden of responsibility for unsafe usage of products onto clinicians because in *Wilkes v. DePuy International Ltd* 'an intervening healthcare professional is a relevant circumstance' in causation.

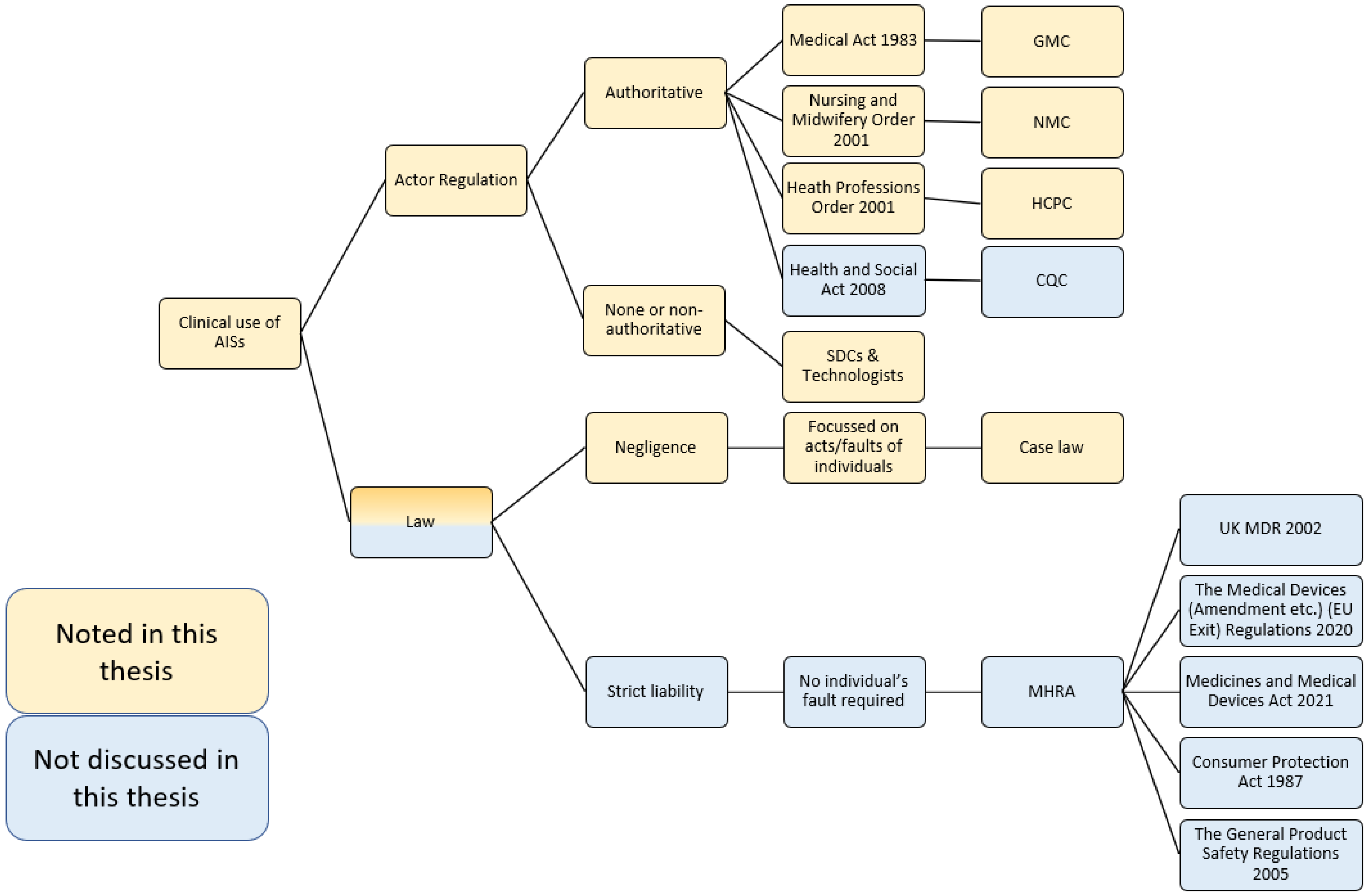
3) In *Howmet Ltd v. Economy Devices Ltd*, the court held that the chain of causation is broken at the moment a user becomes aware of a defect in the product that they then continue to use regardless.

Subsequently, a claim might not be possible using the Consumer Protection Act, and a clinician who uses an AIS - having been warned (e.g. by an SDC) that it will not always output perfectly appropriate information for use in clinical decision-making - might find the legal responsibility for defects placed with them. Because of this, the *Wilkes* and *Howmet* cases bring us back to the negligence work that this thesis discusses and the problem of *novus actus interveniens,* which will be explored later in this chapter.

Additionally, a negligence claim may be more attractive to a patient because they may prefer to claim specifically against the readily identifiable clinicians who had treated them rather than a SDC which is distantly situated away from the bedside. This preference may be, in part, driven by a tendency of the NHS to settle claims in a non-adversarial manner. NHS Resolution resolved 74.7% of claims without formal proceedings in the period of 2020-21, (NHS Resolution (2021a). – suggesting that a claim against the NHS has a good chance of success.

Figure 5 outlines the field of regulation, legislation, and case law regarding AIS use in clinical decision making as discussed in this thesis. This figure is far from exhaustive and fails to take into account other influential and/or authoritative actors in the deployment of AIS in the healthcare; for example, bodies such as NICE which provides evidence-based guidance to guide clinical practice (National Institute of Health and Care Excellence, 2021), or the Department of Health and Social Care's NHSX (NHSX, undated) division which specifically exists to address digital health issues.

Figure 5: The (non-exhaustive) field of regulation, legislation, and case law regarding AIS use in clinical decision making as discussed in this thesis


Clinical use of AISs
Actor Regulation
Authoritative
Medical Act 1983
GMC
Nursing and Midwifery Order 2001
NMC
Heath Professions Order 2001
HCPC
Health and Social Act 2008
CQC
None or non-authoritative
SDCs & Technologists
Law
Negligence
Focussed on acts/faults of individuals
Case law
Strict liability
No individual's fault required
MHRA
UK MDR 2002
The Medical Devices (Amendment etc.) (EU Exit) Regulations 2020
Medicines and Medical Devices Act 2021
Consumer Protection Act 1987
The General Product Safety Regulations 2005
Noted in this thesis
Not discussed in this thesis

This chapter, then, provides legal analysis which examines how the tort of negligence can be applied in the scenario where a clinician uses an AIS's inappropriate recommendation and the patient comes to harm as a result. Firstly, the conditions which put the clinician at risk of carrying the burden of claims in this scenario will be identified along with how an SDC could be able to limit their liability. It will be argued that this situation is unfair to the clinical user as the clinical decision-making space has been modified by the SDC via the deployment their AIS.

A note on vicarious liability

As mentioned in the thesis introduction, this work chiefly assumes the clinical decision maker is an individual clinician. A clinician may, often interchangeably, work as a sole agent (e.g., a self-employed general practitioner) or be employed by an NHS organisation which provides the clinical environment and tools for healthcare. Regardless of their employment status, it is the individual clinician who makes the decisions regarding individual patients at the point of care. Similarly, the term 'SDCs' is dominant in this thesis. Whilst a single technologist may make their own contribution to a project, they would likely be part of a wider organisational effort to create and deploy an AIS for a clinical setting, rather than creating and deploying an AIS entirely alone. Yet, regardless of whether they work alone or within a larger organisation, a single technologist might be individually responsible for their own contribution (e.g., a piece of code) to the construction and deployment of an AIS.

The actions of those acting alone and as part of a team are important when considering vicarious liability. Vicarious liability is where a defendant is held legally responsible, therefore liable, for the tortious acts of another actor (the tortfeasor) (Giliker, 2017). A classic demonstration of the relationship between a defendant and a tortfeasor here would be between an employer and their employee (Giliker, 2017). Giliker (2017, p.273) calls vicarious liability "a rule of convenience", as whilst a claimant may claim against a negligent employee, they will generally sue the employer as the employer generally has the deeper pocket. This makes prior preparation for the possible negligent actions of an organisation's employee's advisable; for example, NHS Resolution provides indemnity schemes for NHS activities (NHS Resolution, 2021b).

Yet, if the AIS is created by an organisation, it remains the SDC's final decision to deploy the technology; it also remains the clinician's choice at the point of care if they will use the output which the AIS is providing them. Thus, for this chapter, despite the presence of vicarious liability, the predominant use of terminology referring to 'clinicians' and 'SDCs' remains in this chapter, as stipulated in this work's introduction, unless specifically relevant to the point being made. Vicarious liability is an interesting area of tort law (especially as the employer may attempt to recover damages from the negligent employee);[14] however, specific questions of *who* within an organisation will be claimed against for specifically *what* will not be explored. To do so would detract from the examination of the legal issues of the tort itself, where this chapter now returns.

Clinical negligence

Clinicians, and the hospitals in which they work, owe a duty of care to their patients. This is noted in *Barnett v. Chelsea and Kensington Hospital Management Committee* and *Darnley v. Croydon Health Services NHS Trust.* Generally, patients are doubly vulnerable, by virtue of their health condition and relative lack of clinical knowledge. Where a clinician has interpreted medical information and proceeds with a treatment plan that they have developed based on their clinical opinion, the clinician owes a duty of care towards their patient in the usual way (the case law underpinning clinical conduct is explained and explored more fully further on this chapter).

The example of the use of IBM's Watson for Oncology in Mongolia, discussed in the literature review, indicates that situations may arise where a clinician might rely upon an AIS to provide clinical recommendations. This poses a novel evidential problem for the claimant, as it is difficult to discern the relative influence of the human clinician and the non-human AIS. The clinician could be seen as a third party, a conduit for medical decisions that have been generated outside of the relationship of care that the clinician has with their patient.

Nevertheless, justice is not served by excluding claims where the clinician has chosen to use an AIS in providing treatment, because the clinician is an autonomous actor who has contributed to the outcome by choosing to use the AISs recommendation. Because clinicians are autonomous actors, some SDCs are adopting a position that presents clinicians as the sole guardians of system safety. As noted in the literature review, the unnamed IBM executive claimed that "Watson does not make decisions on what a doctor should do. It makes recommendations based on hypothesis and evidence based [*sic*]" (Hengstler *et al*, 2016, p.115). Thus, clinicians find themselves trapped in a 'moral crumple zone' (Elish, 2019, p.40) where they become answerable for the AIS because of their choice to use the technology. Assigning full legal liability to clinicians is convenient for SDCs, yet this is grossly unfair on clinicians as it disconnects accountability from the locus of control. The clinician may act as an independent, knowledgeable intermediary between the software's recommendations and the patient, but in practice is encumbered with the responsibility for computer-generated clinical advice over which they have only limited influence.

[14] Either by a breach of the implied term of the employee's contract to use reasonable skill and care (*Lister v. Romford Ice and Cold Storage Co.*), or because vicarious liability is joint and several due to the Civil Liability (Contribution) Act 1978 (Giliker, 2017).

Two key cases underpin the consideration of clinical conduct in almost all negligent treatment and diagnosis claims: *Bolam v. Friern Hospital Management Committee* and *Bolitho v. City and Hackney Health Authority*. Clinical conduct is not usually considered negligent per *Bolam* if it is in accordance with a responsible body of opinion, and thus satisfies the standards of other responsible medical professionals. *Bolitho* requires, further, that the evidence presented by the 'body of medical opinion' supporting a clinician's conduct should have a logical basis. IBM (undated) promotes Watson for Oncology as a combination of the expertise of 'leading oncologists' in cancer care with 'the speed of IBM Watson to help clinicians as they consider individualised cancer treatments for their patients'. This implies that the AIS has the combined knowledge and ability of a large body of senior and responsible professionals, however, IBM does not promote Watson as a decision-maker and positions the clinician as the clinical decision-maker.[15] Yet, if Watson is nevertheless widely accepted as a responsible body of opinion within the clinical community, then relying on it might pass the *Bolam* test. It might also satisfy *Bolitho*, as it is not illogical for a clinician to assume that an AIS designed to aid clinical decision-making could reach better conclusions than themselves.

It was held in *Wilsher v. Essex Area Health Authority*[16] that the duty of care can be discharged by referring to a senior knowledgeable colleague for assistance, raising the question of whether an AIS can be considered as a 'senior knowledgeable colleague'. A clinician might not be successful in arguing that digitised expertise in the form of an AIS is equivalent to human proficiency, but they might be successful in arguing that an SDC had presented their AIS as dispensing reliable expert advice, and thus they were justified in relying on it on the assumption that its reasoning was superior to their own.

Thus, if an AIS such as Watson for Oncology is being presented by its SDC as dispensing expert advice that is a combination of the expertise of leading oncologists – on a level with a senior knowledgeable colleague - it would seem plausible for a clinician defendant to argue that in relying on the AIS, they are in fact (a) acting in accordance with 'a body of responsible medical opinion' (*Bolitho v. City and Hackney Health Authority*) or (b) consulting with (the equivalent of) a senior knowledgeable colleague to achieve the same aim (*Wilsher v. Essex Area Health Authority*).

As per *Maynard v. West Midlands Regional Health Authority*,[17] the law does not presume to differentiate between contradictory clinical opinions. For instance, an AIS may create a

[15] *"I underline that Watson does not make decisions on what a doctor should do. It makes recommendations based on hypothesis and evidence based [sic]"* (Hengstler et al, 2016, p.115)

[16] The case went to appeal in the House of Lords ([1988] AC 1074), but was concerned instead with causation and was silent on the question of referring to senior colleagues.

[17] Lords Scarman (at 639) in the House of Lords said that a "preference for one body of distinguished professional opinion to another also professionally distinguished is not sufficient to establish negligence in a practitioner."

recommendation that might receive less than majority support among the medical community. Such a discovery could indeed represent a scientific advancement. Even if this was contestable, reliance on an unconventional recommendation for treatment would not necessarily be *a priori* negligent.[18] If the system's outputs consisted of recommendations that were illogical, however, the claimant may be able to show that the duty of care had been breached by the clinician who acted on that recommendation, as per *Bolitho*.[19] Indeed, the court might decide that for a clinician to abrogate their personal responsibility and instead delegate clinical decision-making to an AIS is conduct so specious that the claim could proceed on this ground. Consequently, if a clinician wishes to use AI technologies in treating patients and avoid breaching their duty of care, the cautious clinician would ideally be able to fully justify their decision-making independently of the AIS.

Ordinarily, the clinician's relationship with the patient is clear, but using an AIS introduces the 'third' agent of the SDC. This complicates the identification of the entity which is the primary cause of harm. Neither the SDC nor an advisory AIS which does not have physical contact with the patient would have been able to directly cause harm on their own; the clinician would have been the gateway to the AIS being able to harm via its recommendations to the clinician. Hence, 'but for' the clinician's conduct, the AIS's harmful output would not have influenced the patient's treatment at all. This may be plausible reason to assume the clinician is the reasonable cause.

The example of the use of IBM's Watson provides a scenario in which causation may be discussed. Were a patient to come to harm due to inappropriate drug recommendations from an AIS being adopted, it might be possible to argue that 'but for' the AIS's presence, the injury would not have occurred. A generalist clinician without specialist training in a specific illness (e.g., oncology) would not have been able to attempt to treat that illness; they would have been unable to choose or administer any specialist drugs as they would not have known which drugs could have been appropriate for the patient's condition. Consider this hypothetical scenario. Suppose the clinician has no other colleague to refer the patient to, yet if timely treatment is not provided the patient could suffer. The clinician's employing hospital has provided the AIS for this specific purpose. Under these conditions, a reasonable clinician may feel compelled to use that AIS. In this situation, 'but for' the presence of the AIS there would have been no attempt to treat the patient's illness, thus no selection

[18] *Simms v. Simms and An NHS Trust* found that the Bolam test ought not be used to inhibit innovative medical work.

[19] Lord Browne-Wilkinson in *Bolitho v City and Hackney Health Authority* (243) held that if 'it can be demonstrated that the professional opinion is not capable of withstanding logical analysis, the judge is entitled to hold that the body of opinion is not reasonable or responsible'.

or administration of an inappropriate drug, but potentially averting more serious consequences from non-intervention.

In the absence of a valid defence, a claim made by a patient against a clinician who has used a defective AIS as part of their treatment plan may be irresistible. One could argue that the court would quite rightly find that a clinician who follows an AIS's recommendations without careful consideration of the consequences has acted with negligence. Yet it appears neither fair, nor just, nor reasonable (the three *Caparo* criteria to be discussed in the next section) for negligence liability to be extended only to clinicians, given the role of the SDC in a shared endeavour with clinicians in developing and deploying AIS. When an SDC claims that its product is both dispensing state of the art knowledge, but with the additional qualification that the clinician is to make the final decision regarding how to act on this knowledge, it is not unreasonable for a clinician to consider rejecting the use of an AIS until it can be proven that its outputs can be safely relied upon. If an SDC is unable to provide adequate reassurance to clinicians, clinicians would be ill-advised to take the risk of using those systems.

It is the clinician's responsibility to ensure that they are familiar with any tools they use. This includes awareness of the risks of AISs. However, if the clinician cannot scrutinise the AIS because of the 'black box' character of the algorithm, they will be unable to take appropriate steps to mitigate these risks. This merits an assessment of the conduct of the SDC to determine whether they have materially contributed to the overall risk of harm to the patient.

Software developer company's negligence

The discussion of clinical negligence in the above section raised the possibility that the SDC might also potentially hold a duty of care and thereby share liability in negligence. For the purposes of a claim based on joint liability, the conduct of the SDC must be analysed separately. *Donoghue v. Stevenson* provides a distinctive illustration of the duty of care manufacturers owe to the end users of their products. In *Donoghue*, a drink was served to consumers in opaque glass bottles, which rendered safety inspections futile once the bottles had been sealed at the factory. Donoghue allegedly became ill as she had drunk her ginger beer without knowing that the bottle contained a decomposed snail. Extending the principle of opacity by analogy from an opaque glass bottle to an opaque AIS, it could be argued that the SDC relies on the clinician to intervene where an AIS recommends a course of action that is potentially unsafe. The opacity of an AIS may prevent a clinician from checking its outputs, but if the AIS is presented to aid the clinician's decision-making and is subsequently used, then it is positioned to influence the clinician even in the presence of its opacity. At this point, even where a clinician exercises professional judgement, it may be impossible to prove the extent to which the clinician has come to their decision independently, without any degree of influence from the AIS.

Therefore, a court will need to assess whether the SDC has taken all appropriate measures to mitigate the risk of harm to patients where their AIS is deployed in clinical settings.

Duty of Care

As the case law on negligence has developed, the duty of care has been extended 'incrementally and by analogy' with established duty situations (*Robinson v. Chief Constable of West Yorkshire Police*). Defendants cannot act with impunity, expecting that liability will rest with another party and the categories of relationship where a duty of care may be imposed can be expanded as novel situations arise (*Robinson v. Chief Constable of West Yorkshire Police).*

The following offers some ways in which the subject matter with which this thesis is concerned might be considered a 'novel' duty of care for the SDC. The 'novel' duty situation depends on the facts of the case being considered, for example, the way that the AIS is presented to the clinician for use. An SDC may present the AIS for clinical use in a form that is not embodied as a component of a physical product, and is a separate addition to the device in which the AIS is installed; this might create two problems. Firstly, it would prevent a claim against the manufacturer of the device on which the AIS is installed via a product liability approach, as the manufacturer would not have developed the AIS (as per the discussion regarding The Consumer Protection Act at the start of this chapter). Secondly, the SDC might not be treated by the court as a manufacturer as they had not manufactured a device for their AIS to be used on, therefore again preventing the use of a product liability approach. As an alternative, the courts might focus on the AIS itself, and the SDC that created and deployed it, rather than the device that the AIS is installed on. Here, they might take the position that the relationship between the SDC and the patient is sufficiently equivalent to that of a manufacturer and the patient, and would therefore be covered by *Donoghue v Stevenson*. The courts took a narrow view of what constituted a 'novel' case in *Robinson v. Chief Constable of West Yorkshire Police* and *Darnley v. Croydon Health Services NHS Trust*. Both cases recognised that there is no single general principle (in this jurisdiction) to test all situations to ascertain whether a duty of care is owed, and preferred the liberal view of existing precedent. However, a duty of care in the clinical decision-making context is not necessarily obviously owed by analogy by the SDC to the patient. Historically, the clinician would have made their decision alone, but the introduction of the AIS by the SDC denotes a new and additional actor in the decision-making space which is cited specifically to influence the actions of the clinician in a way that is without current analogy (*Darnley v. Croydon Health Services NHS Trust)*. Ergo, this is what makes the situation 'novel'. Whilst a court would be unlikely to conclude that no duty of care is owed by an SDC to a patient, it might not be entirely obvious how that duty of care may be recognised due to the actions of the clinician being more prominent. Yet, all of this is speculative; we do not know if the courts will treat this situation as a novel duty of care as this scenario has not yet been tested. Because of this uncertainly there is value in considering how the *Caparo* criteria may be applied, in case the courts did recognise that the SDC owed a novel duty of care to a patient.

The *Caparo* criteria, noted in *Caparo Industries v. Dickman,* provide guidance as to where the court may be persuaded to broaden the scope of negligence liability and determine if a duty of care exists. They ask: (1) whether the harm resulting from the defendant's conduct is reasonably foreseeable; (2) whether the relationship between the parties is sufficiently proximate in law; and (3) if it is 'fair, just, and reasonable' to impose a duty of care on the defendant. The following discussion sketches out the potential liability of the SDC towards the patient using this framework.

Foreseeability

The first *Caparo* component of foreseeability seeks to determine if it is reasonable for an ordinary, reasonable person (*Glasgow Corporation v. Muir*) to foresee that their careless actions may result in harm to others (*Page v. Smith*).

SDCs cannot reasonably say that they did (or could) not foresee that patients would be affected by their system's outputs, given that their system was specifically designed to advise on patient care. The relationship between the patient and the AIS is undiluted even if mediated by a clinician, as the patient's data are processed by the system directly. A skilled professional in possession of knowledge who wilfully ignores it is '*prima facie* negligent' when it is foreseeable that an injury could occur (Rowland and Rowland, 1993, p.240). This extends to technical professionals in the employ of the SDC.

The SDC could argue that intermittent episodes of malfunctioning are not unprecedented (Petricek, 2017). AISs operate according to observations latent in the data, rather than the experiential and applied knowledge base which clinicians possess. SDCs are not to be expected to have medical expertise, but a negligence claim might not be unreasonable if an SDC has released software designed for use in clinical environments.

An alternative foreseeability scenario also exists. Should the AIS have a high frequency of accurate outputs desirable for clinical decision-making, the clinician may find their attention wanes when monitoring the effect of the system's recommendations as applied to patients. This phenomenon can be described as an 'atrophy of vigilance' (Freudenberg, 1992, p.19). When the device works consistently without issue, the user might begin to trust it uncritically, even when they know that they should not. This is of course in conflict with the clinician's duty of care, which would compel them to pay attention when using any kind of tool, but atrophy of vigilance is certainly foreseeable.

If atrophy of vigilance can result in death or personal injury in the above safety critical situations, then death or injury might be considered a foreseeable consequence of AISs which provide rapid solutions and aim to lessen the cognitive burden for system users. Trusting an AIS which is usually reliable is a foreseeable consequence in the lifespan of an AIS; thus, it could be posited that there is an open door for courts to find that SDCs owe a duty of care if harm eventuates due to a clinician's foreseeable loss of attention while using their system. Were the courts to agree, a positive obligation from SDCs could be required to take human factors into account and design the system to ensure safety in the high-pressured clinical environment. Absence of such holistic design features could be taken as evidence of a negligent omission.

Proximity

The second *Caparo* component of proximity asks if the defendant is positioned sufficiently close to the claimant for that relationship be considered a legal relationship (*Hedley Byrne & Co Ltd v. Heller & Partners Ltd*). Where there is a legal relationship, actors need to take precautions to prevent harm, rather than those who may be harmed needing to take precautions against being harmed (*Stovin v. Wise*).

To avoid liability claims, the SDC would attempt to position the clinician as the primary guardian of safety. Proximity, as understood in the tort of negligence, need not be solely geographical, instead it describes:

> *"such close and direct relations that the act complained of directly affects a person whom the person alleged to be bound to take care would know would be directly affected by his careless act."*
>
> *(Donoghue v. Stevenson, 581)*

Within the clinical environment, this duty of care towards patients is held by both medical and non-medical staff as found in Darnley (*Darnley v. Croydon Health Services NHS Trust,* 2018).

It is conceivable that a similar relationship might be considered to exist between an SDC and a patient if the SDC provided their system to be used specifically to aid patient care. The SDC's position might be considered analogous to other professional roles within the clinical environment, for example radiologists or haematologists. These specialised clinicians can advise their colleagues on an interpretation of a medical image or on the appropriateness of administering a unit of blood product to a specific patient; and they frequently do this without ever having met the patient themselves. Specialised clinicians who are not directly at a patient's bedside are expected to follow their code of professional conduct, as do their fellow clinical colleagues (GMC, 2020; HCPC, 2016; NMC, 2018). In law, specialists are subject to the same duty of care as generalists, though the expected *standard* of care may differ.

This matrix can be applied to the SDC that develops an AIS which accepts an input, processes it according to a specialised algorithm and then generates a recommendation which purports to embody clinical expertise. However, the SDC itself is a third party detached from the bedside. Thus, an AIS may make recommendations which prompt a user's actions, but this does not lessen a user's duty to take care that their actions shall not harm another. The participation of the SDC clearly complicates an otherwise straightforward assessment of where the duty of care lies.

'Fair, just and reasonable'

It might be argued both that it would be reasonably foreseeable to an SDC that their AIS would affect patients and that an SDC is sufficiently proximate to a patient. This might not, *however*, not be enough to award an SDC with a duty of care due to matters of legal policy regarding fairness, justness, and reasonableness, which I shall explain now.

Where a duty of care has been denied by the courts on the basis of the second component of proximity, it is argued by Witting (2005) that the reasoning for the denial is often based in Caparo's third component of fairness, justness, and reasonableness:

> *"the concept of proximity 'masks' the real policy-based reasons for arriving at particular duty determinations and that it is preferable that these reasons be expressed openly"*
>
> *Witting, 2005, p.34.*

The imposition of a duty of care by the courts is a normative decision (Witting, 2005) and the imposition of that duty needs to be fair, just, and reasonable (*Caparo Industries v. Dickman*). The courts seem to have approached this as a matter of legal policy (Witting, 2005); legal policy being recognised by Lord Millett in *McFarlane v. Tayside Health Board* as "our more or less inadequately expressed ideas of what justice demands" (Keeton *et al*, 1984, p.264). This means that concern lies with the question of whether a legal relationship between two parties *should* be recognised based on their proximity before their interaction resulted in the damage (Witting, 2005). This is because a duty recognised will not only affect the particular case that is before the court, but also those who are not before the court – i.e., those persons who may be future claimants or defendants (Witting, 2005). Witting (2005) offers various legal policies which are considered by the courts, for example:

- Floodgates: A flood of claims could result from a new duty of care being recognised:
    - Cases are carefully scrutinised by the courts to carefully manage issues which may allow a flood of claims to arise, such as *McLoughlin v. O'Brian* where it was recognised that a duty of care is owed to those who own property or are near the scene of an

accident, rather than those persons who are miles away from the accident and had not heard about it until hours after the event.

- Indeterminacy: The financial effect that a duty rule may have, the large number of people that the duty rule would affect, and the fairness of the distribution of that duty rule; i.e., would it be fair to hold defendants liable to an unknown number of claim/ants?
    - Witting (2005, p.40) claims that "defendants should be able to weigh up the costs of taking precautions against the possible size of claims that could be made against them; that they should be able to predict the number of persons that their negligence might affect."
- Loss spreading: The House of Lords indicated in *Smith v. Eric Bush* that liability can be covered via professional insurance.

Witting (2005) recognises various problems with these policies. For example, a policy of denial when concerned with floodgates and indeterminacy may result in victims of negligence being denied compensation. Also, a policy for loss spreading leads to the problem that those who are harmed are treated more favourably by those actors who have insurance than those who do not. Decisions that are based on such legal policies are at risk of error when the scope of the policy considerations are so wide; they can potentially lose relevance to the claimant and defendant's dispute and "less is the likelihood that they will speak unequivocally in favour of or against duty" (Witting, 2005, p.42).

Rather than courts openly denying a duty on policies such as those just described, Witting (2005) posits that duties are determined by proximity factors concerned with each party's relationship to each other. This has the benefit of tests for proximity providing greater consistency in legal decision-making, which contrasts to the uncertainty offered by decision-making based in policy; where the courts cannot determine how parties will react in the future based on their decisions made using policies as described above (Witting, 2005). "The law cannot be remade for every case" (Hobhouse LJ in *Perret v. Collins*) and where proximity is used rather than policy, some degree of certainty can be achieved.

The analysis that is offered in this work is limited as this thesis is working with scenarios which have not yet taken place. Simply, we don't know what the courts will decide until a case is decided as it is impossible to speculate what, if any, legal policy may be applied to the scenario to which this thesis is concerned. Whilst Witting (2005) is concerned about proximity factors, the court may ultimately accept that it *is* fair just and reasonable for a duty for care to be awarded to an SDC, but still for a claim to fail in the context of the causation stage of a negligence claim – this point of potential failure will be discussed later in this chapter.

Aside from the above discussion regarding legal policy, further analysis of the component of fairness, justness, and reasonableness is offered below.

Firstly, there is no general duty of care to prevent damage being inflicted by third parties (*Smith v. Littlewoods Organisation Ltd)*. There is more to be said about this in the context of clinical care, as per Lord Sumption's judgment in *Woodland v. Swimming Teachers Association*. Here a non-delegable duty of care is identified as owed between an employer and their employee, a school and its student, and a hospital and its patient; i.e., that a non-delegable duty of care is owed by the hospital to the patient when a patient suffers harm due to the actions of a third party contracted by the hospital. However, this is discussed by Beuermann (2017, p.25) who notes that "it is relatively unusual in tort law for one person to be held strictly liable for the wrongdoing of another." Beuermann reports that there have been few cases which consider the non-delegable duty of care of a hospital to a patient, and that of those cases that have been heard, the negligence considered was that of a hospital employee or the hospital itself rather than a third party. *M v. Calderdale & Kirklees HA* has been the single case heard by a county court judge where a patient successfully sued for non-delegable duty of care. Here a health authority's community health centre referred a patient to a hospital in the local area for an abortion which was performed negligently. *M v. Calderdale & Kirklees HA* was commented on in *A v. Ministry of Defence* by Lord Phillips of Worth Matravers MR as a finding not representative of the current state of English law. This was because *M v. Calderdale & Kirklees HA* was based on *Gold v. Essex County Council* and *Cassidy v Ministry of Health* where, in both cases, it was the hospital themselves that carried out the treatment rather than third parties. Beuermann (2017) concludes that Lord Sumption erred when he included such hospital relationships in non-delegable duty of care and, rather, the hospital-patient relationship provides only that an opportunity for wrongdoing may occur and, as per *Bazley v. Curry*, it is not enough for a defendant to merely provide an opportunity for wrongdoing to take place.

The introduction of the AIS by the SDC is unique in that the AIS is software and that the SDC is not physically present or directly interacting with the patient comparably to how other third-party actors might. To influence a patient's outcomes, the AIS would need to be used by a clinician, and that clinician would likely be employed by the hospital to provide care for patients. If the AIS is unable to dispense treatments without the cooperation of a clinician, this is a reason for denying the extension of the duty of care to encompass SDCs. But the AIS has been designed as a tool with the specific purpose of influencing the actions of clinicians, which will directly affect the health status of the patient. The general rule is that there is no duty of care for persons to prevent harm being caused by third parties, yet a 'special relationship' may be possible, which would override the conduct of others; as in, for example, the recent case of *ABC v. St George's Healthcare NHS Trust*. Here, the Court of

Appeal found that a claim based on a ‘special relationship’ between an NHS trust and a patient’s unborn grandchild was sufficient to be arguable at trial (Mackenzie, 2017) and was recently confirmed on the facts in the High Court. Thus, it is not inconceivable that the courts might consider that there are grounds to argue that a ‘special relationship’ exists between the patient and the SDC, even though there is no case law to illustrate a duty of care between the SDC and the patient at present. If the system has been deployed with that specific purpose, it could be argued that it is unreasonable that the SDC’s legal responsibility for the effect of the system is negated by the clinician using it.

Secondly, an SDC may argue that an AIS is no substitute for skilled, clinical decision-making and that the law only imposes a standard of care commensurate with the level of specialism which the defendant holds themselves out as possessing.[20] If an SDC wished to transfer legal responsibility for using the AIS to the clinical user, they would therefore present it as merely ‘assisting’ clinicians in their decision-making, but would not claim that it is of the standard to substitute a clinician in the speciality that the system advises in. Based on this premise, it might be argued that it is unfair and unreasonable to assign negligence liability to the SDC which created and supplied that system.

A clinician would, however, be unlikely to consider using an AIS that they believed would make recommendations which are inferior to their own calculations; to do so would be illogical, counterproductive and would expose a patient to needless risk. In the example of IBM’s Watson for Oncology, the system is portrayed by the SDC as possessing the expertise of ‘leading oncologists’ (IBM, undated). The reported use of this system by UB Songdo hospital (Ross and Swetlitz, 2017) is concerning if they had been informed that Watson makes recommendations rather than clinical decisions, thus resulting in the risk of system use being borne by its clinical user. UB Songdo Hospital must have been convinced that Watson was sufficiently sophisticated to perform that role; in the absence of universally accepted standards, individual institutions must assess each AIS’s adequacy on a system-by-system basis.

As mentioned above, a non-delegable duty of care has not been found to exist involving patients, hospitals and third parties, but that does not mean that the duty of care between the patient and the third party does not exist. This is exemplified in *A v. Ministry of Defence*. Here, the Ministry of Defence had changed its arrangements for those in their service and their dependents when stationed in Germany. Rather than healthcare being provided by British Military Hospitals, arrangements were made for treatment to be obtained from German hospitals. An obstetrician acted negligently which

[20] See *Philips v. William Whiteley Ltd* where the court held that a jeweller did not claim to be of the same standard as a surgeon and that the appropriate standard of care was of a reasonable jeweller undertaking ear piercing and not the reasonable surgeon.

resulted in a child, ‘A’, being born with severe brain damage and resultant cerebral palsy. The Ministry of Defence was found to not hold a non-delegable duty of care for care given by the German Gilead Krankenhaus hospital, but this did not mean that there was no recognised duty of care. German law has the facts of negligence claims investigated by expert commissions, in this case they found that there had been medical malpractice, and the hospital’s underwriters admitted liability. Additionally, the English experts instructed by the parties in this case also agreed that clinical negligence had occurred.[21]

The duty of care between the patient and a third party is also exemplified in *Farraj v. King’s Healthcare NHS Trust*. Kings had sent a sample of foetal tissue from the patient (residing in Jordan) to their London hospital. The sample was tested by a third party laboratory which confirmed that a sample of foetal tissue was free from a genetic condition; the test result was wrong, and the child born was found to have the condition. Kings was not found to be strictly liable for the actions of their third party contracted laboratory, but the failure of the laboratory to communicate to Kings the technician’s doubts that the sample contained foetal tissue was negligent.

A preference for extending the duty of care to the SDC might exist if there is sufficient foreseeability of harm and a convincing relationship of proximity between the SDC and the patient. This could be too challenging for the courts to accept as the clinician is acting at the bedside whereas the SDC is not. Yet, the SDC is acting for the benefit of the patient notwithstanding that the clinician is the intended recipient of the AIS’s recommendations, and there is an interesting analogy here with the case of *Smith v. Eric S Bush*. The defendant surveyor was instructed by a building society to report on a property being purchased by the claimant. In addition to the surveyor owing a duty of care to the building society to carry out the inspection with due skill and care, the surveyor owed a duty to the third-party purchaser on the basis that the defendant would know there was an ‘overwhelming probability’ that the purchaser would also rely on that report. Similarly, there is an ‘overwhelming probability’ that if the recommendation given to the clinician is negligent, its impact will be felt by the patient. This analogy reinforces the argument for extending the duty of care to the remote patient. Similarly, if the AIS is a ‘black box’, meaning that the clinician cannot interrogate its reasoning, it might be ‘fair, just and reasonable’ that the SDC should owe a duty of care to the patient too, independently of any negligent conduct on the part of the clinician.

[21] If the family had claimed in Germany, the damages awarded would have been no less than that offered in English proceedings. However, the family of A claimed against the Ministry of Defence as they wished to take action in England where they all now live and where the losses that the damage caused would be experienced.

The *Caparo* principles were not intended to cover all future scenarios, but to guide the court's consideration of novel fact situations. Therefore, an argument is made in favour of imposing a duty on the SDC in addition to any claims against the clinician so that risks are appropriately managed and contained.

Breach of Duty

As previously mentioned, clinical conduct is not considered negligent if it satisfies the *Bolam-Bolitho* calibration of the standard of care. If a duty of care is found to exist for the SDC, their system would need to be in accordance with a responsible body of opinion and the expert evidence supporting the defendant's conduct or decision must 'withstand logical analysis'. The following considers how this calibration of the standard of care could apply to the SDC.

The AIS might advise contrary to the expectations of the clinician, but that might not necessarily be negligent as long as the system's recommendations are safe and therapeutic for the patient (as per *Luxmore-May v. Messenger May Baverstock)*. In safety critical areas, it is desirable that an SDC strives to follow, and even surpass, the relevant standards for medical devices (Rowland and Rowland, 1993). Healthcare is recognised as a safety critical area; any incorrect advice from an AIS has the potential to harm the target patient.

An SDC could demonstrate discharging their duty of care through observation of the standards and codes of practice relevant to their profession (Rowland and Rowland, 1993). The notion of observation of standards is supported in comparative case law from New Zealand. In *Bevan Investments v. Blackhall & Struthers*, the defendant engineer had achieved the expected standard, and so it was held that upon rational analysis the court could conclude that no negligence had occurred.

Specific to litigation regarding healthcare contexts in the UK, Heywood notes multiple cases [22] which indicate that "there is an ever-increasing body of case law revealing that where guidelines are introduced in evidence, not infrequently, matters are resolved in favour of defendants" (Heywood, 2021, p.65). Yet Heywood observes that, whilst guidelines offered by reputable bodies provide a benchmark by which the courts may evaluate conduct, a practitioner operating inside or even outside of a guideline does not automatically render them either exculpatory or inculpatory. Rather, the court should afford more attention to questioning the suitability of the guideline "in a *Bolitho* sense" (Heywood, 2021, p.65). In *Jones v. Conwy and Denbighshire NHS Trust,* the patient attended the hospital with a sinus infection and orbital cellulitis. Clark J identified that there was a lack of a clear

[22] *Price v Cwm Taf University Health Board, Barry v Cardiff and Vale University Local Health Board, Rich v Hull and East Yorkshire Hospitals NHS Trust, C v North Cumbria University Hospitals NHS Trust, Cowley v Cheshire and Merseyside Strategic Health Authority*

consensus rule in the medical literature that a CT scan should be done when orbital cellulitis is suspected with accompanying uncertainty about whether the source of the cellulitis is pre-septal. Whilst Clark J was apparently convinced that following the clinical guidance regarding CT scanning in this scenario was equivalent to responsible practice, Heywood (2021) is critical. Heywood (2021, p.65) proposes that guidelines need to be subjected to a more rigorous '*Bolitho*-justifiable' examination by the courts; ergo that more attention needs to be given to the suitability of the guideline when it is applied. In the circumstances of *Jones v. Conwy and Denbighshire NHS Trust*, that would have meant that it would have been more logical, responsible, and defendable for the CT scan to have been performed which would have allowed the intra-cranial pus to have been found.

Jones v Conwy and Denbighshire NHS Trust is an example of how not all guidelines are conclusive enough to be relied upon. The judgment of *C v North Cumbria University Hospitals NHS Trust* noted that those who act according to guidelines should be safe from negligence charges as they have not acted unreasonably. However, Mr Justice Green also noted that, even when guidelines have professional and regulatory approval, they may still be incomplete or incomprehensive. That being the case, the surrounding facts and circumstances will also be considered by the court, as per *C v North Cumbria University Hospitals NHS Trust.*

Comparatively, in the USA, *Daubert v. Merrill Dow Pharmaceuticals Inc* goes one step further; here strict objective requirements were set for scientific data to be used in cases; specifically, that the data presented must be reliable and relevant to the scenario being considered.

A claimant may ask why a guideline was or was not used, and a practitioner may wish to explain why they had chosen to use that same guideline to demonstrate that they have complied with the standard of care (Samanta *et al,* 2006). To aid judges in cases where guidelines need to be evaluated, Samanta *et al* (2006) propose four questions that may be considered:

1. Is the guideline *Bolam*-defensible? – if the conduct was not what a reasonable practitioner would have done, then the standard of care has not been met.
2. Is the guideline *Bolam*-justifiable? – does the body of opinion relied upon have a logical basis?
3. Is the guideline *Daubert*-valid? – is it reliable and relevant?
4. How does the guideline apply to this scenario? – how were the guidelines applied by the practitioner?

The use of these questions allows judicial decision-making to examine the use of evidence-based guidelines and explore via testimony how the practitioner's judgement was exercised (Samanta *et al,* 2006).

The problem with novel clinical applications (e.g. AISs), though, is that comprehensive and authoritative guidelines for practice are not always immediately available when those novel approaches are being developed or enter the healthcare environment.

As earlier noted, codes of conduct, practice guidance, and professional standards are in place for clinicians; however, technical standards are neither mandatory, enforceable, nor sufficiently advanced yet in the domain of AI-assisted medicine to provide specific, definitive guidance to SDCs.

As noted in chapter 4's literature review, codes of conduct and guidance for SDCs and technologists exist in many forms, both generally and specifically in the healthcare setting of England and Wales, yet these exist without teeth or unification. These need to be enforceable and healthcare-sector-specific, or even health-condition-specific, for SDCs before medical experts and the public might be reliably assured of the safety of the AI devices and algorithms that they produce.

The chief item of guidance regarding AISs in healthcare in England and Wales has been the Department of Health and Social Care's (2021b) Guide to Good Practice for Digital and Data-driven Health Technologies. Yet, whilst this guidance speaks of behaviours expected from developers and those deploying AIS, does not stipulate the obligations of these stakeholders to take care that their actions of creating and deploying an AIS will not harm the patients to whom their AIS shall affect. This contrasts with the clinical professions whose codes stipulate that registrants must act to preserve safety and are enforced by each profession's regulators.

UK clinicians are aided in their practice by organisations such as the National Institute of Health and Care Excellence (NICE) which publishes evidence-based guidance on clinical conditions and treatment pathways (NICE, 2020). NICE has started to create briefings about innovations which use AI, however, NICE has not yet evaluated AIS for use in treatment pathways. Instead, NHSX, a unit dedicated to digital transformation of the NHS, is reportedly working on standard setting in this area. This work is supported by the release of NICE's (2021) Evidence Standards Framework for Digital Health Technologies which describes the standards of evidence that should be available for a digital health technology to demonstrate its value to the UK's health service.

Whilst there is not a comprehensive body of knowledge from which the courts may draw comparisons at this moment in time, initiatives such as these are beginning to build that knowledge base and the standards of conduct which SDCs should adhere to for the courts to draw upon and apply tools such as Samanta *et al*'s (2006) aforementioned four questions to when considering cases in the future. Although the standard of care in negligence and the standards set out in the various sources of published guidance are not necessarily the same, as described above, courts often rely heavily on professional guidance to inform the standard the law should apply. For example, in *Montgomery v. Lanarkshire Health Board*, the Supreme Court expressed that doctors' duties to inform patients are closely aligned with General Medical Council guidance on the matter. The equivalent applied here could be that the SDC and the clinician ought to ensure that AISs developed and deployed ought to reach the standards stipulated by the guidance provided by authoritative sources, such as those aforementioned from NHSX, NICE, the Department of Health and Social Care, and the clinical professional regulators.

If the AIS is thought to be so risky that its outputs need to be verified by a clinician, the court might find it hard to accept that a system's outputs are effective substitutes for the professional standards of clinical staff (and that it thereby does not satisfy the *Bolam-Bolitho* standard of care). AIS outputs are not the same as the evidence-based medicine upon which modern clinical practice is grounded. However, the clinician may not have the skills to appraise the AIS which has been offered to them. Additionally, there is no peer-reviewed evidence base which the clinician may draw upon to ensure that using the AIS would result in an improvement in care. These two issues create a problem of non-translatability from the body of knowledge offered by the AIS to the clinical environment. When presenting an AIS for appraisal and use by clinicians, an SDC would need to be circumspect in their conduct to avoid misrepresentation of their product as equivalent to the work of 'the ordinary skilled man exercising and professing to have that special skill' (*Bolam v. Friern Hospital Management Committee)*. Advertising exaggerations regarding AIS's may shape liability under section 3(2) of the Consumer Protection Act and give rise to claims under contract law for misrepresentation, but the crux of the matter is that the conduct of SDCs as interdisciplinary specialists could be used as evidence that the SDCs ought to have known that defects in their system would be highly likely to inflict harm, putting them in breach of their duty of care for not taking active steps to avoid such harm.

Causation

This chapter has already touched upon causation in the earlier discussion of factual causation. Broadly speaking, the two major views are that either (a) 'but for' the malfunctioning AIS the clinician would not have been prompted to apply the harmful recommendation to the patient or (b) that a clinician openly acknowledging that they do not have the skills of a specialist clinician does not relieve them of a negligence claim should harm eventuate when they have used an AIS which makes recommendations in that specialist area. However, if the court holds that the SDC shares responsibility with the clinician for the patient's harm, the court must consider how multiple causes may be related.

When multiple factors and/or multiple tortfeasors are determined as having caused injury, the courts use different methods to determine the cause responsible for the injury. The tortfeasors' ‘independency’ approach finds that each factor had caused a single damage independently of other factors (Noee *et al*, 2016, p.221). By contrast, the tortfeasors’ ‘separate impact’ approach differs from tortfeasor’s independency as multiple injuries are attributed to separate tortfeasors (Noee et al, 2016, p.224). Tortfeasors’ contribution exists when multiple tortfeasors all acted and the sum of the acts are treated as creating the damage (Noee et al, 2016). The use of these three approaches can make it challenging to predict the outcome of a claim (Noee et al, 2016).

Causation is considered in two stages in terms of ‘factual causation’ and ‘legal causation’; each shall now be discussed in turn.

Factual causation determines if there is an uninterrupted sequence of cause and effect events which link the defendant’s actions to the harm that the claimant has suffered (Hodgson, 2008). The patient may be able to satisfy the causation element of a claim by proving that the negligence ‘materially contributed to the damage’, and if successful, the defendant may be liable in full for the whole of the damage, notwithstanding that they have only been proved to contribute to it (*Bonnington Castings v. Wardlaw; Bailey v. Ministry of Defence; John v. Central Manchester and Manchester Children’s Hospital University Hospitals NHS Trust)*. However, where the damage suffered is treated as ‘indivisible’, the parties might argue that liability on the basis of being responsible for a ‘material increase of risk’ is the more appropriate claim (*Fairchild v. Glenhaven Funeral Services Ltd)*. If argued, the material increase of the risk approach to causation might be applied where both the clinician and the AIS independently make the same error but as each confirms the other, it is impossible to prove which entity contributed more to the harm suffered by the patient. But, both aforementioned arguments depend on the SDC’s AIS being accepted as an authoritative source. If the recommendations given by an AIS is not accepted as such, it could be argued that the clinician should have made their own independent check on the appropriateness of an AIS’s recommendations.

However, there are two reasons why the *Fairchild v. Glenhaven Funeral Services Ltd* judgment might not have an application in this thesis’s AIS scenario. Firstly, *Fairchild* was specific to its scenario in the context of occupational asbestos exposure and mesothelioma. This was confirmed in the speeches in both *Fairchild* and post-*Fairchild* cases (e.g. *Barker v. Corus*, *Sienkiewicz v. Grief*), therefore making *Fairchild* non-applicable to the AIS scenario that this thesis considers.[23] Secondly, in cases involving AIS use, patients could be affected via issues such a diagnosis being delayed due to the clinician’s use of the AIS which may lead to their treatment being either delayed or being inappropriate for them.

[23] Even – and this is very much an imaginative argumentative stretch - if the AIS was being used to treat mesothelioma, the AIS would not have been the source of the asbestos causing the mesothelioma, therefore still not *Fairchild* specific.

This might inspire a claim for the loss of chance for a better outcome for their condition. Jones (2006) notes loss of chance in their discussion of *Gregg v. Scott* where the claim was for a loss of chance for an increased range of treatment and improved survival chances when a clinician had negligently diagnosed a lump in the patient's armpit as benign when it was malignant. Jones notes that (as per comments made by Lord Hoffmann in *Barker v. Corus*) allowing loss of chance claims such as *Gregg v. Scott* would have allowed *Fairchild* to be extended to all medical negligence cases. This makes it unlikely, but not impossible, for a 'material increase of the risk' claim to be applied to an AIS scenario. Yet, despite these two reasons, future cases are difficult to predict and could potentially go either way as it is illogical to have rules for a single disease when there are other analogous cases.

If a court were to approve of the 'material increase of risk' approach to proving causation, it is necessary to outline the limitations on obtaining compensation that this entails. According to *Barker v. Corus UK Ltd*, liability for a 'material increase of risk' is to be apportioned severally, meaning that a patient claimant or a clinician co-defendant might not be able to claim the appropriate contribution to damages from the SDC. Several liability is recommended by the European Commission's Expert Group on Liability and New Technologies whose report directly considered liability for AISs. Yet, under a several liability regime, the claimant must succeed in their claim against each defendant separately to receive compensation in full for the negligently inflicted injury, which adds additional unnecessary costs and minimises the prospect of obtaining full compensation for the harm suffered (*Barker v. Corus UK Ltd)*. This could discourage claimants from pursuing justice from an SDC and mean that the incentive for SDCs to fulfil their duty of care is weakened. A return of joint liability shall be argued later in this thesis, to enable patient access to a timely remedy.

In the scenario this thesis presents, an AIS is unable to inflict harm without the conduct of the clinician as an intermediary. It would therefore be advantageous to the SDC to argue that the conduct of the clinician is a new intervening act; *novus actus interveniens*. However, according to *Webb v. Barclays Bank plc and Portsmouth Hospitals*, if a clinician acts negligently and makes the claimant's injury worse during the course of treating them, whilst they may find themselves attributed liability to the damage resulting from their own actions,[24] that does not mean that the defendant's liability for the original injury ends (Jones, 2021). This is evident in medical negligence case law; Jones (2021) referred to *Webb v. Barclays Bank plc and Portsmouth Hospitals* and *Rahman v. Arearose* when summarising that medical negligence does not necessarily break the causal chain between the defendant's original

[24] In *Webb v Barclays Bank plc and Portsmouth Hospitals* the claimant had fallen due to the negligence of their employer. The surgeon treating her advised an above-knee amputation which was negligent. Liability for the damage attributable to the negligent amputation assessed at 25% to the employers and 75% to the Trust (tidily explained in Laing and McHale, 2017).

negligence causing the injury and the claimant's loss. As such, actions of third parties do not generally break the chain of causation unless the intervening conduct is so outrageously negligent that it would not be fair for the initial defendant to continue to carry responsibility for those later acts (*Spencer v. Wincanton*). Specifically to medical negligence case law, Hodgson (2008) notes that, as per *Webb*, only medical treatment that is grossly negligent through being a completely inappropriate response to the injury should result in the causal chain being broken between the claimant and the defendant. This is the point where a clinician is at risk of being held liable for harms that may eventuate due to the use of an AIS in their clinical decision-making.

In *Horton v. Evans* a pharmacist was held to have acted negligently due to not questioning a doctor's prescription which was eight times the patient's usual dosage of dexamethasone. This case could be seen as analogous to a clinician not questioning an AIS's recommendation. As such:

- if a clinician did not competently recognise that an AIS output would be inappropriate to the point of being harmful to the patient of whose care they were deciding,
- then subsequently followed that recommendation,
- and the patient came to harm as a result,
- the courts may find the clinician's action of following the AIS's recommendation as grossly negligent through being completely inappropriate (as per *Webb*).

Ergo, the clinician that follows inappropriate AIS recommendations without reasonably questioning them to ensure their appropriateness could find themselves liable, similarly to *Horton v. Evans*.

Thus, if acting solely on the basis of the defective AIS recommendation, the clinician may find themselves in the position of being causally responsible. In this case, the court may be minded to hold that these are two separate instances of negligent acts. Instead of the *Performance Cars* approach where the harm caused by the initial act is deemed to continue, here the situation is reversed. Under the *novus actus interveniens* doctrine, any 'extraordinary' latter conduct of the clinician could obliterate the defects latent in the AIS, which may remain undiscovered. If the spirit of *Bolam* is that the clinician would need to have achieved the standards of other professionals, then the clinician may have failed to achieve this standard if they had failed to identify that an AI's recommendation was inappropriate for a specific patient.

Does the clinician professing to their patient that they do not have the skills of a specialist clinician relieve them of the usual standard of care? Likely not when considering the objective standard of care as illustrated by *Nettleship v. Weston*; if 'the certainty of a general standard is preferable to the vagaries of a fluctuating standard' (*Nettleship v. Weston* at 707) then the standard expected of a clinician offering specialist treatment for a specific condition ought be determined of an acceptable standard as per a specialist clinician's practice, rather than of a non-specialist clinician offering specialist treatment of which they are not qualified. This principle was confirmed in the recent case of *FB v. Princess Alexandra Hospital NHS Trust* where it was found that, when considering liability, the standard of a doctor's care is not to be adjusted to take into account their limited experience.

Legal causation considers the extent by which the defendant should be required to pay for damages which their conduct has played a significant role in causing (Hodgson, 2008). Suppose the clinician and the AIS independently make identical errors. Specifically, if a defective AIS produced an erroneous output which could harm the patient if followed, is the effect of this error eliminated by the clinician's independent conduct? In *Performance Cars v. Abraham*, it was held that the first act may obliterate the second. One may be able to say that both actions were tortious and conclude that the effect of the first act (the negligent development of the AIS) continues in spite of subsequent negligence (*Performance Cars Ltd v. Abraham*; *Heil v. Rankin*). The SDC's negligence would take priority in this instance, because reference to the AIS would precede the clinician's conduct. However, this might unduly relieve clinicians from acting in the patient's best interests and raise a moral hazard, encouraging reckless behaviour. As such, this 'consecutive cause' approach may not provide the best means for balancing the responsibility of the clinical and technical parties in ensuring that the AIS is developed and used in a way that protects patient safety. Furthermore, it is questionable whether the Performance Cars doctrine is applicable here as the 'conduct' of the SDC, as expressed through the AIS tool, and that of the clinician are not truly independent of each other.

In the absence of a definitive pronouncement by the court on the causation issues raised in this chapter regarding the presence of the AIS, it is impossible to accurately predict how far causation principles will be central to the future development of the law. Nevertheless, this chapter has shown that the common law is not closed on these matters, meaning that there is a real possibility that SDCs may be liable in negligence for defective AISs in clinical settings.

Using volenti as a defence

The characterisation of IBM Watson for Oncology as a recommendation engine (Hengstler et al, 2016) suggests that the technical community may adopt a strategy of excluding their liability in negligence. This may result in an SDC requiring a clinical user to accept a contract reflecting this exclusion. However, it is not the clinician who might suffer harm from the use of the AI, it is the patient. When considering user contracts, the patient is not the user of the AIS, the clinician is; the patient's relationship with the AIS would be via the clinician using it, would likely be transitory, and related only to the immediate and specific clinical needs which the AIS is positioned to help with. As such, an individual patient would likely not have been directly involved in the selection and adoption of an AIS to the clinical environment – that process would have been a transaction between the SDC and the clinician, rather than the SDC, the clinician, and the patient. As the patient is likely not involved in the choice and deployment of the AIS used in their care, they would likely not be party to a contract between the SDC and the clinician for the use of an AIS. However, despite any pronouncements that a clinician retains ultimate responsibility for patient care, in law this is not the SDC's choice to make. A defendant cannot exclude liability for negligently caused death or personal injury (Unfair Contract Terms Act 1977, s 2(1)), so this tactic might prove to be an inadequate defence in a negligence claim brought by a patient.

Thus, a defendant SDC's option for a full defence might be limited to arguing that the patient had consented to the risk of an AIS being used in their clinician's decision-making which would thereby affect the delivery of care, i.e. *volenti non fit injuria*. In medical contexts, this principle contains the following elements: consent must be given voluntarily by the patient (*Smith v. Charles Baker & Sons*) and they must be of sound mind (*Kirkham v. Chief Constable of Greater Manchester*). The patient is also entitled to be provided with information about all the relevant factors so that they can formulate their decision (*Chester v. Afshar*).

Questions could be raised about whether it is possible for a patient to be sufficiently informed to voluntarily agree to the use of a 'black box' AIS, where its processes are largely unknowable and inscrutable in precise detail. It is reasonable that a patient would want to know if a system they were about to use might produce faulty outputs which might harm them before consenting to its use. A prudent patient may wish to demand that the SDCs quantify the risks before engaging with their products. With this knowledge, they may choose to not permit the use of an AIS in their care. On the other hand, if the patient did choose to permit use of the AIS knowing of the risk of potentially harmful recommendations being produced, they would be wise to consider a low threshold to seeking verification by a clinician prior to accepting the treatment which the AIS had recommended. *Volenti* in any case is a very unlikely defence in a medical negligence action; it is only rarely successful and in very limited circumstances (McHale and Laing, 2010).

The present state of the law does not account for the way that SDCs might perpetrate harms from a distance. In conventional negligence liability, the principles of proximity (with regard to the duty of care) and remoteness (in respect of the damage caused) are used to delimit the legal responsibility for acts or omissions to the parties that are most directly connected to the harm. This arises out of the notion that 'fault' ought to be the basis of liability (Cane and Goudkamp, 2018); the broader the definition of liability, the greater the likelihood that 'fault' is applied to parties whom society might otherwise perceive as morally innocent or whose conduct is justifiable and excusable.

Another basis for awarding damages is that it provides a deterrence signal to potential tortfeasors (Schwarcz and Siegelman, 2015). Ultimately, claimants would rather that their injuries had never occurred, and the threat of being forced to pay financial compensation is one way to encourage parties to consider their actions carefully and take measures to avoid causing harm. Closely related to this is the idea that liability promotes the efficient 'internalisation' of costs in risky activities (Faure and Partain, 2017, p.103). Therefore, the defendant is generally the party who could have avoided the harm at the least cost. Hence, compensation is a redistributive obligation that seeks to restore economic equilibrium among participants.

It is this latter view that reflects the position of each of the three key stakeholders. The clinician has a clear obligation to act in the best interests of their patient. The SDCs ought to design their system so that is 'fail safe'; minimising the risk of harm when defects occur and forestalling the 'atrophy of vigilance' phenomenon. Likewise, the patient could have taken steps to avoid harm by interrogating the safety aspects of the AIS which their clinician has chosen to use but was prevented from doing so due to the 'black box' character of these systems (thus their consent is vitiated).

Preparing should the worst happen

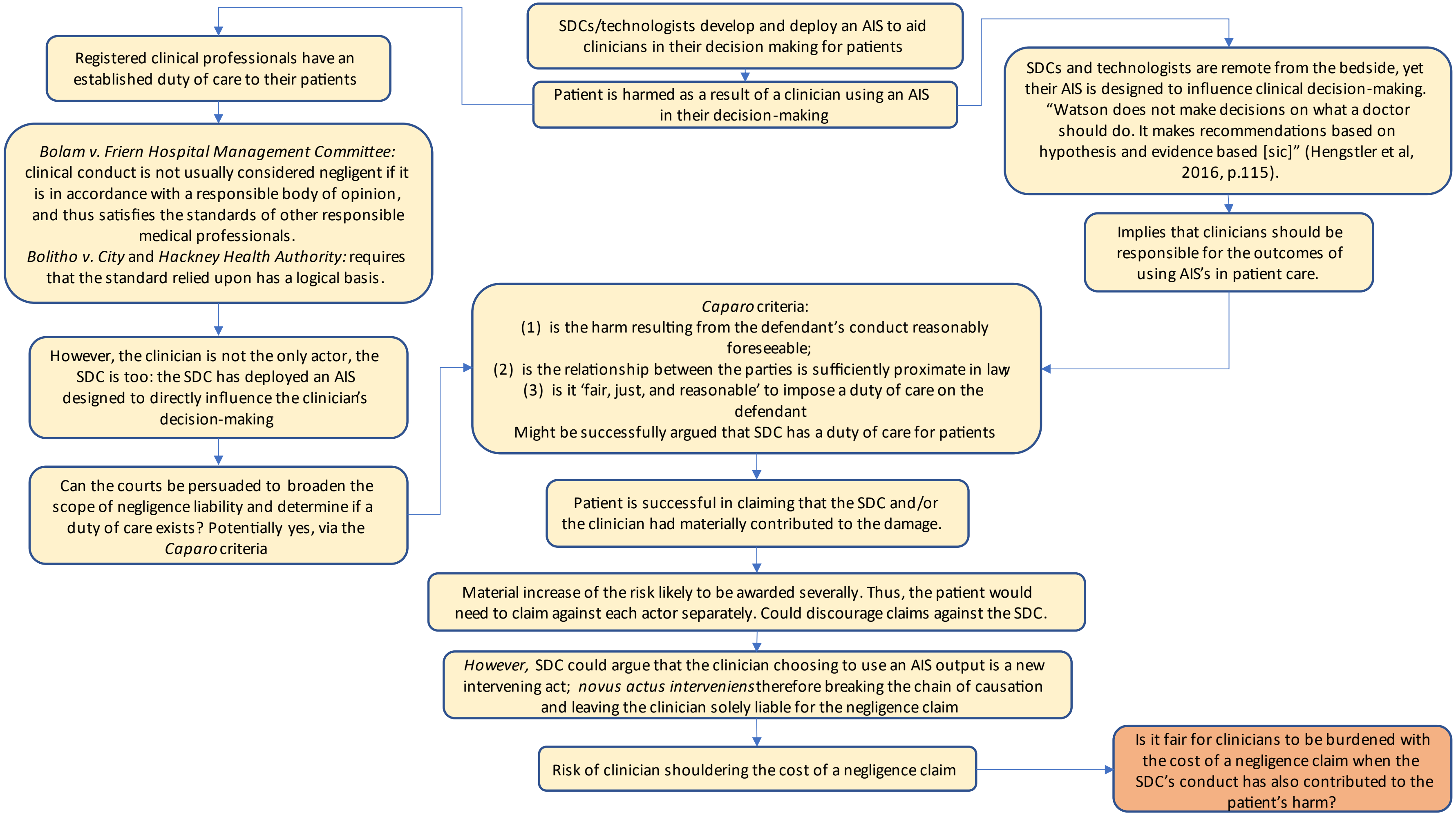
The patient is clearly deserving of compensation if they have suffered unnecessary harm, so the question becomes one of how the patient claimant's loss is best distributed and rectified. There are grounds to argue that liability for damages ought to be shared among the stakeholders in medical AI devices and algorithms, involving both the technical and medical teams as they jointly contribute to the overall risk of harm. Yet, there is a desire to innovate and provide an environment where beneficial technologies can be tested and deployed rapidly in front-line care.

The UK government proposes that liability insurance may help to balance the risk between actors and provide clear accountability among participants in sensitive sectors (Government Office for Science, 2016). If an actor can envisage a duty of care arising from their actions, insurance can help to defray some of the costs of engaging in risky activity. Indeed, in some instances, purchasing insurance to cover the minimum expected liability is compulsory (for example, the Road Traffic Act 1988, ss 143 and 145). But to insure against harms, there needs to be a clear model of how restitution for that loss is to be implemented.

Conclusion

The literature review asked how, according to current law in England and Wales, will legal liability be allocated between clinicians and SDCs when AISs are used in clinical decision-making? This chapter has outlined legal analysis to theorise how the tort of negligence can be applied in the scenario where a clinician uses an AIS's inappropriate recommendation and the patient comes to harm as a result. This chapter has been summarised into figure 6, which demonstrates the legal analysis presented (as opposed to a negligence decision tree).

Figure 6: Speculated negligence claim pathway



SDCs/technologists develop and deploy an AIS to aid clinicians in their decision making for patients

Patient is harmed as a result of a clinician using an AIS in their decision-making

Registered clinical professionals have an established duty of care to their patients

Bolam v. Friern Hospital Management Committee: clinical conduct is not usually considered negligent if it is in accordance with a responsible body of opinion, and thus satisfies the standards of other responsible medical professionals. *Bolitho v. City* and *Hackney Health Authority:* requires that the standard relied upon has a logical basis.

However, the clinician is not the only actor, the SDC is too: the SDC has deployed an AIS designed to directly influence the clinician's decision-making

Can the courts be persuaded to broaden the scope of negligence liability and determine if a duty of care exists? Potentially yes, via the *Caparo* criteria

SDCs and technologists are remote from the bedside, yet their AIS is designed to influence clinical decision-making. "Watson does not make decisions on what a doctor should do. It makes recommendations based on hypothesis and evidence based [sic]" (Hengstler et al, 2016, p.115).

Implies that clinicians should be responsible for the outcomes of using AIS's in patient care.

Caparo criteria:
(1) is the harm resulting from the defendant's conduct reasonably foreseeable;
(2) is the relationship between the parties is sufficiently proximate in law;
(3) is it 'fair, just, and reasonable' to impose a duty of care on the defendant
Might be successfully argued that SDC has a duty of care for patients

Patient is successful in claiming that the SDC and/or the clinician had materially contributed to the damage.

Material increase of the risk likely to be awarded severally. Thus, the patient would need to claim against each actor separately. Could discourage claims against the SDC.

However, SDC could argue that the clinician choosing to use an AIS output is a new intervening act; *novus actus interveniens* therefore breaking the chain of causation and leaving the clinician solely liable for the negligence claim

Risk of clinician shouldering the cost of a negligence claim

Is it fair for clinicians to be burdened with the cost of a negligence claim when the SDC's conduct has also contributed to the patient's harm?

The law addressing liability in the use of artificial intelligence for clinical decision-making is complex and ill-defined, especially for non-legally minded stakeholders. The current state of affairs is unfair to patients who may be harmed due to the use of AIS, because it is difficult to assess with any certainty which claims would succeed or fail in this area as case law on the points raised in this article is non-existent and existing legal frameworks might not be desirable (House of Lords: Select Committee on Artificial Intelligence, 2018). Clinicians are at risk of shouldering the burden of a negligence claim, even though the SDC has designed an AIS to directly influence the decision-making of the clinician. *Novus actus interveniens* seems to offer protection to the SDC while leaving the clinician vulnerable to negligence claims; this seems unfair to clinicians.

Whatever shape it ultimately takes, a model for restitution would need to be legally suitable (explored above) but also fair to all stakeholders; and this fairness requirement leads us to ask if ethical responsibility can be determined for the consequences of the use of AIS in clinical decision-making. In response to this question, the next chapter will examine ethical responsibility as applied to this thesis' scenario. The outcome of that analysis will serve to inform a fair model of restitution.

Chapter 6: How does ethical theory inform the allocation of responsibility for the use of AIS in clinical decision-making?

This chapter seeks to conduct an ethical analysis of the scenario, described above, of a clinician using an AIS to inform their decision-making. This will then be used to draw conclusions which may serve to guide the ongoing actions of those who choose to develop and/or use AISs in this context, both at an individual ethical level as well as at a level which may influence the community moral standard.

As just discussed in the preceding chapter, due to *novus actus interveniens*, current negligence law seems likely to put legal responsibility upon the clinician in the event of patient harm should an AIS be used in clinical decision-making. This is due to the clinician's proximity to a patient (by being at the bedside) and their choice to use an AIS's outputs. The SDC is not physically present to supervise the clinician who is using their AIS and cannot intervene if their AIS dispenses an incorrect output or is used incorrectly. SDCs have designed their AISs in a way that simultaneously can directly influence clinical decision-making but writes them out of direct contact with a patient's clinical activities, taking advantage of the clinician's traditional proximity (it is impossible to know if this arrangement has been intentional or unintentional). Here, the clinician, positioned as the final decision maker, is used as the safeguard - or 'moral crumple zone' (Elish, 2019, p.40) - between the AIS and the patient. Indeed, the literature review found authors claiming that an AIS's user is responsible for evaluating its outputs before using them (Pouloudi and Magoulas, 2000) and that clinicians using AISs should maintain responsibility for outcomes (Whitby, 2015, Sukel 2017b). But the clinician's decision-making processes are being knowingly and deliberately influenced by the SDC's AIS; and this allows the SDC to benefit from their AIS being used, but without having to bear any legal liability should the risk eventuate. However, the literature review also found authors supporting shared responsibility between stakeholders (Pouloudi and Magoulas, 2000; Whitby, 2015) and interdisciplinary collaboration (Pouloudi and Magoulas, 2000). Yet, my legal analysis showed that, whilst there is scope for an SDC to also be held negligent in the future, due to a lack of legal precedent and the aforementioned *novus actus interveniens*, legal responsibility in a negligence claim likely remains with the clinical user for now. The legal analysis has provided a position which might be later adopted by the courts, but this speculative legal position might not be ethically justified. Thus, this chapter asks how can ethical responsibility for the consequences of the use of AIS in clinical decision-making be determined and allocated?

This chapter considers this scenario through an ethical lens using ethical theories to consider where responsibility should be allocated between the SDCs who create AISs and the clinicians who use them. I use this analysis to build argument, alluded to above, that there is scope for SDCs to carry responsibility and to argue that it is not just clinical users who should be responsible for the outcomes of using an AIS. Theories of justice, causation, and personal moral responsibility shall be described and applied to support this claim. I shall a) establish how we may allocate moral responsibility, then b) outline how the legal position describe above might be challenged from an ethical perspective.

A note on ethics and morals

Ethics and morals are often used interchangeably in bioethics, but they are two distinct terms. For the purposes of this thesis, morality is a term used "normatively to refer to a code of conduct that, given specified conditions, would be put forward by all rational people." (Gert, 2020). This definition is applied on a community-wide basis rather than on an individual level. Conversely, ethical theory is defined by Gert (2020) as the guide which an individual adopts to determine their own life. This guide must be viewed as a proper guide for others as well as for themselves, and as such ethics is often discussed as a community wide rather than personal viewpoint. Regardless of the specifics of the definitions, any code of conduct can be personal, and may also be in alignment with behaviour which is accepted by society generally. This thesis engages with issues which affect individual stakeholders, but the provision of healthcare is an issue which also affects wider society, especially affecting patients. As ethics and morality are used alike, it is difficult to consistently use one term or another in this chapter. For this reason, when I cite an author's works, I have employed their use of the terms ethics and morals and applied those to my analysis, but I allow those terms to be interchangeable. In this way I preserve an author's chosen words and attempt to avoid the trap of mistakenly deriving an incorrect meaning from their materials.

Why do ethics and morality matter to the conduct of law?

Ethics and law exist as separate disciplines which often examine similar issues (albeit with differing approaches). A fully comprehensive discussion of this subject is outside of the scope of this thesis, but I do recognise that there are multiple views and tensions.

Consideration of ethics and morality allows people to think about why a position may be right or wrong. The ability to rationally think through a course of action might encourage people to act in a way that is accepted as correct by others in their society. In this way, cooperation within a society is eased when ethical beliefs are shared. Giving rational arguments for our moral beliefs allows others to understand them and, perhaps, to come to share them. Understanding and shared belief allow society to function peacefully and cooperatively. Indeed, ethics is recognised by Bryson (2019) as "the means by which a society maintains itself." However, ethics may offer only guidance for society. On its own, a code of community morality cannot compel all people to act in a particular manner;

ultimately, ethics has no 'teeth'. This differs from law, which uses coercive measures to compel people (e.g., a sanction, such as a fine or jail) to act accordingly.

Finnis (2001, p.18) describes natural law as the identification of "conditions and principles of practical and right-mindedness, of good and proper order among persons, and in individual conduct." A natural lawyer's position would claim that law and morality have a connection; this is contrasted with the legal positivist approach which would claim that there is no connection (Greenawalt, 1998). Brownsword (2008, p.12) highlights that the understanding of the relationship between ethics and law is a longstanding jurisprudential question; do we accept that there is a connection between law and ethics because law is a "sub-species of moral reason"; or should we follow the view of legal positivism and deny that link? Ethics and law do cover similar ground. They both ask, "what is my duty" and "to whom is that duty owed?"; but ethics differs from law by asking "what should we do?" rather than "what can we do?" (Sullivan and Reynolds, 1998, p.620). Holmes (1918) notes that when the rules of behaviour can be enforced by others, an actor is encouraged to behave in a particular fashion. Thus, when punishment can be predicted, coercive potential distinguishes ethics from law.[25]

In the practical application of law, courts have described themselves as one "of law, not of morals" (Lord Justice Ward in *Re A (Children) (Conjoined Twins: Surgical Separation*, p.4) which indicates rejection of ethics in practical application of law. Conversely, according to Huxtable, *"*…just as it can be difficult to define law in a way that does not beg moral questions, so too it can be hard to define bioethics without some reference to law." (Huxtable, 2016, p.96-97).

Huxtable and Ost (2017) note that narrative approaches are used to explore, give meaning to, and interpret human events. However, a narrative approach does not always translate well into law; court decisions may appear illogical when ethical problems are interpreted into and resolved by legal structures. Thus, the legal interpretation might conflict with the ethical interpretation of the story being examined. Even if a stakeholder's action is technically legal, it may not be deemed ethical;[26] therefore, the individual who has suffered an ethical harm may have no legal redress. Just because a body of law has developed in any given area does not mean that it must be accepted as the only valid argument and solution to a given problem.

[25] "…for legal purposes a right is only the hypostasis of a prophecy - the imagination of a substance supporting the fact that the public force will be brought to bear upon those who do things said to contravene it" (Holmes, 1918)

[26] A completely unrelated example here could be the controversy between the legal application of the death penalty despite parties arguing that it is immoral. The same may be argued vice versa, that some actions may be performed ethically yet not legally, e.g., polygamy.

This thesis created an opportunity to examine an ethical and moral narrative provided in the context of the scenario of a patient coming to harm due to a clinician using an AIS in clinical decision-making. The clear determination of ethical and moral responsibility may present an opportunity for stakeholders to consider their duties to patients in a novel way. This opportunity may prompt stakeholders to act differently to ensure patient safety, thus offering the potential for reducing the risk of harm and aiding the fair allocation of responsibility for harms should they occur. Additionally, it might help inform the formation of a fair system of legal redress.

Currently, due to a lack of case law or a specific statutory instrument, it is unclear if an SDC would be found to have a duty of care to a patient if a clinician used their AIS, though I have argued above that there is scope for SDCs to potentially be subject to a duty of care.

An ethical duty of care held by SDCs which strengthens a legal duty of care which prompts SDCs to be additionally careful in the creation and deployment of their AISs might have a beneficial cascade effect. This could benefit patients by reducing their risk of harm due to AISs subsequently being more carefully designed and deployed. This would also benefit clinicians through sharing their burden of ethical responsibility when the AIS is used in clinical decision-making, but only if legal liability followed suit. Better AISs might result in an overall increased incentive for clinicians to use them (and patients to demand them). The encouragement of patients and society to trust in AISs might thus promote their increased uptake - to the benefit of the SDC.

Current theories of ethics

My starting point for examining the allocation of moral responsibility is justice. Justice is one of the four highly influential principles of bioethics as set out by Beauchamp and Childress (2013). The others being autonomy (self-governance), non-maleficence (abstinence from harming others), and beneficence (contribution to another's welfare). Whilst relevant, these principles are not central this thesis, but each shall be touched upon later in this chapter as they arise in the discussion. The following is a brief survey of different accounts of justice that are relevant to this thesis.

Distributive justice is concerned with burdens and benefits in society being fairly distributed (Beauchamp and Childress, 2013). This concept is relevant to this thesis as it seems there is an unfairness (an injustice) in the distribution of the legal responsibility between stakeholders. If considered through an ethical rather than a legal lens, the SDC has scope to hold moral responsibility for harmful effects of AIS use in clinical decision-making. However, rather than the SDC embracing and acting upon this responsibility, it appears that the SDC has the option to take advantage of a situation that allows them to avoid responsibility. The SDC should be as interested in justice as the clinician because, as we shall see, the SDC is currently open to being accused of not taking their fair share of moral responsibility when their AISs are deployed.

To know whether the current legal distribution of responsibility is just, I must first have an account of what justice is. Beauchamp and Childress (2013, p.226) note that justice has been historically discussed in terms of fairness, what is deserved, and entitlement, and go on to define justice as "fair, equitable, and appropriate treatment in light of what is due or owed to persons" (note the similarity to the *Caparo* criteria discussed in the previous chapter). But, in ethics, how does one decide what is fair, equitable and appropriate?

There are several linked theories of justice that I can use to try to answer this question, and I will discuss them under four broad headings:

1. Utilitarian approaches
2. Egalitarian approaches
3. Communitarian approaches
4. Contractarian approaches

The texts of each of these theories are generally concerned with the fair allocation of resources, rather than justice in the allocation of responsibility. This has made discussing the theories using the works of other authors challenging; however, as will become evident, I have distilled the vital elements of each theory and discussed those elements in the context of justice in the allocation of responsibility.

The first three theories I shall touch upon only briefly. This is because these three lead to the fourth approach of contractarianism, on which the most detailed discussion is focussed (for reasons that will become clear).

Utilitarian approaches

Beauchamp and Childress (2013, p.354) describe consequentialism as "the act that produces the best overall result as determined by the theory's account of value." Utilitarianism is the most prominent of consequentialist theories; and the 'consequence' that contemporary utilitarians tend to be concerned with is positive notions such as 'wellbeing' (Beauchamp and Childress, 2013, p.354) or 'happiness' (Bentham, 1983, Singer, 2011). Utilitarianism accepts only one principle of ethics: utility; thus, one should act to achieve maximal "positive value over disvalue - or the least possible disvalue, if only undesirable results can be achieved" (Beauchamp and Childress, 2013, p.354-5). When adopted by society, utilitarianism may allow institutions to deliver the greatest satisfaction to those whom they affect (Sidgwick, 1907).

When viewed broadly and uncritically, the maximisation of positives doesn't immediately seem unreasonable. If a person is freely acting to achieve their own interests, arguably they will balance losses and gains in their lives to achieve rational ends which reflect their own greatest good (Rawls, 1999). If the effects of an individual's actions are localised to that individual and that individual's actions are not affecting others, then those actions may well be broadly permissible. But deeper consideration needs to be made when actions affect other person(s), especially when the consequences of those actions are negative and affect the wellbeing of others; indeed, it is hard to think of many significant actions that have no effect on others at any stage. As Rawls puts it:

> *"Since the principle for an individual is to advance as far as possible his own welfare, his own system of desires, the principle for society is to advance as far as possible the welfare of the group, to realize to the greatest extent the comprehensive system of desire arrived at from the desires of its members. Just as an individual balances present and future gains against present and future losses, so a society may balance satisfactions and dissatisfactions between different individuals. And so by these reflections one reaches the principle of utility in a natural way: a society is properly arranged when its institutions maximize the net balance of satisfaction summed over all the individuals belonging to it."*
>
> *Rawls, 1999, p.21*

Yet, whilst it may appear that a society should organise itself to achieve the most '*good'* (Rawls, 1999), a basic criticism of utilitarianism comes to light: it does not automatically follow that just because an act creates *good* that the act is *right.* Frankena (1973) frames this latter view as a 'teleological' theory of utilitarianism: an act is *right* and should be chosen above any available alternative act when more *good* (value) than *evil* (disvalue) would be produced by the act.

It is helpful to illustrate this with a vignette. A person might choose not to wear a mask in a pandemic as they wish to wear make-up and avoid spots or pressure sores from the mask (the *good* here is to avoid skin damage). But this may then result in the disproportionate *harm* (the disvaluing act, or *evil*) of contracting and subsequently transmitting a pathogen (e.g., COVID-19) to others. The good and harm of a scenario can be identified and interpreted differently by different persons dependent on many factors ranging from personal experience, to exposure to different educational experiences, to the values of the social group that they inhabit. Hopefully individuals would recognise that mask-wearing is beneficial to all persons at a time of pandemic crisis, but there is a potential for a wide variety of views about what right is at an individual level. For example, a person may decide that the personally acquired good of not wearing a mask outweighs the burden of the harm that they might expose other persons to. If individual actors are left to attempt to balance risks without guidance on how to fairly balance those risks when their actions affect others, it may lead to the sum-total actions of a society becoming harmful thanks to a non-unified approach of how to act in a given scenario. In this masking vignette, an individual's actions in prioritising their interests might be individually right and achieve good for them, but the consequences of their actions may leave others to unfairly suffer a burden. Conversely, if society forces persons to wear masks for the greater good without recognising and weighing the various impacts of the burden of mask wearing, then it risks alienating those for whom mask-wearing is a disproportionately harmful obligation (for example, those with brittle asthma for whom mask-wearing may trigger a life-threatening attack).

The above illustrates that it is helpful when societies and members of societies are not left to self-regulate their actions without guidance on how to achieve the balance of satisfaction between members whilst ensuring that that balance is both good and right. Rawls (1999) notes that as an individual determines and balances future gains and losses, so does society; but society balances against different individuals, rather than the individual choosing for themselves. In justice, utilitarian obligations value social utility maximisation (Beauchamp and Childress, 2013) and typically direct society in a manner which is somehow legally enforceable (for example The Health Protection Regulations 2020 which promoted national wellbeing during the height of national crisis by mandating the wearing of facemasks via legal requirement).

Utilitarianism is useful for undertaking cost-benefit analysis when choosing between two given options. Returning now to the consideration of AIS use, it could be argued that using an AIS is more cost effective and therefore a more preferable option than employing a clinician. Using a hypothetical scenario, a utilitarian approach to justice might entail that it is acceptable for an AIS to output recommendations that might result in harm to a theoretical 5% of a population if that system otherwise provides recommendations effectively and appropriately for the needs of the remaining 95% of patients. This could be argued as acceptable because using the AIS has a maximising effect (i.e. the AIS that produces maximal welfare overall). Although 'just' on utilitarian grounds, a person in the 5% may feel particularly hard done by; why should the 5% carry the burden of harm when 95% carry the benefit? This is where prioritisation of the greater good seems to outweigh the interests of smaller affected populations, and these smaller affected populations are then at risk of being disproportionately affected by a harm that might not affect others. Thus, that harm is smaller and easier to disregard as the good that has resulted from the 95% of patients who have benefitted from the AIS vastly outweighs the harm endured by the 5%, thus making AIS use right by a utilitarian standard.

But utilitarianism fails to consider that stakeholders are likely to have aims other than social utility maximisation. A hypothetical patient in the 5% who suffers a known risk of avoidable harm (where the other 95% would benefit) may legitimately ask 'why me' and question whether using this approach is just to them as individuals. Additionally, persons in the 95% will have loved ones in the 5% group, thus one cannot assume that the 95% will be happy for harm to be permitted to befall the 5%, even though they are personally to benefit from the good. The cost-effectiveness of AIS use is justified when societies employ the utilitarian reasoning that "correct distribution…is that which yields the maximum fulfilment" (Rawls, 1999, p.23) and that "there is no reason in principle why the greater gains of some should not compensate for the lesser losses of others" (Rawls (1999, p.23). Such views show how utilitarianism, when considered only in the context of maximising purchases and goods (the context in which Rawls discusses it), highlights the needs of society rather than that of the individual. Indeed, Rawls (1999, p.24) states that "utilitarianism does not take seriously the distinction between persons." Yet, disregarding the distinction which leads to inequality between persons allows space for harm to eventuate (as just discussed). This makes utilitarianism inadequate for those who would be harmed due to the use of AISs as it does not challenge that inequality. Smart and Williams (1973) describe negative utilitarianism whereby suffering is minimised rather than happiness maximised. They advise that we "worry about removing misery rather than about promoting happiness" (Smart and Williams, 1973, p.28-29). When considering the use of AIS in clinical decision-making, there is a clear desire to promote happiness by providing a useful AIS for clinicians to use. However, if a minority of patients - however small that minority may be - risks being harmed, then the theory of ethics used must consider not only those who stand to benefit from the use of AISs (SDCs), but also those patients and clinicians who would be burdened should harms eventuate. To help those at risk of harm from the use of an AIS in their care, theories based in equality may be more helpful.

Egalitarian approaches

Beauchamp and Childress (2013, p.256) describe the central tenent of egalitarian theories being that all humans must be treated equally as they have equal moral status. From an egalitarian standpoint, justice's fundamental goal is equality (Gosepath, 2021) and Temkin (1986, p.100) opines "that it is a bad thing – unjust and unfair – for some to be worse off than others through no fault of their own." In the case of the patient who has been harmed due the clinician choosing to use an AIS in their care, the patient is a subject of the harm, and not at fault for it. Aristotle's formal equality principle states that "equals are to be treated equally and unequals unequally" and that "injustice arises when equals are treated unequally and also when unequals are treated equally" (Pojman, 1995, p.2). Where persons are equal in one aspect of their lives, it is rational to consistently treat those persons equally unless there are sufficient reasons to do otherwise (Berlin, 1955-6). Indeed, it can be entirely irrational when the inequality results from a risk of direct harm that one party may be subjected to and the other not. The value of avoiding such irrationality is captured well by Philippa Foot:

"The existence of a morality which refuses to sanction the automatic sacrifice of the one for the good of the many . . . secures to each individual a kind of moral space, a space which others are not allowed to invade."

Foot, 2002, p.102.

Arguably, treating the 95% who might benefit the same way as the 5% who might be harmed amounts to treating unequals equally as it risks the automatic sacrificing of the interests of 5%. What makes this problem even more interesting is that those populating the 5% might not even recognise that they are in that group – it might not be predictable (for example, due to the opacity of an AIS) that individuals are at risk of that harm until that harm befalls them. If stakeholders (patients) do not know for certain which group they occupy, it is in their interests to ensure that all members of both the 5% and the 95% group are treated in a way that recognises and responds to the existing inequality. But, specifically, how can inequality be recognised in the patient stakeholder group?

Dworkin (1981) notes that equality values a distribution of resources that is free of envy. It is not unreasonable for 100% of patients who are already living with health issues to expect an AIS to positively aid a clinician's decision-making rather than to threaten their wellbeing. In reflection of this, it may be said that the egalitarian approaches which are centred in equality directly challenge the utilitarian approaches which would allow for casualties even though a larger number of people would benefit. Anderson (1999) notes that persons should "stand in relations of equality to others." (p.289), as such, it is reasonable to say that there is an inequality when a hypothetical 95% benefit and another hypothetical 5% experience the eventuation of harm from the use of an AIS in deciding their care. Thus, egalitarian approaches ask challenging questions of utilitarian distribution. In this case, it asks whether there are legitimate reasons (e.g., a difference between the 95% and 5%) that warrant allowing the creation of two groups which will be treated differently by way of either being helped or harmed. If they are alike in morally relevant ways then they should be treated alike, and not to do so would be unjust. In principle, 100% of patients are morally alike, as their sole uniting characteristic is that they are patients requiring care. The harm that they might experience from the use of the AIS would not be determined by their moral characteristics, just by the possession-by-chance of their physical maladies. Beauchamp and Childress (2013) highlight that there are no egalitarian theories which require that all persons should receive equal sharing of all social benefits, and that the dominant egalitarian theories "identify basic equalities while permitting some inequalities". If it is accepted that patients are morally alike, then it seems unfairly burdensome on the hypothetical 5% of patients who would be harmed by an AIS informed intervention. Such an inequality is unacceptable because the burden on the hypothetical 5% of harmed patients is too great. This could easily become a spectacularly large problem when scaled up from a small patient population to a national one and would be significantly harder to ignore or justify harms on a whole-population basis constituting millions of people.

On reflection of the above, it could be interpreted that I am arguing for rejecting the use of AISs in clinical decision-making because, for as long as AISs are unable to give 100% perfect outputs which will always benefit the patients that they serve, there will always be a risk of a patient being harmed due to the use of an AIS. However, as per chapter 2, that risk already exists in healthcare due to the existing imperfections of human clinical decision-making. As per chapter 1, clinicians have already adopted AISs in some areas (e.g., Watson for Oncology) with the aim of improving their professional practices, and there are more AIS applications in development. Currently, the choice to accept or reject the use of any tool in healthcare remains in the domain of the clinical professions, who then offer the benefit of the use of those tools to patients. Unless clinicians and patients reject the use of AISs until they are 100% risk-free, the risk of harm due to the use of AISs will remain – thus, this risk of harm is a burden that is currently partially accepted by society. This is at odds with the ethical theory just discussed and those individuals within our society who would carry the effects of that burden may very well not wish to do so. Anderson (1999, p.294) states that "justice demands that the claims that people are entitled to make on others should be sensitive not only to the benefits expected on the part of the claimants but to the burdens these claims place on others." Whilst benefits of AIS use in the clinical environment can accrue to society at large, there needs to be clear recognition of the potential risks which may result in harm for a few whilst others are receive a benefit, and consideration of how those burdens will be managed when they eventuate.

Whilst the risk of the burden of harm from inadequate clinical decision-making exists, society has provided practical legal pathways (i.e., negligence, as outlined in chapter 5) where harmed parties may attempt to address - thus equalise – their unequal exposure to harm. Harmed parties can apply pressure on those responsible for the harm by claiming financial compensation, thus incentivising those responsible to address unequal burdens. The provision of legal routes allows the opportunity for restorative justice via societal mechanisms and reflects the principle of "equal respect and concern for all citizens" (Anderson, 1999, p.289). However, as we have already seen in this thesis, the route for compensation is not established and this marks an opportunity to restore equality in the event of harm by exploring how responsibility for harms can be allocated and compensated for. This opportunity highlights the need to search for widespread agreement about the allocation of responsibility, i.e., what is owed to whom and by whom in a relational account of justice specifically concerning the SDC, the clinician, and the patient when AISs are used.

I do not accept the inequality of some being harmed where others can benefit. I do not accept that it is fair for untrodden legal routes be depended upon to address these inequalities when they arise – I demand an alternative – for these reasons, I have rejected an egalitarian approach. This is where communitarian and, specifically, contractarian approaches come in. These approaches search for agreement when questions of justice arise between stakeholders; here there is an opportunity to seek and reach agreement *proactively* before AISs are widely adopted, rather than *reactively* after the harm has occurred. The communitarian approach seeks agreement within the community and uses social norms in a society to form its standard of justice (Bell, 2020) whereas contractarianism seeks negotiated agreement between involved stakeholders.

Communitarian approaches

Humans tend to live in communities with other humans. Our contact with other persons offers and provides meaning in our lives, but a consequence of that contact can be that the community in which we inhabit informs, influences, and shapes our moral and political judgements (Bell, 2020).

In reflection of this, communitarian approaches derive the conception of good from whichever of the "diverse moral communities" is being considered (Beauchamp and Childress, 2013, p.258); what is owed to individuals and groups depends on that community's standards and promotion of the common good. Cox (1997) identifies communities as institutions infused with social norms. Whilst this initially appears a straightforward approach, communitarianism is ill defined, not least because the concept of what good *is* shifts within the multitude of human societies across the globe. That which may be accepted as good in one culture may be unacceptable in another. This makes it challenging to make sweeping generalisations regarding an issue which may be accepted by every community. For example, in Western society clinical decisions are made with an emphasis on patient autonomy; this differs from the consideration of the perspective of the whole family in the East as "an individual is regarded as a smaller self within a larger self, specifically the family" (Cheng-TekTai, 2013, p.64). Approaches to achieving justice on a global scale might never be universal due to regional differences, however if communities actively embrace the issues regarding justice when they present themselves, a communitarian approach permits solutions to be formed which answer specifically to the needs of the population that the community seeks to serve. Whilst this might result in large amounts of energy being expended to reinvent a conceptualisation of justice for every community that is affected by an issue, the output of a formulation of justice that reflects the values and norms of the members of that community might increase the satisfaction of that community, thus (hopefully) increasing wellbeing as a result. As such, a communitarian approach has the potential to be recognised as *just* and might please more persons at a local level, rather than a universal formulation of justice being created and enforced which fails to recognise and account for those differences which exist within the communities that subdivide human society. If the larger and globally relevant issues have been addressed at the highest level (e.g., as per chapter 5, the obligation of a duty of care between persons) then it makes sense that some of the smaller issues (e.g., how that duty might be formally recognised and administered for) to be determined at a local level which reflects a community's perceptions and norms.

Taylor (1979) argues in favour of communitarian approaches by outlining that the prioritisation of an individual's rights over society is problematic as individuals are not developed without the influence of social structures such as family and community (much like the view from the East as noted above). This does not mean that individual needs are completely ignored, nor does it mean that existing social norms remain static. Rather, that the application of a social norm may not be appropriate to certain individuals – for example disadvantaged stakeholders - and exploration of *why* that application is inappropriate may lead to the realisation that the norm in question is potentially non-beneficial for the remainder of the community too, the recognition of which could lead to positive changes in the accepted social norms. For utilisation of this approach to be plausible, an individual examination of each issue as it arises by key representative members of the community must be undertaken to allow the community the opportunity to reconsider its norms and improve them incrementally by updating its norms through the consideration of a collection of relevant scenarios. Bell (2020) exemplifies this with the Indian caste system; here a non-liberal society was described just as per standards set within the community itself (as per Walzer, 1983) yet Bell (2020) notes contemporary thinking from members within the Indian community now views the caste system as a past legacy that members ought now overcome. This example shows that a community's conceptualisation of *good* is not always a fixed and shared value within every global community and that a given social norm may change over time within individual communities.

Arguably, there are some social norms that achieve a conceptualisation of *good* that ought to be universally shared regardless of the accepted social norms of a community in question. Whilst universally set rules could be applied to attempt to ensure that a minimum standard of justice is achieved throughout the entire population of global humans to reduce the risk of a local standard not meeting a universally accepted minimum,[27] there remains the issue of what that standard would be and the fundamental reasoning behind it. The concepts of justice are influenced by factors such as Western or Eastern thinking. As such, conceptualisations of justice will be variously accepted or rejected depending on factors such as the geographical and/or cultural norms where the resulting

[27] Indeed, such standards do exist in the form of the various international Human Rights declarations. This started with the Universal Declaration of Human Rights, which was then applied regionally - for example The European Convention on Human Rights and the American Declaration of the Rights and Duties of Man. From the regional declarations, individual countries have ratified Human Rights instruments into their own laws.

rules are applied. With this problem in mind, the communitarian approach permits the adoption of certain key universally accepted concepts (e.g., fairness) along with the flexibility of development within the community as its values and standards change (e.g., what is a just way to deal with those who act unfairly). In this way communitarian approaches may benefit a community's members by promoting growth in the community's social values as they develop over time (e.g., updating the rules of practices to reduce or eliminate the unfair effects on members of a community). As this thesis has been written from a Western perspective by an individual whose influences have been rooted in socialised medicine and an accompanying exposure to Western philosophical and legal structures, the issues and solutions offered later in this thesis may be rejected by those who inhabit different communities. For example, there might be no culture of negligence claims in another geographical/cultural area. As such, the issue of concern might not be the allocation of responsibility for harms caused. The predominant issue instead might be the challenge of the clinician's intellectual authority from an AIS; for example: the prestige and position of the clinician might be negatively affected due to the arrival of an AIS. As such, a clinician's highly valued role in society may be challenged and result in their being socially demoted and devalued. To add insult to injury, this may even be in the context of the authority of an AIS being yet to be earned. But, whatever the issue regarding justice that is being discussed within a community, there is value in representative members of society discussing that issue and determining the conception of good that they wish to achieve and the means by which they may attempt to achieve it. Such discussions promote the ongoing sharing of evolving ideas, understanding, and problem solving that refine the conception of good that Beauchamp and Childress (2013) recognised and allows that conception to evolve as the community's accepted norms do.

Applied to the central problem of this thesis, a community may decide that their social norm will be to reject the use of AISs in clinical decision-making if there is a risk of harm. However, if a community does choose to use AISs and accept that risk, then the consequences of the risk eventuating need to be considered. The eventuation of risk does not only affect the patient, but also the clinician (as we have seen in the preceding chapter 5) - as a result, the clinician might end up carrying the burden of legal liability for any patient harm, despite the harm not being all of the clinician's making and when others could also carry the burden of harm. If others (i.e. the SDC) refuse to carry their fair share of that harm, it is *prima facie* unjust towards the clinicians forced to carry that burden. The process of restoring justice to the clinician presents an opportunity for a community to evaluate the actions of whomever else is affected alongside the clinician, e.g., those involved in the creation of the AIS in question and to allow those involved to carry their fair share of responsibility, even when they might not have been expected to do so before.

If observing a communitarian approach, discussion is needed between community members and stakeholders when examining a community standard or social norm. This discussion could lead to collaboration which could lead to alterations (and subsequently improvements) in the community standard or social norm in response to changing practices (e.g., the adoption of AISs in healthcare). This has the potential to be particularly effective when specific and localised issues are considered within a single community (even localised at the level of, say, an individual hospital) – where those who are directly affected are able to offer a meaningful and constructive voice which is heard, taken account of, and incorporated into the thinking which then shapes how future issues are addressed. Contractarian approaches could aid that discussion. The outcomes of stakeholder discussions could be reflected in a contract that embodies the new conceptualisation of *good* that is recognised by the community in question – e.g., specifically to the use of AISs in regard to the new collaborations between SDCs, clinicians, and patients in healthcare practices - and it is this approach to which the most detailed discussion in this chapter is devoted.

Contractarian approaches

Contractarianism covers two areas; one is centred in political theory regarding the consent of governed persons to accept the claim of legitimate authority of their government, the other is the moral theory resultant from the political theory that mutual agreements - or *contracts* – gives rise to this consent (Cudd 2021). For the purposes of this thesis, my discussion centres on the contractarian claim that contracts result in agreed moral norms with accompanying normative force being derived from that agreement (Cudd, 2021).

The contractarian approach embodies aspects of Kant's Formula of Humanity which I will explain before exploring contractarianism in more depth.

The Formula of Humanity declares:

> *"So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means."*
>
> *Kant, 1998, p.38*

Kant illustrates the Formula of Humanity by describing a man who wishes to borrow money but knows that he cannot repay the lender back in the agreed time. To take the money using a promise to repay that he knows he will break is wrong (Kant, 1998). It is wrong because one must act not just as agreed, but also because one's actions should be what is morally required (Beauchamp and Childress,2013, p.362). Here, people should not be treated as means to an end (e.g., the acquisition of money), but as ends in themselves (e.g., that they are people who should be treated as autonomous moral agents)

(Kant, 2011). The Formula of Humanity is respected when one agent gets the free and voluntary agreement of the other. When contracts are agreed and executed, the involved parties are not being used, because the agent has chosen to engage in the way set out. This is what makes employment, trade, and sex, for example, permissible.

To explain the relevance of Kant's Formula of Humanity within the context of this thesis, I'll summon again the quote from the unnamed Executive Consultant for IBM (quoted in Hengstler et al, 2016, p.115): "*I underline that Watson does not make decisions on what a doctor should do. It makes recommendations based on hypothesis and evidence based* [*sic*]"

Here, the SDC wants their AIS used, wants the clinician to use it, and benefit from its use. The SDC also dictates how an AIS is to be used, ensuring that they are able to benefit from its use whilst stipulating that they do not have to carry any burden of responsibility if use of the AIS leads to harm. The AIS is being presented by the SDC and the SDC has stipulated rules for its use, but it is not clear whether clinical users have consented (or could consent) to this set up. I have found no consultations in the literature or the media when researching this thesis which describe clinicians and SDCs jointly considering the AIS the allocation for responsibility for AIS error, nor negotiations where they collectively decide how much value is to be placed on the AIS's recommendations when the clinician makes their decisions. Instead, the SDC has released their AIS, dictated its limits, and dictated how the clinician should act when using their AIS, apparently without consideration of foreseeable routes to harm beyond indicating that they are not responsible for its clinical use.

It seems here that SDCs may be attempting to take advantage of the moral requirement for clinicians to make giving the best care for the patient their primary concern (as underlined by their enforceable professional codes of practice: GMC, 2020; NMC, 2018; HCPC, 2016). As this thesis has already identified, there is no enforceable compulsory code of practice for SDCs or technologists.

The contractarian approach respects the Formula of Humanity by treating clinicians as autonomous moral agents and - rather than excluding them - allows them to be involved and voice their preferences during the process of agreeing to what they are responsible for, and why. Ideally, the process of drawing up the agreement would not limit stakeholder involvement to clinicians and SDCs; patients ought also to be involved as the use of the AIS very much affects them. As such, SDCs may wish to take the opportunity to use the agreement process to express that they view healthcare as aiding patients and improving care, thus avoiding criticism of their motivations for releasing an AIS being entirely financially driven. To do this, they could adopt the comparative standards of their clinical counterparts and ensure that their AIS's outputs offered the same (or higher) level of appropriateness and relevance to the patient being treated as the decision-making of the specialist clinician which their AIS is trying to emulate.

Sugden recognises contractarians as those who…

> *"seek to derive principles of morality by analysing the problem that would be faced by rational individuals in a state of nature. Such individuals, it is argued, would recognize that they could best further their separate interests by agreeing to abide by certain rules, prescribing cooperation and mutual restraint. But what makes these rules moral?... If it can be shown that rational individuals in a state of nature would agree to follow impartial rules, then contractarianism has generated a system of morality."*
>
> *Sugden, 1990, p.768-9*

Plainly, contractarian approaches allow stakeholders to discuss the problem at hand and to jointly and rationally agree on a plan of how actors ought to act; the agreement is the 'contract' of contractarian approaches. Instead of allowing one stakeholder to dictate and enforce their rules upon others, which is an injustice, promotion of discussion between all stakeholders creates the opportunity for the solution to be influenced by and agreed by, and thus be just to, all stakeholders.

The agreement reached might not be a final robust solution; it could be the first in many iterations before a decisive solution is formulated, or maybe no clear solution is ever found. Ives (2014) describes that by trying to reach *the* solution, stakeholders might find *a* solution which is closer to the desired resolution. If no solution is found, this does not mean that the process of stakeholders communicating and exploring an issue was a useless activity; instead, it could be viewed as part of the "process of 'noble failure'" (Ives, 2014, p.304, and Huxtable, 2012). The nature of problems can change over time due to the constant fluctuation of technological advancement and social change; noble failure allows for stakeholders to persistently endeavour to find solutions through this ongoing flux (Ives, 2014, p.304).

A "nexus of contracts" (Cox, 1997, p.401) between stakeholders symbolises how agreements are made within a community. Through a nexus of contracts a community can agree that any solutions that they decide upon will ensure that certain principles will remain constant; e.g., that patient safety comes first. If patient safety is paramount, then it is easier to argue that there is value in supporting the clinician who is providing care to those patients. Focusing on the good of an individual stakeholder group through a nexus of contracts could involve some kind of loss to other members of that community, for example the SDC then must spend more to develop their technologies before deployment and make them more understandable to the clinical users. But, if a standard is set which declares that every stakeholder is treated well, the community as a whole, i.e. all stakeholder groups, would be treated well by each other. Thus, if the SDC invested more into their AIS before deployment, the initial cost might be higher, but there could be a corresponding higher amount of AIS adoption, and therefore profit, as a result. This is something that rational moral agents could agree on, and the agreement ensures that all agents are treated as ends in themselves, as not means.

Yet, it is questionable whether these negotiations are even necessary. There are a vast array of clinical tools, from defibrillators to ventilators, that inform the clinician's care decisions for which they alone have historically assumed moral and legal responsibility, without having had these explicit discussions with the devices' manufacturers. So why would negotiations suddenly be necessary when considering the use of an AIS?

The novel factor is apparent here when the AIS is designed and presented to directly influence a clinician's decision-making, as clinical decision-making (the process of weighing-up different facts and arriving at course of action) has historically been the preserve of the clinical professions. SDCs have been noted to say that they "envisage a world in which most care is “"protocolized"—that is, in which clinical decisions on the best treatment options are suggested to physicians by an automated decision algorithm (that weighs up various clinical facts) informed by advanced analytics." (Champagne & Leclerc, 2015). This differs from other clinical tools, e.g., heart monitoring or blood test results, which provide information to the clinician, but it is the clinician who weighs that information and decides how to proceed. An AIS which is developed to specifically influence the decision that the clinician shall ultimately make regarding a patient's treatment is precisely designed to enter the clinician's decision-making space.

Having outlined the broad rationale for taking a contractarian approach – that being it allows a negotiated justice that can be agreed upon by rational stakeholders – we need to consider how that approach can be enacted. To do this, I will consider three philosophers who offered theories of contractarian justice: Hobbes, Rousseau and Rawls. In the discussion that follows, I have at times drawn heavily from Rawls's analysis in his "Lectures on the History of Political Philosophy" to identify relevant points of Hobbes and Rousseau before I then discuss Rawls's contribution (which builds upon Kant, Hobbes, and Rousseau's contributions).

Hobbes

Hobbes' *Leviathan* (2018) describes all persons as equal "in the faculties of body and mind" (p.112). Regarding the body, he explains that whilst one person might be strong enough to overcome another, a "confederacy" of allies might achieve the same goal that one person alone might not. Regarding the mind, prudence and experience is equally bestowed to all men over time "in those things they equally apply themselves unto” (Hobbes, 2018, p.112); i.e., one person might feel intellectually superior to another, but all may become wise if they put in the time to acquire wisdom.

With this equality between persons noted, comes the realisation that a common feature in populations is persons wanting the same ‘ends’ to achieve a better life. Yet, “if any two men desire the same thing, which nevertheless they cannot both enjoy, they become enemies” (p.113).

A population without government is known as Hobbes’ state of nature (Hobbes, 2018; Lloyd, 2018), and, in that state of nature, persons may find themselves quarrelling for one of three prominent reasons:

> *“first, competition; secondly, diffidence; thirdly, glory. The first maketh men invade for gain; second, for safety; and the third, for reputation”*
>
> *Hobbes, 2018, p.114*

If persons are to be in Hobbes’s state of war because every man finds himself as an enemy to every other, then life will be “solitary, poore [*sic*], nasty, brutish, and short” (Hobbes, 2018, p.115). Hobbes’s drastic account is criticised (Holm, 2017) as he appeared to have merely made observations of human nature rather than using historical accounts back up his arguments. However, Hobbes’s description of war allowed him to go forward and argue that such a predicament between persons should be avoided as all men agree that “peace is good” (Hobbes, 2018, p.147). Hobbes envisaged that the establishment of a sovereign authority would facilitate peace; indeed he stated that it didn’t matter how a person came to power, be it “naturall or civill [*sic*]”, what was important was most men were “united by consent” in that person (Hobbes, p.76). The political features of Hobbes work are less significant to this thesis, however the moral implications hold more relevance. Whilst it would not be wise to follow Hobbes’s suggestions on the intricacies of how to construct a contract between persons in the modern world,[28] it can be seen that Hobbes’ description of contracts allows persons to set out and agree what each party will do to satisfy the other. By suggesting the use of contracts, Hobbes has described that justice is achieved by allowing parties to negotiate for what they want in order to meet their own rational ends, whilst the contract serves to ensure that the terms of the activity are mutually agreed and therefore have the opportunity to be fair.

To apply Hobbes, were the principles of the social contract employed, it may be possible to argue that it was rationally self-interested, but unreasonable, for an SDC to allow a clinical user to be the one who would be held responsible in a negligence claim (due to the clinician being the actor who

[28] For example, he stated that it was acceptable to uphold your commitment to paying ransoms for your life if you committed to that contract when afraid/being extorted (as per p.128, Hobbes, 2018).

performed the *novus actus interveniens* which led to patient harm). But the SDC may recognise that it would be rationally (in their interest) to act reasonably. Should SDCs chose not to act reasonably (or at least in a way that others consider reasonable), then no contract would be created, and stakeholders may choose to reject using the AIS, resulting in SDCs losing the market for their product.

As an alternative, before an AIS was deployed, SDCs could invite stakeholders to cooperate with them by taking the opportunity to discuss the introduction of the AIS and how the burden of carrying the responsibility for harms which might eventuate would be allocated. By initiating discussion, negotiating, and agreeing to a social contract, this burden of responsibility could be distributed between stakeholders in a manner that is both rational and reasonable.

This section has described Hobbes' (2018) principles of a social contract and explains why it is logical for people to adopt those principles when planning activities with others. Rousseau described the environment needed for social contracts to thrive.

Rousseau

According to Rawls, Rousseau (2002) described goodness in human nature which makes it possible for stable social and political arrangements to be created (Rawls, 2007). To Rousseau (2002), the principle of equality was highly significant; that every person has the same status as an equal citizen (Rawls, 2007). This was reflected in his argument that social and political institutions must be arranged to facilitate cooperation which is achieved by social contracts; this results in equality (social and political), moral freedom, and independence being ensured (Rousseau, 2002; Rawls, 2007). Rousseau's (2002) terms of social cooperation are known as the 'compact' and have four assumptions:

- That persons need to be equal members of the social group.
- That a person's interests are advanced through social interdependence with other persons.
- That persons have free will to act as their desire to act whilst using valid reasoning.
- That persons can understand, apply and act as per the social compact.

Rousseau defines 'will' as "the capacity for deliberative reason" (Rawls, 2007, p.223). Our common interests result in a social bond which forms the general will of society; without common interests general will dies (Rawls, 2007). General will only considers common interests and not private ones (Rawls, 2007), and according to Rousseau this is acceptable as everyone shares fundamental interests, and social cooperation ensures those interests are met in the form of basic laws (Rawls, 2007). Laws create social conditions which identify common interests and thus an environment conducive to the achieving of the common good (Rawls, 2007).

From Rawls's presentation of Rousseau, I can highlight the importance of reciprocity and equality and the common good. From Hobbes I can draw the distinction between individual rationality and the reasonableness that must frame a social contract. These values underpin social stability and make justice possible, and they are taken by Rawls to underpin his theories.

Rawls

Rawls's approach is the most recent example of contractarian theory (Miller, 2021)[29] and one of his key concepts is equality; equal citizenship gives people status within the social world (2001, p.3). Rawls posits that if one wishes to act in a just manner, a stakeholder must act reasonably and envisage first how their actions may affect others; i.e. attempt to reconcile their aims and interests with other persons using terms which both could acknowledge as legitimate if they swapped places (Keating, 1995, p.312). This would lead to those wishing to act reasonably restraining their self-interest and using cooperative principles (Keating, 1995, p.312). For this to be achievable, Rawls argues that social cooperation is required which enables justice to be conceptualised in a democratic society (2001, p.5). He identifies three essential features of social cooperation (Rawls, 2001, p.6):

1. Firstly, that recognised rules aid the guidance of social cooperation which in turn regulate the conduct of those who cooperate in the societal structure.

2. Secondly, that rules are fair when everyone accepts them; creation of a cooperative environment of mutuality or reciprocity allows participants to benefit when standards are publicly agreed.

3. Thirdly, that participants shall wish to somehow advance their own position.

As applied to this thesis:

1. The responsibility for the role of clinicians is guided by their defined authoritative professional codes of conduct and they work with patients as per their duty of care; this is not the case with SDCs.
2. The community of stakeholders (especially clinicians) may decide that it is unfair for SDCs to enter the clinical decision-making space without rules reflecting that they also carry their fair share of responsibility for the consequences which may befall a patient from AIS use.
3. Collectively, all participants wish to advance their position by doing their job well (clinicians), having their products used (SDCs), and improving their health (patients).

[29] He described his "fundamental ideas of justice as fairness" (Rawls, 2001, p.12) as a predominantly political theory which is not part of a defined moral domain such as utilitarianism (p.14). I have chosen not to reject using his theory based on this as my logical use of interdisciplinary approaches may serve to improve my arguments rather than detract from them.

a. SDCs not accepting that responsibility could affect the quality of the AIS offered for clinical decision-making, therefore affecting the patient and their position.

b. Patients might be unwilling to accept the use of AISs in if not all stakeholders involved are actively recognising and mitigating the effect of their role in the clinical decision-making space.

c. Clinicians will be less likely to adopt the AIS offered if they alone bear the responsibility for negative outcomes.

If the end result is that key stakeholder groups are dissatisfied with the position that they find themselves in, then there is high risk of a breakdown in the relationship between them in regard to the use of AISs. Indeed, this amounts to a modern quarrel, and subsequent breakdown, which Hobbes (2018) predicted.

Rawls positions his theory of justice in terms of a narrow application in political theory rather than a comprehensive moral theory; yet he states this whilst placing political theory within the domain of moral theory (Rawls, 2001). Rawls found that utilitarianism cannot offer "a satisfactory account of the basic rights and liberties of citizens as free and equal persons, a requirement of absolutely first importance for an account of democratic institutions" (Rawls, 1999, p.xii), thus explaining why he did not present a universal principle (Wenar, 2021). Instead, he surmised that "the correct regulative principle for anything depends on the nature of that thing" (Rawls, 1999, p.25). Thus, whilst Rawls's distinction between political and moral theory is noted here, his theory is still general enough to be applied to the issues raised by this thesis. Indeed, the non-specific nature of his work makes for a good flexible foundation when negotiating contracts between parties.

Rawls's theory of justice is not a perfect ethical model to determine the attribution of moral responsibility between stakeholders, not least because of his adoption of the "veil of ignorance" (to be discussed shortly), but more because it could be over-idealistic. As well as Hobbes and Rousseau,[30] Rawls was also strongly influenced by Kant. Rawls noted Kant's assertion that agreements could only arise from a coalition of all involved (Rawls, 2007, p.15). However, to develop a contract with agreement of *all* concerned can be impossible if stakeholders number in their thousands. This situation is far from implausible when considering the number of patients, clinicians, and SDCs who could be affected by the introduction of AIS in clinical decision-making. To overcome this, Rawls offers Kant's solution of the hypothetical contract: that it is possible to determine what it is that rational people would hypothetically agree to. To take this one step further, I suggest that a representative sample of persons (rather than all persons) might engage in the cooperative process so that an

[30] He outlines their work in his Lectures on the History of Political Philosophy (Rawls, J. 2007).

arrangement may be determined. Whilst this would still amount, strictly, to a hypothetical contract insofar as it does not have the agreement of everybody affected, relevant stakeholders could nominate a voice which speaks for their group, thus allowing their voice to be heard and some actual agreement reached.

It seems obvious, however, that each stakeholder will likely have differing views on responsibility in this context. Rawls' 'veil of ignorance' (discussed next) might help here, not because it makes one party understand the perspective of the other, but because it provides a device by which one may impartially consider and weigh the individual perspectives of each group's interests.

Rawls's (2001) 'veil of ignorance' asks the decision maker to determine a solution to a problem of justice whilst not knowing the position they themselves occupy in society; Rawls calls this the 'original position'. If the decision maker does not know their condition, (e.g., whether they are rich or poor, sick or healthy) Rawls suggests that the decisions that they make would be just for society rather than favouring their own position (Rawls, 2007 p.17-19) - the idea being that all people are equals and making decision from behind the veil of ignorance prevents too much weight being given to advancing the one's personal position or interests, and focuses on general societal interest that would ensure those who are worse off would be looked after. Rawls expected the veil to be used as a thought experiment rather than taken literally (Rawls, 2001, p16-17), thus stakeholders could make use of the veil of ignorance as a tool to aid discussion.

My initial reaction to this concept was that it is impractical and idealistic, as individual and group held biases and influences will not be put to one side when making decisions just because the thought experiment demands it. As we saw in chapter 2, biases can be unconscious, and thus challenging to account for in decision-making. Yet, given that Rawls intended the veil to be used as an abstract thought experiment for public- and self-clarification (Rawls, 2001), one may use the veil to make conscientious efforts to minimise the effects of bias and vested interests. Rawls notes that a given person will presumably reach various age milestones; all of which will have varying healthcare needs. This means that requirements at one stage of life must be balanced against those required at another stage (Rawls, 2007). Nonetheless, in practice, this is a very difficult position to take; it is impossible for any person to not be influenced by their socio-economic group, their own health experiences, and that which they have witnessed of others. One may be asked to put aside such influences, but it is not unreasonable to suspect that, at some level in every individual, some of these influences and biases would remain and affect a decision maker's final choice.

Edwards and Deans (2017) describe the Rawlsian account of ethics as one where an ethicist would "play a substantial role in specifying and applying the content of public reason" rather than actually deliberating with the stakeholders concerned. This is a notion that I disagree with, as stakeholders ought to be involved in negotiations and the formation of agreements so that they may ensure that their interests are clearly represented. So, at this point, another approach is worthy of consideration for use during the stakeholder co-operation process. Edwards and Deans (2017, p.61) raise the option of a Habermasian (1990) approach when policy is being decided. They describe stakeholders reasoning freely and equally within an appropriately structured space to reach legitimate decisions. Elements of Rawls and of Habermas could be employed to encourage stakeholder deliberation and negotiation to reach an agreement regarding the use of AISs. This mixed approach could be inclusive of direct stakeholder inclusion, whilst promoting self-representation, communication, and negotiation between stakeholders. It could also increase the potential to reach equality by using such tools as the veil of ignorance – and the adoption of ethicists who are external to the stakeholders to help deploy the veil - so that stakeholders may consider each other's opinions without disadvantaging each other (despite that tool's flaws[31]).

Without employing the veil of ignorance, the SDC's actions may be biased by their interest in protecting their beneficial position of their AIS being used while being insulated from responsibility (by the clinician) for any negative outcomes. Such a position allows SDCs to have more than their share of freedom from responsibility. SDCs holding this view would be damaging to clinicians as they would then be burdened with carrying all responsibility for the use of the AIS when they are not the only actors in the decision-making process. This view would also be damaging to patients, as there is risk of an inadequate AIS being released and used as a result of the SDC failing to adopt meaningful responsibility.

The veil of ignorance could be a useful tool for the SDC to quantify the potential effects of their intended role in healthcare as a whole: to allow them space to consider the consequences of the deployment of their AIS, and to facilitate their empathy towards other stakeholder groups when they consider their own position and actions. Indeed, if neither clinicians nor SDCs knew the position that they occupied, the rational solution would be developing the AIS to the point of eliminating or mitigating risk to mutually acceptable levels prior to AIS deployment and the equal sharing of responsibility for the consequences of its use.

[31] Whilst reading and selecting appropriate ethical theories to use when preparing this thesis, I reached the stance that "all models are wrong, but some are useful" (Box, 1987, p.424). Nozick (1974, p7-8) makes a similar point about the usefulness of erroneous theory. Whilst the veil of ignorance is seen critically because people will always be influenced by their own position, the concept of the veil of ignorance can encourage them to consider others.

On reflection, I am clearly a biased stakeholder within this debate as I am a clinician. In recognition of this I have attempted to adopt the veil of ignorance to some degree in this thesis by speaking in very general terms of the SDC, the clinician, and examples of AISs being used, and not giving immediate preference to any particular set of interests. I have referred throughout this thesis to a specific scenario (IBM Watson for Oncology) but I have not attempted to obtain further details beyond that which is available in the literature. I could have made it an objective of this thesis to enquire why a specific organisation has taken its actions and opinions, but instead of focusing on a specific SDC organisation, I have chosen to attempt to step back and explore a non-specific scenario (of a clinician using an AISs output which leads to patient harm). By attempting to avoid allowing one specific vignette to influence the moral consideration of the entire application of AISs in all forms of clinical decision-making, I looked instead at the issue of assigning responsibility to stakeholders when AI is used to inform clinical decision-making as a whole. By doing this I am attempting to not have an identifiable interest in the situation and to derive general principles from specific cases. This prevents specific details from one scenario being used to reach an unbalanced and overly specific position. If my deliberations only fit those scenarios identified, this thesis would not generate a generalisable view which may help stakeholders in the future when responsibility and justice are considered in this context.

To recap, the veil is a device to determine the basic principles of a system of justice and Rawls finds that decisions made under the veil will tend to advantage the most vulnerable in society. Using it reminds the thinker to make decisions in a way that guards against partiality and considers all interests equally rather than simply advancing one’s own. Here, one can see the Kantian influence on Rawls’s work, in the sense that Rawls follows Kant in seeking impartial reasons for action (as expressed most vividly in the Formula of Universal Law) – “*I ought never to act except in such a way that I could also will that my maxim should become a universal law*” (Kant, 1998, p.15)

Several approaches to arriving at the just allocation of responsibilities have been examined now, and each approach provide insight into how burdens and benefits in society could be fairly distributed. But to actually apply them, we need to know who the interested and relevant parties are who carry these risks and benefits. Whilst all members of society could have interests in the use of AISs in their healthcare system, not all will be directly involved or potentially have responsibility for consequent harms. And, as already outlined above, it is the actions and interests of the SDC and the clinician that is my focus. Although we have now considered how we might achieve a just allocation of responsibility in principle, the discussion has been limited to procedural aspects rather than substantive. The contractarian approach I have adopted tells us how we can find justice through agreement, but it does not tell us what is reasonable for people to agree on. For that, we need to consider the conditions under which moral responsibility can be reasonably be allocated.

This next section examines this problem and begins with an exploration of 'but-for' causation.

"*But-for*" when determining causation

This thesis has already touched upon *but-for* causation when discussing legal causation and remoteness in chapter 5. We return to *but-for* causation now for further examination within the ethical and moral context rather than the legal.

Causal models are used to understand the what, why, and how of an action with the purpose of allocating responsibility (Lagnado and Gerstenberg, 2017). They look at what *actually* happened (including a person's motivations and beliefs whilst acting) as well as what *could* have happened to determine what *should* have happened had a reasonable person acted (Lagnado and Gerstenberg, 2017).

Causal models are not without their critics. Hume (2014) posits that we can never actually observe cause, we merely theorise it, which make us prone to causal error when trying to pin down exactly what cause led to what effect. One way around this problem, which has something of a Kantian flavour, is that we might not be able to see or feel a connection, but that the theoretical structure of causality and dependence require the idea of a connection. A theorised connection between cause and effect can be declared a valid connection if it has been shown as universally demonstratable and that all judgements on the connection are in agreement. As applied to this thesis, a cause and effect connection's validity is brought into question if there is a lack of agreement of the existence of that connection between all stakeholders. Through outlining the relevant arguments which determine causation, it is hopeful that this thesis shall promote agreement via the use of reason when stakeholders consider causation in the use of AISs in clinical decision-making.

The *but-for* test is the standard test for causation. There is '*but-for'* causation if it is true that if the actor had not acted, the result would not have eventuated (Lagnado and Gerstenberg, 2017). There does seem to be *prima facie* case for '*but-for'* causation being attributed to the SDC for any outcome arising from the AIS they develop. The SDC develops an AIS to aid clinical decision-making; the clinician may take advice from the AIS, the AIS's output is one which, if taken, is harmful to the patient, the clinician follows the AIS's advice, and this results in the patient being harmed. This outcome would not have occurred '*but-for'* the SDC developing the AIS. But the same can also be said for the clinician who chose to use the AIS. *But-for* the actions of both the SDC and the clinician, the patient would not have been harmed. Because many actors may stand in a *but-for* causal relation to an outcome, it is difficult to identify the agent who is overall responsible for the outcome. A solution to this could be to hold the clinician and the SDC jointly responsible for the use of an AIS's outputs.

Clearly, the *but-for* test is not a perfect mechanism of allocating moral responsibility for an outcome; my critique is as follows.

But-for is over-inclusive

There could potentially be unlimited *but-for* factors that might contribute to a patient coming to harm or avoiding coming to harm. Here are a few possible examples:

- It might have been possible for the patient to have taken preventative action to preserve their health some years prior. This action might have prevented their presenting illness from developing that necessitated the clinician's aid in the first place. *But-for* their inaction, there would be no harm.
- The clinician could have taken the day off as holiday thus avoided causing harm that day. *But-for* that decision there would be no harm.
- The SDC could have avoided harming the patient by developing that AIS in a different area of clinical care, thereby avoiding that patient. *But-for* that decision there would be no harm.
- A fire could have broken out that day, forcing an evacuation and consuming the AIS and preventing that clinical decision being made, but a porter stopped the fire spreading using an extinguisher. But for that porter's actions, there would be no harm.

Many of the reasons above are insignificant, being distant or coincidental factors. They might have nothing to do with the actor's behaviour towards the injured party (Lagnado and Gerstenberg, 2017), and therefore cannot be seen as morally significant.

Conversely, an actor's causal contribution to a situation might be small and distant, but their moral responsibility for the effect may be great (Mumford & Anjum 2013). For example, in clinical nursing practice one nurse may check the calculations of another prior to drug administration. If the checking nurse agrees with the calculations of the administering nurse and signs the drug chart to confirm, their small contribution would have permitted the administering nurse to give the drug. This small act of clinical administrative gate-keeping is loaded with professional, as well as moral, responsibility.

Distance becomes an important consideration when determining morally relevant *but-for* causation, especially when an actor's contribution might be separated from the effect of their actions through either circumstance or the influence of other actors. Rather than only considering the traditional physical or temporal meanings of distance, distance can be conceptualised through the effect of actions. An action which creates an effect also results in a degree of causation being attributed to that actor. Menzies and Beebee (2019) use a simplified account of Lewis (1973) to explain why:

> *"Where c and e are two distinct possible events, e causally depends on c if and only if, if c were to occur e would occur; and if c were not to occur e would not occur…*
>
> *…Where c and e are two distinct actual events, e causally depends on c if and only if, if c were not to occur e would not occur."*
>
> *Menzies and Beebee, 2019*

To apply Lewis's rule to this thesis's key scenario of harm eventuating to a patient due to clinician relying on an incorrect AIS output: if the harm was causally dependant on the AIS being deployed by the SDC then there must be a causal link between the harm and the SDC. It follows that if the AIS was not deployed, then there could be no causal link between the harm and the SDC. We have seen in the last chapter that distance and causation are dealt with legally by using *novus actus interveniens*. *Novus* breaks the causal link when the clinician uses the AIS to determine care. There are practical advantages in law to allowing causal links to be broken rather than allowing every possible causal link to remain and be claimed against. But ethically there is no reason for that link to not be examined so that the impact of the SDC's actions can be considered as part of a holistic assessment of the events and actions which may have caused a given harm.

It could be argued that holding a person morally responsible for standing in a simple *but-for* causal relationship to an event is over-inclusive, and too much of an imaginary stretch. However, there is an argument for adopting a highly inclusive approach when considering *but-for* causation in relation community wellbeing.

Consider the following example: if someone lights a bonfire in their garden to burn treated wood, resulting in acrid smoke that causes a neighbour to suffer an asthma attack who dies as result, one could plausibly argue that that the burner caused that harm in a way that makes them morally responsible for it. *But-for* their decision to light the fire, the asthma attack would not have happened. The burner, however, may claim that there was too much distance between their actions and the effect of the asthma attack. For example, the wind had to be blowing in the right direction for the neighbour to be affected; the neighbour happened to have asthma; the neighbour happened to be outside as the smoke blew past; the neighbour did not have the right medications to hand because the repeat prescription had been delayed; and the ambulance was delayed by two minutes because of dense traffic. Here, the concept of distance creates space between the purported cause and the effect, and involves multiple other links in the causal chain. The end result of the action could have been different had any number of different events taken place in between, and this arguably reduces the moral significance of the action that set it all off. There is clearly distance between the putative cause and effect that calls into question the burner's moral responsibility for the death, but a key question to ask is whether or not the burner had any positive obligation to avoid creating acrid smoke in the first place, and whether it is reasonably foreseeable that producing acrid smoke could harm another person. It is reasonable to think that anyone living in a built-up area who clearly shares air with others has a responsibility to avoid creating acrid smoke (be it formally via local restrictions, or informally via either consideration of others or from specific individual residents' requests), and it is certainly foreseeable (a concept we shall return to later) that burning treated wood would create acrid smoke that could be harmful if breathed in. In this way, we might consider the burner morally responsible for their neighbours' death despite the presence of distance.

If it is possible for specified actions (or omissions) to affect members of a community in specific ways, then those actions could be anticipated and planned for. Employing the above example, a community could dictate that garden bonfires must only take place in specified hours or not at all, and that they must not produce acrid smoke, and this creates obligations for members of that community. Similarly, SDCs, clinicians and communities could collectively identify and plan for potential *but-for* causes when an AIS is deployed, including consideration of what prior positive obligations are and should be in place to avoid a harmful consequence.

Pragmatically, actors in a health service are not limited to the SDC, the clinician, and the patient. The potential list of additional actors who would have an interest in the adoption of AISs in healthcare could be enormous and include actors such as service regulators, patient and public pressure groups, hospital trusts, and NHS governing structures (such as the overarching Department of Health and Social Care). These additional actors may help create the healthcare environment in which care is delivered, but it is the SDC, the clinician and the patient who are the primary actors. They are also the proximate actors, whose actions directly impact on one another - the distance between them being much less than between other actors. Hospitals can adopt their AISs, patients can lobby to have them used, but it is the SDC who has programmed and provided the AIS and advocated its use, and it is the clinician who is the final decision maker at the point where harm could eventuate. There is, of course, the risk that I have oversimplified by only focussing discussion on these two actors, but, even though it would be reasonable to consider the actions of additional actors, it is unmanageable within the confines of a single thesis.

Distance may be a limiting factor when attempting to mitigate the over-inclusiveness of *but-for* causation, but it is reasonable to assess distance alongside foreseeability and the existence of prior obligations. If an actor has significant distance from the effects of their actions, but it is foreseeable that their actions would have an effect despite that distance and/or if they had an existing obligation to refrain from the causal action, the actor would have made a morally significant causal contribution to that effect. Foreseeability is addressed in greater depth later in this chapter.

But-for is imprecise, difficult to prove, and can result in pre-emption

'*But-for'* causation compares what hypothetically could have happened against what actually happened, and if the action is analysed carefully, one may show proof of causation (Lagnado and Gerstenberg, 2017).

For example, the clinician might have made the wrong choice by following the AIS's outputs, but they could have made the exact same decision under their own cognitive power, regardless of whether the AIS's output was being utilised. Because we cannot be certain what decision would have been made, or what would have happened if a different decision was made, in that counterfactual scenario it is difficult to know for sure whether the clinician was the *but-for* cause

Clinicians can make their decisions alone or as part of a team, and this complicates the *but-for* model. Consider the following examples. If in this scenario the clinician would have usually consulted with a colleague to check their thinking prior to making and acting on their clinical decision, and if the AIS is directly replacing a colleague's opinion, and if the clinic would not have embarked on the course of action without the confirming opinion from the colleague, then sharing the causation between the clinician and the SDC/clinical colleague seems justifiable because both stand in a morally relevant *but-for* casual position. Where the clinician makes their own decision, alone, they take sole responsibility for their actions and stands alone in a '*but-for'* position. However, if a clinician consults with a colleague or an AIS, the AIS's output/colleague indicates a course of action that confirms a faulty decision that the clinician would have embarked on anyway, not only is the clinician falsely reassured that their plan of action is correct as confirmed by the AI/colleague, but a window of opportunity where a different decision could have been made might be lost. It is unclear in this scenario whether their failing to correct the faulty plan of action positions the AIS's output/colleague in a *'but-for'* causal relation. Here, the outcome would have been the same if there were an AIS/colleague or not, but the AIS/colleague would have failed to help the clinician. This failure would be the exact opposite of what consulting an AIS/colleague would have been created to achieve.

This failure can result in the problem of pre-emption when considering *but-for*. Pre-emption is where an alternative action could have generated the same result as the original action – essentially one action that would have led to effect 'E' is pre-empted by the other, which also leads to effect 'E' (Lagnado and Gerstenberg, 2017). This may result in the *but-for* test giving us the wrong answer (Lagnado and Gerstenberg, 2017) – in this case, whilst an AIS's output/colleague might have been involved in the clinical decision being made, their contribution did not affect the end result. This is because the clinical decision-maker had already decided their plan of action. Therefore, in the absence of a convincing argument from the AIS's output/colleague, this plan of action would not have changed. The AIS's output/colleague had not influenced a change in events, they had merely not stopped what was already about to take place.

It might be that one wants to assign responsibility to an SDC for the outcome of the use of their AIS's harmful outputs, but if the clinician would have made the same wrong decision autonomously, independently, and without consideration of AIS outputs, the SDC cannot be held responsible because the effect would have been the same whether the AIS had been involved or not.

This creates problems for determining morally significant *but-for* causation, because where pre-emption occurs it seems to be matter of simple moral luck whether another party is implicated and, clearly, if the same effect would have occurred anyway then consulting the third party did not '*but-for*' cause the effect. The problem, however, is that when pre-emption occurs it is difficult to be sure, in any sense of word, that the same decision would be made, and the same effect would therefore have followed without the consultation (as it all remains hypothetical).

But-for risks resulting in overdetermination

Overdetermination happens when two or more actors simultaneously perform the same action, and each action, individually, would have been enough to cause the harm. According to simple *but-for* causation, either both or neither individual actor caused the harm, as if one actor had abstained the other actor would have caused the harm (Lagnado and Gerstenberg, 2017). This differs from pre-emption as both actors perform the same action (Lagnado and Gerstenberg, 2017). Here, *but-for* reasoning concludes that although harm has clearly been caused, no individual has caused that harm. In Lagnado and Gerstenberg's words:

> *"The textbook case of overdetermination is when two people (A and B) independently and simultaneously shoot the victim, and either shot alone was sufficient to kill the victim. On the but-for test, neither shooter is a cause of the victim's death, because if A had not shot, the victim would still have died from B's shot, and the same is true for B. But it is counterintuitive to conclude that neither shooter caused the death. What makes this different from pre-emption cases is that each shooter does exactly the same thing and we want both to be judged as causes of the death."*
>
> *Lagnado and Gerstenberg, 2017, p.571*

Overdetermination means that no individual can be held morally responsible for harms caused, even if the motivations of both actors is to cause harm.[32]

Overdetermination was initially a concern of this thesis as, together, the clinician and the SDC's AIS reach the determination of a course of action and this action causes harm. However, it is not obviously overdetermination when we consider that the SDC's and clinician's actions are not the same. They are, in fact, quite different. The clinician has direct contact with the patient and the final decision to use the AIS's output, but the SDC has influenced the clinician via the SDC. Whilst both actors have 'shot a gun', they were different guns and different times, and the harm only eventuates after the second shot.

Even if the actions are not the same, however, they are still relevant. Stapleton (2008, 2009) uses an extension of *but-for* of "contribution" to show this. Here, the actor's actions and omissions are weighed for their contribution to the harm which has resulted. Using this mechanism, it might be argued that both the SDC and the clinician have contributed to the harmful outcome, but in different morally significant ways; one designed the AIS and provided it to be used, the other used the AIS. The result of both of those contributions is that the patient was harmed – and both actions were necessary for the harm to have come about. The harm would not have occurred but for the clinician using the AIS, but the clinician could not have used the AIS but for the SDC placing it into the decision-making space. Being able to identify the part each actor played in the outcome of a series of events means that if it is found that someone is found to have acted in a way that contributed to harm in some way, then allocation of that actor's moral responsibility for that harm can be made.

A flaw with the extension of contribution may lie with how the actors are treated once their contributory acts have been determined. For example, a clinician and an AIS read the same X-ray image and they reach the same conclusion, but using different processes to reach that conclusion (i.e. human thought and an AI process). If both the clinician and the AIS missed signs of cancer in that image, the harm would be a lost opportunity for possible treatment. Both the clinician and the SDC's actions would both have been 100% wrong, but to split moral responsibility equally would mean the clinician is 50% responsible and the SDC 50% responsible. This accounts for 100% of the harm, but each actor is only held partially responsible for a wrong that was 100% theirs. An alternative could be that that the clinician is held 100% responsible and the SDC 100% responsible, but this is also flawed as the patient has not suffered 200% harm. However, it would not really matter either way if, instead of being used as a finger-pointing exercise, or to calculate the extent of liability and contribution to

[32] A good example of this at work is firing squads. This is exactly the reason it was a squad, rather than an individual executioner.

financial compensation, we use allocation of moral responsibility to create opportunities for actors to improve their actions. So here, as the clinician and the SDC have each both contributed 100% of their own faulty actions, they each undertake to 100% identify the fault in their own processes. If each actor were to accept their own contribution to a final negative outcome, then there is an opportunity for them to make amends to a harmed patient through performing actions of restitution. Moral restitution could come in the form of contrition from the actor followed by behavioural changes which aim for a reduction of the eventuation of future risks, for example new practice/policies regarding the development and use of AISs. Financial restitution is a challenging problem that cannot be ignored, and I offer a solution to this in chapter 7, but whatever solution is adopted would ultimately have a monetary cost to the SDC and the clinician to ensure that a harmed patient's needs are provided for. A dual approach of moral and financial restitution would benefit the patient population by reducing future risks, and financially motivating the clinician and the SDC to take steps to prevent that consequence from repeating.

Addressing the problems of *but-for* causation

Pre-emption does not appear to be an issue when determining *but-for* causation in the context of this thesis, as the clinician is required for the outcomes of an AIS to reach the patient. If the clinician does not consider the AI outputs, then that causal route is blocked; the AIS cannot reach the patient without using the clinician.

Over-inclusivity and overdetermination need to be addressed, however, before utilising *but-for* causation. Regarding over-inclusivity I am, again, content to limit the number of stakeholders considered to just the SDC and the clinician as they appear to be the two key contributors at the point where harm may eventuate when care is decided. If harm has occurred, the actions of the SDC and clinician are morally relevant if they have both acted, even though those actions are not the same. The *but-for* extension of contribution could be useful here as it considers the motivations and obligations behind an actor's actions and the role they plated in bringing about the harm – allowing for more than one party to make morally significant causal contribution. I suggest, and now shall demonstrate, that employing theories of moral responsibility in the context of this thesis justifies using the extension of contribution model.

Linking morally significant causation and moral responsibility

It is not possible to stop every person from acting in a way which causes harm others. People will make their own choices and there is scope here for a discussion of how one person's acts can give rise to the opportunity for others to also act and the outcome of those acts to cause harm to a third party. Arguably, as Eddie Izzard notes, whilst one agent might be clearly causally and morally responsible through their actions, there may well be both causal and moral responsibility on the part of the agent who enabled that action by providing means and/or opportunity.

> *“The National Rifle Association says that, "Guns don't kill people, uh, people do." But I think, I think the gun helps. You know? I think it helps. I just think just standing there going, "Bang!" That's not going to kill too many people, is it?”*
>
> *Eddie Izzard, 1998*

A gun would be harmless *but-for* the person pulling its trigger. However, *but-for* the gun manufacturer providing the gun, a shot would never have been fired. The gun manufacturer made the weapon available, but did not decide in what capacity their product was to be used; the end user was the final actor rather than the manufacturer. There is certainly a *but-for* causal responsibility on the manufacturer for harm caused by one of their weapons, but simple causal responsibility differs from moral responsibility. Thus, we now need to consider in what circumstances moral responsibility follows from causal responsibility. This has been hinted at above where we talk about morally significant causation, but it deserves more in-depth exploration.

As applied to this thesis, an SDC may be causally responsible for the consequences of creating and distributing an AIS which has caused harm, but what exactly was their *individual* contribution? If the clinician uses an AIS output which they should have recognised as faulty, are they personally morally responsible for harms which eventuate as a result? If the clinician is the one ‘shooting the gun’, can the SDC be personally morally responsible? Is it fair that the clinician be held personally morally responsible for using the AIS and following its recommendations if that is exactly what the AIS was designed to do? If both the clinician and the SDC could both be held morally responsible, should that responsibility be assigned jointly? Or should each stakeholder be considered separately?

To address these questions, I shall first identify and define what personal moral responsibility is so that it can be factored into the reasoning of how responsibility may be allocated to stakeholders.

Personal moral responsibility

The full spectrum of debate surrounding moral responsibility is complex (Fuscaldo, 2006), and too large for this thesis of consider in full, thus discussion shall be limited only to this thesis’s direct concerns. To begin, I shall first provide a definition of personal moral responsibility, and then examine a distinction between forward and backward looking moral responsibility, which will be essential in considering how moral responsibility can be allocated to both clinical users of AISs and the SDCs which create the AISs.

As noted in chapter 3, a person’s account of their actions is linked to responsible behaviour, which is characterised by the “common norms which govern conduct” (Oshana 2004, p.257). Personal moral responsibility is the individual's obligation or duty to ensure that something is acted or obtained, and this individual's burden is attached to them due to the role that they fill within the context being discussed (Zimmerman, 1992). It seems that Zimmerman's personal moral responsibility is *personal* because it's the actions of the individual which are being taken into account and this definition lends itself well to considering specific scenarios rather than the general consideration of societal norms. Zimmerman's definition was envisaged as being assigned to individuals, but stakeholder groups sharing the same characteristics, aims, values and goals might each be awarded similar or identical levels of personal moral responsibility. As noted earlier in this thesis, the term 'SDCs' is dominant as, generally, an organisation would create and deploy an AIS for the clinical setting, rather than a single technologist operating alone. Whilst it may seem odd that a company could have anything 'personal', it is an individual organisation and (unless individual technologists are singled out for scrutiny) is treated as a united entity – thus 'personal' to that organisation. Employing Beauchamp and Childress's (2013) principle of non-maleficence, it is reasonable to assume that the obligation of one person to another would be, at the very least, to ensure that the first did not harm the second. This could be extended to utilising another of Beauchamp and Childress's (2013) principles, beneficence, which would oblige the first person to contribute to the second's welfare.

Personal moral responsibility can be broken down into two types (Zimmerman, 1992):

1) Prospective (forwards looking) personal moral responsibility can be identified as moral or legal or defined by a set of rules. Duty of care (having personal moral responsibility towards an individual) is an example of prospective responsibility.
2) Retrospective (backwards looking) personal moral responsibility is to be personally morally responsible for an outcome. Here approval or disapproval of actions would be expressed morally or legally.

Yeung (2019) nods to prospective and retrospective responsibility in their report considering the implications of artificial intelligence use within a human rights framework. They rightly state that "only if both the historic and prospective dimensions of responsibility are attended to can individuals and society have confidence that efforts will be made first, to prevent harms and wrongs from occurring, and secondly, if they do occur, then institutional mechanisms can be relied upon to ensure appropriate reparation, repair and to prevent further harm or wrongdoing" (Yeung, 2019, p.49).

Such mechanisms are very much in place for the clinical professions. Regulation exemplifies the adoption of both prospective and retrospective responsibility, as evidenced by the development of the professional body's codes of conduct and adoption of 'fitness to practice' hearings to enforce those codes. Hearings allow concerns about a registrant's behaviour in practice to be raised (therefore the individual being held retrospectively responsible) and/or if the registrant is fit to practice further (therefore the individual being held prospectively responsible). A negative outcome for the professional could include removal from the profession's register and preventing the clinician from taking up clinical work (HCPC, 2018; NMC, 2018a; GMC, 2021a), thus providing a mechanism for the profession to prospectively prevent patient harm.

Although personal moral responsibility is embraced by the clinical professions, SDCs do not have an established and formalised professional duty of care as they are not regulated as the clinical professions are. This lack of regulation means that they are not burdened with the consequences of professional regulation. SDCs are not formally held accountable by their profession and therefore not officially penalised by a regulator should they fail to take care. Given that SDCs are novel actors in a decision-making space historically occupied solely by clinicians, it is important to ask whether they *should* be held similarly accountable.

I shall now further unpack prospective and retrospective moral responsibility to see how they are useful approaches when considering personal moral responsibility for SDCs. This will include examining the conditions under which an actor can be morally responsible for an outcome in which they are causally implicated.

Prospective personal moral responsibility

When considering prospective personal moral responsibility, SDCs are attempting to seek to influence patient care by creating and deploying AISs which they claim are able to aid the clinician in their decision-making. If the SDC is placing themselves in a place of clinical authority, and aim to influence the clinical decision-making space, they need to accept that by entering that space the contribution of their AIS will directly affect the patient to whom the decision-making relates. As a result, it is arguable that SDCs should be similarly prepared to adopt their clinical counterpart's position of adopting prospective personal moral responsibility towards patients, thus a duty of care. Therefore, should an SDC release an AIS which aids clinical decision-making, a prospective personal moral responsibility towards the patient could be owed even if it cannot currently be enforced.

Patient safety is a central value in healthcare, and so if technologists are entering the clinical space, they must adopt the duty of keeping patients safe - just as the clinical professions do. For example, if the AIS that the technologist supplies is at risk of either giving incorrect outputs to the clinical user, or if the AIS fails to recognise that it is inadequate for the situation that it is involved in, the AIS should advise the user to call for better help. To not do so would amount to a technologist failing to take care to deliver an AIS which is safe enough[33] for the targeted patient group who are to receive interventions based on the AIS's outputs; here patients could be placed at risk and made vulnerable to harm. It seems clear, at least in principle, that the technologist has a duty of care (even though this duty has not been formalised through professional regulation membership) and prospective personal moral responsibility to their target patient group.

As prospective personal moral responsibility has been argued to be owed not only by clinicians but also for SDCs, it is arguable that, as well as clinicians, SDCs should have personal moral responsibility to deal with the outcomes of adverse events arising from the use of their AISs. Responsibility to deal with the outcomes of adverse events is encompassed in retrospective responsibility.

Retrospective personal moral responsibility

Retrospective personal moral responsibility is not concerned with a duty which is yet unfulfilled, but of one failing to fulfil a duty owed to another (Zimmerman, 1992, p.1089). The consequences of behaviours that this thesis is concerned with are primarily the eventuation of any kind of patient harm, but there are also secondary harms which may come to the clinician, the SDC, and healthcare as a whole. If AISs are eventually to be widely adopted in healthcare to aid clinical decision-making, then the reputation of this technology will also reflect upon those who are involved in delivering and using

[33] Some words on the judgement of how safe a system is. "Safe" is an ambiguous term. Does safe mean that an AIS will never make a mistake? That its outputs are 100% accurate every time that it is used? Or does safe mean that an AIS is correct a prescribed percentage of the time? One could compare the safety of an AIS to its approximate human counterpart, but humans are fallible and not correct 100% of the time. If anything, humans are suspicious of other humans who claim to be 100% correct at all times.
IEEE (2017, p.49) helps here by recognising the challenge of 100% accuracy: "because designers cannot anticipate all possible operating conditions and potential failures of AIS, multiple additional strategies to mitigate the chance and magnitude of harm must be in place." Indeed, Holms et al (2021, p.175) predict that "AI systems will be introduced when they make fewer errors than HCPs, not when they are perfect."
Instead of demanding that an AIS is completely safe, for the purposes of this thesis I'm going to state here that an AIS must be "safe enough" before exposing it to clinical use and I suggest that, at some point there needs to be a considered formal agreement on how to identify when an AIS is safe enough. It is reasonable to assume that agreement needs to be negotiated between regulatory, professional, and lay stakeholders to ensure that all agree of what safe enough is and that everyone is comfortable with the level of security and risk that is afforded by that definition when AISs are used in clinical decision-making. Once "safe enough" has been identified and stipulated it is easier to suggest that it is irresponsible for a system to be released which does not meet that agreed definition.
I am concerned with the scenario of the SDC who has released an AIS for use which has not been formally agreed as safe enough nor transparently rigorously tested. My concern is relevant and valid as IBM's Watson team appear to have done exactly that in Mongolia.
SDCs may argue that novus actus interveniens puts the clinician in line to hold legal responsibility for the use of the outputs of their systems; this position has permitted the lack of negotiation in defining and confirming what safe enough means in an AIS used in clinical decision-making. This lack of shared decision-making has allowed SDCs to stipulate what safe enough is in terms which might be fine for them, but might not be fair or practical to the end user.
This footnote is in danger of turning into a typology of safety, but when I refer to an AI system being safe, I mean that it is "safe enough" to use without unduly risking patient harm. Formal identification of what "safe enough" *is* is outside of the scope of this thesis.

it. Retrospective consideration of the negative effects of using AISs in clinical decision-making could influence the choices that a future patient or clinician may make when choosing an individual's care route; an AIS which has been deployed or used carelessly could affect the future development, deployment, and uptake of other AISs. This could ultimately harm the society it is trying to serve due to a lost opportunity for safe and considered utilisation of this technology.

Retrospective responsibility is established with two conditions. A person is responsible for the consequences of an action if (and only if):

1) the agent was free to act otherwise/acted voluntarily and
2) the consequences of the action were reasonably foreseeable (Fuscaldo, 2006).

Fuscaldo holds that if a reasonable person could expect the occurrence of consequences following an action, then those consequences are foreseeable (Fuscaldo, 2006). In the following, I shall concentrate on voluntariness first, and then consider foreseeability.

Foreseeability is considered not only in conceptions of ethical responsibility, but also in legal responsibility (it is part of the *Caparo* Criteria discussed in chapter 5). The following section explains why foreseeability is the key aspect of retrospective personal moral responsibility. In the following section I discuss potential scenarios where foreseeable erroneous AIS outputs could arise and consider how stakeholder responsibility could be assigned.

Voluntariness and Foreseeability

Firstly, to be morally responsible the actor must be acting act voluntarily, and they must be freely able to act otherwise. Fuscaldo (2006) acknowledges that the discussion around the freedom of an actor's actions is vast, but for the purposes of this thesis, let's assume that SDCs are not being forced to make AISs to be deployed for use in clinical settings, and they could choose not to. If SDCs are free to act, they meet the voluntary condition for moral responsibility. Similarly, the clinician appears free to not use the AIS which the SDC is offering, and so also meets this criterion. However, as we shall see later in this section, there seem to be factors that may restrict this freedom, particularly in light of possible patient demand and pressure to make use to all the tools that are at their disposal. We will come back to this.

Secondly, we can only be morally responsible for the consequences of those actions when the consequences were foreseeable; Fuscaldo (2006) defines foreseeability as consequences which a reasonable person may expect to occur.

These two conditions appear to be plausible and certainly seem intuitive. Allow me to outline a hypothetical example: the wrong patient's notes are reviewed in an outpatient clinic consisting of patients with similar characteristics (e.g., two female patients called Mary Jones) and homogeneous medical issues (e.g., hand eczema). The clinician might not recognise that the clinical record that they are now working with is for a different patient. Should an AIS be used to examine the patient's clinical record, the use of the incorrect information may well result in the wrong treatment recommendation being offered to the consulting clinician. If the clinician believes that they are viewing the correct clinical record, they may find that they draw the same conclusion that the AIS did. If neither the AIS nor the clinician somehow recognise that the wrong clinical record has been identified and the wrong treatment option is pursued, there is risk of a negative consequences for the patient in question; for example, a missed opportunity for treatment, or incorrect drug/dose prescribed. It is hard to view this as a foreseeable event when there is an existing process to correctly and routinely pair patients with their records and when that process has historically worked well at every prior appointment; however, it is not impossible for an honest mistake to be made. Thus, if there is an absence of foreseeability, it might be argued that just because an actor has causally contributed to a consequence does not necessarily mean that they are morally accountable for it (Fuscaldo, 2007). In this case, personal moral responsibility would not be allocated to the clinician or the SDC (unless they had been the ones to negligently incorrectly identify the patient)[34]

A weakness of Fuscaldo's account of moral responsibility is that it is assumed that the actor is a rational decision maker. Barret (2004) makes useful observations that add specificity to Fuscaldo's account in this regard, suggesting that "moral responsibility assumes a capacity for making rational decisions, which in turn justifies holding moral agents accountable for their actions." Let's break this claim into two parts.

Firstly, what makes a decision rational? Uzonwanne (2016, p.1) defines rational decision-making as "facts and information, analysis and a step-by-step procedure to come to a decision." Thus, a rational decision can be broken down by the decision maker to explain how that conclusion was reached. Reyna and Rivers (2008) identify reaching one's goals according to one's own values as the aim of a rational decision.[35] To build on this, let's lean on Rawls: to be rational, one logically acts for one's own interests; to be reasonable, one's actions are "fair-minded, judicious, and able to see other points of view" (Rawls, 2007, p.54). That's not to say that one's own goals must be to act selfishly; they could easily include or be directed towards helping others achieve their goals. That which contributes to a

[34] This indicates the potential for a responsibility gap, as who would then be burdened with responsibility for this consequence? This is a problem which society ought to engage with as the determination of the allocation of this burden reaches beyond the key stakeholders of the clinician and the SDC which this chapter is interested in. However, this issue is addressed by the solutions presented in chapter 7 where responsibility is *shared.*

[35] One's own goals do not have to act selfishly; they could easily include or be directed towards helping others achieve their goals.

rational decision may be in a fluid state; a Bayesian approach allows that a person's beliefs may be updated as more evidence is considered within a given context (Godfrey-Smith, 2003).

Secondly, what is it that makes one person morally responsible for another? As already identified above, personal moral responsibility is the individual's obligation or duty to ensure that something is acted or obtained (Zimmerman, 1992) and the obligation of one person to another would be at the very least to ensure that the first did not harm the second. Rousseau explains that there is a need for reciprocity in a just society (Rawls, 2007). In Rousseau's society which promotes valid reasoning, equality and social interdependence (Rawls, 2007), it is fair to argue that at the very least that one actor must not harm others.

Given that we have established that an SDC has both prospective and retrospective responsibility – just like the clinician – we can draw on the account of moral responsibility provided by Fuscaldo and Barret's observations to consider how those responsibilities might be fulfilled.

To take prospective responsibility first, an SDC is obliged to think about the consequences their AIS could generate. The SDC would need to rationally consider what is known about both possible and likely harmful consequences and break down how their decision to deploy their AIS was reached. They would then need to demonstrate that they had acted to mitigate the risks of harm that were identified.

Fuscaldo's condition of foreseeability is doing the key work here, and the overarching duty of care (prospective responsibility) requires the possible consequences of the AIS to be carefully considered by the SDC. In doing this, the SDC can show that they acted rationally, but in doing so they set up the conditions for meeting the foreseeability criterion. This makes them *prima facie* liable for retrospective moral responsibility and also allows them, in principle, to show they have taken action to mitigate the foreseeable risks, which may be enough to protect against liability.

If, in meeting their duty of prospective responsibility, the SDC shows that certain consequences are foreseeable, this then paves the way for them being retrospectively morally responsible if those foreseeable harms eventuate.

That duty, of course, must be discharged diligently and rationally. Even if it was not foreseen, the key question is whether or not it was reasonably foreseeable by a rational and diligent actor.

The existence of a prospective moral responsibility means that the SDC has a duty of care to mitigate against foreseeable risks of harm. One way a SDC may choose to do this is by warning the user of a risk of an output which could be harmful. This appears to be an attempt to shift the burden of personal moral responsibility on to the user. This might be legally allowable[36], but it does not seem entirely morally permissible given that it is reasonably foreseeable that users of an AIS are imperfect. As mentioned in chapter 5, an AIS might have a high frequency of accurate outputs which may lead a clinical user to suffer an 'atrophy of vigilance' (Freudenberg, 1992, p.19). A loss of attention when a clinician has found an AIS to have been historically trustworthy is understandable, but could easily lead to patient harm. It is fair to allocate personal moral responsibility for harmful consequences to a reckless clinician who uses a glaringly dangerous AIS recommendation, but when the AIS is incorrect and the mistake is more subtle then it's unfair to allocate all responsibility for consequences on the clinical user. In simply stating 'user beware' the SDCs may discharge some of their personal moral responsibility, but some responsibility would seem to remain due to their AIS's ongoing presence in the clinical decision-making space and that it is clearly foreseeable that clinicians could potentially unintentionally allow a harmful AIS output to reach a patient because of a built-up trust in the system's performance. If a system is designed to influence clinical decision-making, and it is foreseeable that an AIS may dispense a harmful output, and it is foreseeable that a clinician might unintentionally accept and use that output, then it is foreseeable that a chain of events could occur that would allow harm to reach the patient. Thus, a freely acting SDC discharging their prospective duty of care holds a proportion of moral responsibility for those foreseeable consequences - particularly if it does not include feasible safeguards against it. This does not relieve the clinician of their moral responsibility for harm which occurs, but it does increase the space for the SDC to join the clinician.

Careful use of an AIS application can be challenging when its processes are opaque to the user; for example, I noted in chapter 3 that in some applications of machine learning it is impossible to understand how the AIS reaches its conclusions. One issue with opaque AISs deployed for clinical decision-making is that the outputs are not always predictable and may change if the AIS learns with each new experience and then applies that learning to each subsequent use. This means that the AIS's outputs may not always be reasonably foreseeable, but they are *foreseeably unforeseeable* and therefore might offer a potential source of recommendations which could be harmful to patients.

The SDC could still, however, argue that they have designed the delivery of the AIS to include the requirement of a clinician to be present to manage outputs which are not foreseeable and possibly incorrect. Again, in doing so, they have handed over the personal moral responsibility for the consequences of using the AIS to the clinician who is using them for clinical decision-making and in so

[36] Applying *A v National Blood Authority and Worsley v Tambrands Ltd*, an AI system whose technologist has warned users that its outputs should not be solely relied upon which then delivers the wrong output to the user maybe considered a standard product as there is the risk of an erroneous output being given to any user, and that user had been warned.
This is product liability rather than negligence law, so this case was not discussed in chapter 5.

doing use the clinician as a moral buffer zone. But the fact that the AIS is opaque, and is provided to the clinician on the grounds that it is able to do things and reach conclusions that the clinician cannot, means that the clinician may not be in a position to take on that role of safety gatekeeper. The clinician could be told that both the AIS is a superior clinical decision-maker, and it is not possible to understand its superiority because its reasoning cannot be scrutinised. Under these circumstances, they are unable to evaluate an AIS's strengths and weaknesses, thus they are unable to safely weigh the value of its outputs. If the clinician is not able to evaluate the safety of the AIS's outputs and patient safety is at risk, the clinician would be justified in rejecting using the AIS at all; thereby fulfilling their duty of care to maintain patient safety. The SDC fails in their duty of care to the patient in this scenario if they cannot safely utilise its outputs, and the fallible clinician is put in a position where they are forced to underwrite the safe deployment and use of the AIS. Given this, the SDC might wish to withhold novel technologies until they are safe enough, rather than releasing an AIS which risks causing harm to patients and thereby damaging the image of the SDC, the AISs, and their users when they are still in their infancy.

It is not unreasonable for clinicians to wish to use AISs when it helps them to help their patients, but if the consequences of AIS use include inequal and burdensome allocation of personal moral responsibility, that might be enough reason for the clinical professions to reject AIS use. Prospective and retrospective responsibility can be allocated to SDCs, thus there is value in exploring how the resulting duty of care could be created at a professional level for SDCs. Equality in the recognition and carrying of stakeholder responsibility would go a long way to rectifying this unjust moral problem.

It would be useful to identify and apply the theory discussed above to scenarios of foreseeable harms that could take place if an erroneous output from an AIS reached a patient. The following section does just that and discusses the issues that arise with each example.

Practical application of ethical theory

During the process of researching and writing this thesis I have considered various potential negligent acts which may occur due to the use of foreseeable erroneous AIS outputs. I present these now to explore the practical application of the ethical theories this chapter has so far discussed. Whilst I recognise that this will not be a conclusive list of causes of erroneous AIS outputs, I shall limit discussion to exploring technologists distributing an AIS which is not perfectly accurate, user error (involuntary and voluntary), and atrophy of vigilance.

I will be drawing on an example of AIS which was identified in the literature review, IBM Watson for Oncology. I shall use this example to demonstrate how issues of justice, prospective and retrospective responsibility with the key aspects of foreseeability, and voluntariness arise and intertwine. Discussion of solutions to these issues shall be found in the next chapter.

Technologist’s deployment of an AIS which is not perfectly accurate

To recap from earlier in this chapter, an SDC is *de facto* causally responsible for harms caused by the use of their AIS’s outputs. They might also be personally morally responsible for harms arising from the use of their AIS. This personal moral responsibility is both prospective (they have a duty to ensure that they do not harm others, this implies a duty of care towards patients) and retrospective (when use of their AIS results in harm to a patient to whom they owe a duty of care). It is the SDC’s choice to deploy an AIS into the clinical environment and, should they choose to deploy, I suggest that they be prepared to hold a degree of personal moral responsibility for the consequences of that action; but the amount of personal moral responsibility held by them must be reasonable and appropriate. SDCs stand in a *de facto* causal relationship, certainly, but they would not be morally responsible for all harms that arise from their technology or its use. For example, if the AIS was used negligently by the clinician, or if it was intentionally given wrong information, and harm resulted to a patient, the responsibility might reasonably lie elsewhere.

The potential for SDC’s AISs to be inaccurate is a foreseeable issue but this has not been openly discussed by SDCs. Assuming the comment reported by Hengstler *et al* (2016, p.115) is representative of SDCs, there is evidence that SDCs will state that the clinician makes the final decision about any treatment that shall be delivered. This kind of position implicitly accepts that AISs may produce erroneous outcomes whilst also using the clinician as a ‘buffer zone’. Were a patient to be harmed due to the use of an erroneous AIS output, the SDC could argue that the clinician is the specialist holding the duty of care and that their duty to safeguard extends to the tools that the clinician used (e.g., the AIS). It is reasonable to assume that this is an attempt to use the clinician’s traditional allocation of professional and personal moral responsibility to underwrite their own activities. This approach apparently could work in negligence claims (as we saw in chapter 5), but it does not stand up under analysis using the theories of justice and moral responsibility which I have outlined in this chapter. I shall address justice below; the allocation of moral responsibility will be discussed in the sections which immediately follow.

As earlier mentioned, there has been no public discussion between the SDCs and the clinical users about how responsibility for the positive or negative outcomes of using an AIS will be distributed. If the AIS in question is useful, clinicians will likely feel pressure to use it, but justice demands that clinicians are involved in developing the rules around its use. This lack of negotiation amounts to a lack of social cooperation which is a necessary ingredient in a just society (Rawls, 2001, p.6). Social cooperation is arguably needed for the function of any society; the more cooperation there is, the better that society does. In taking this position, even if this position has been unconsciously taken, SDCs are not acting justly towards clinical users. The lack of opportunity to publicly negotiate their position or to formally accept or reject this burden represents an action of injustice of the SDCs onto the clinicians.

The easiest way for a clinician to object would be by simply not using the technology. However, if clinicians chose to reject using an AIS due to not wanting to hold sole moral responsibility for the consequences of using the SDC's AIS, the clinician may face pressures from patients, their employers, or the clinical system in which they worked to use the AIS offered. This pressure may encourage them to use the AIS at their own risk and to suffer the injustice of being held responsible should a patient be harmed as a result[37]. The injustice arguably does not lie in not being involved in the development and deployment of an AIS; the injustice lies in being excluded from discussions regarding who will be responsible for the AISs use and being given no choice but to take responsibility for the outcomes if they chose to use it.

To rectify this injustice, there must be discussion and negotiation between stakeholders of the rules by which the SDCs will deploy their AISs and the clinicians who will use them. When considering the inclusion of stakeholders in these deliberations, it is important to consider who will be involved: should all stakeholders be consulted? Or is a representative sample enough? How big should that representative sample be? Which stakeholder groups should make up that sample? There is an opportunity and a need here to consider the opinions of all affected stakeholder groups, not just clinicians and SDCs, for example: representative patient groups, clinical regulators, and relevant specialist bodies. The questions outlined above would be served well by a consultative qualitative research approach, and I outline the value and need for consultation prior to AIS adoption further in chapter 7. As noted in chapter 1, there is insufficient space to address all stakeholders in this thesis, but they must still be considered nonetheless.

User error

This chapter has argued that an SDC who has created an AIS is causally responsible for the consequences of the use of their AIS. The clinician joins the SDC in causal responsibility if they were the one who used the AIS's outputs to inform the patient's care.

Applying Fuscaldo (2006, p.71), who argued that "we are morally accountable for the intended and unintended reasonably foreseeable consequences of our free actions", it is reasonable to assume that

[37] I hope that my thesis may make some small contribution to help stop that from happening.

an SDC may foresee that there is a risk of a user using the AIS erroneously which may result in an AIS output which could be harmful if it reached the patient. (e.g., inaccurate data input leading to inaccurate data output). The clinical user could have used the AIS outputs erroneously either involuntarily, or voluntarily. I shall approach each of these separately.

Involuntary user error

According to Fuscaldo's two conditions, the clinical user is held personally morally responsible if their action was voluntary, whether intended or not.

The clinical user is causally responsible for their actions when using the AIS and has both prospective and retrospective moral responsibility towards the patient – both professionally and personally. The SDC would likely not be present in the clinical environment and thus unable to intervene to prevent the clinician from erroneously using an AIS output which causes harm to the patient. Even if the SDC had been present, they may well have not had the expertise to know the output could be harmful.

But the risk of user error should certainly be foreseeable to the SDC. As such, the SDC would be acting irrationally if they did not institute reasonable safeguards[38] to prevent user error. Of course, safeguards are not fool-proof and cannot offer guarantees. But an SDC is acting in a morally responsible way if they have placed sufficient safeguards which protect harmful AIS outputs from reaching the patient. If the safeguards were found to be inadequate and a patient was harmed as a result, it would be a breach in the SDC's duty of care for them not to reconsider their safeguarding to prevent similar harm from eventuating again. If the safeguards were foreseeably inadequate, the SDC would be morally responsible of the harm, at least in part.

Thus, the clinician would be personally morally responsible for their part in using an AIS output which caused harm to their patient. The SDC would also be personally morally responsible if they had failed to install reasonable safeguards to guard against this foreseeable user error.

Given that there are innumerous combinations of actions which could result in involuntary user error, it is not possible to create a simple rule by which to treat errors; each erroneous event would need to be carefully considered and the contributions of each actor carefully weighed; for example, what if the user was drunk? Here, it would be unfair for the SDC to assume any responsibility for harms caused if the clinician had made the final decision to use those harmful AIS outputs.

[38] I am going to refer the reader back now to an earlier footnote about "safe". "Safeguards" is as ambiguous as saying "safe". In the same way to when I refer to an AI system being safe, I mean that it is "safe enough" to use without unduly risking patient harm, when I refer to a system as having "safeguards", I mean that reasonable safeguards have been considered and negotiated between regulatory, professional, and lay stakeholders to ensure that all agree as an appropriate level of protection against harm from arising and that those safeguards have subsequently installed into the AI system.

The clinician should carry their fair share of personal moral responsibility, but the SDC must also carry their fair share if the AIS has been a factor in eventuating harms.

Atrophy of vigilance

This thesis's focus has so far been on the potential of the risk of patient harm eventuating from erroneous AIS output; but what if the AIS is initially so successful that people start to trust and rely upon it too much? Even when an AIS is known to have flaws, and the user has been warned that it cannot be 100% relied upon, if experience suggests it is reliable then a user may begin to place more trust in it than is warranted.

Trust can be problematic as it is a fallible state which reflects the trustee's confidence in the positive outcome for the trustor (Holmes and Rempel, 1989). When the trustor adopts trust (consciously or unconsciously), their critical analysis of the service which the trustee provides may be affected. During the initial period of use when an AIS is first released, it may be under intense scrutiny and may be challenged frequently. Once this period has passed, so long as the AIS performs well and consistently, the user may acquire a degree of confidence in the AIS. Were the AIS not being used, the clinician would have continued to conscientiously make their own decisions independently. The user may unconsciously start to apply Hume's uniformity of nature principle where it is assumed that "the future will resemble the past" (2014, p.37). This is problematic as the user may start to assume that if the AIS has worked well in the past, then it will continue to work well and may not need to be supervised as closely as was initially thought. What follows is 'atrophy of vigilance'; which describes how confidence increases over time and thus attention wanes, even though the overall risk has remained the same (Freudenburg, 1992).

Atrophy of vigilance is a foreseeable risk when deploying AISs for clinical decision-making. If the clinician is to safeguard the patient by ensuring that the AIS's output selected does not cause harm, that safeguard will be affected if the clinician's vigilance deteriorates.

Certainly, there is a prospective personal moral responsibility upon the clinician; they must remember to remain critical of an AIS's outputs to protect their patient from possibly erroneous outputs. As things stand, the clinician is the one professionally charged with judging the worthiness of an AIS's outputs and, due to their specialist knowledge, they possess the full foresight of the positive and negative implications of acting on the output provided by the AIS. If they made a mistake and the patient's safety is compromised, then they may have to explain to their professional body why they had failed to keep their patient safe. The clinician would be expected to maintain their attention and not allow their vigilance to wane over time. If their vigilance did wane, the clinician would have failed in their prospective responsibility.[39] As earlier identified, this failure is recognised in law as the tort of negligence.

Ethically though, it is not clear that it is fair to solely condemn the clinician if they had developed atrophy of vigilance in an AIS they had learned to trust. An AIS which is designed to help with clinical decision-making is, by its nature, created to influence the user's thinking processes and is designed to be reliable and accurate. Notwithstanding the clinician's duty of care, atrophy of vigilance may be unconscious, unintentional and happen gradually over time, and clearly counts as foreseeable for the SDC, and is therefore something they have a prospective moral responsibility to guard against (just as the clinician does). If it is foreseeable, if would also mean they have retrospective responsibility for harms if they fail to guard against it.

Voluntary user error

Malicious or fraudulent use of the AIS is another potential source of erroneous AIS outputs which could lead to patient harm. The main difference between this and simple user error is in the intent of the user whose choice to use the AIS's outputs has resulted in patient harm eventuating.

If the clinical user (who is causally responsible due to using the AIS and who holds a prospective and retrospective personal moral responsibility towards the patient) intentionally and maliciously chose their course of action (e.g., intentionally inputting inaccurate data into an AI powered system then using the system's outputs) they would have not acted in a morally responsible fashion, thus the responsibility for harms caused would fall to them.

The majority of the burden of personal moral responsibility would fall on the clinical user in this scenario as their intentional action would have been the prime cause of the foreseeable harm which eventuated. It is reasonable to say that it is not fair for the technologist to be held completely responsible for the free behaviour of others. But this does not mean that the SDC would not be completely free of any personal moral responsibility in this scenario. In the same way that the

[39] This principle stands true for other used of AISs away from healthcare: Uber settled out of court when one of their car's safety drivers failed to intervene when their vehicle hit a pedestrian (Woodhall, 2018). It appears that Uber had recognised their prospective moral responsibility; that they had a duty of care to other persons. This was demonstrated by Uber employing safety drivers as they had foreseen that their vehicles were not perfect and that they required supervision to be safely on the roads. Uber's employee had been delegated personal moral responsibility for the vehicle's safety, the employee had become distracted, and a third party was injured as a result. By making the pay-out, Uber re-claimed retrospective personal moral responsibility as the vehicle would not have been on the road if they had not paid the employee to do so.
Tesla has made no such settlements when one of their cars crashed. The car's owner/driver had been warned by the vehicle to pay attention when it was in Autopilot mode and therefore it is the driver's own fault if they were harmed as a result of not doing so (Lambert, 2018). If anything, the Tesla driver had a prospective moral responsibility to all other road users: they should have been paying attention to prevent their car from injuring others (no others were hurt).

aforementioned involuntary user error is a foreseeable risk, there could be a foreseeable risk that the clinical user would cause a voluntary user error. Again, here the SDC would also be personally morally responsible if they had failed to install reasonable safeguards which attempted to prevent foreseeable actions from the clinical user. This precaution could be comparable to mechanisms in other products which prevent malicious acts, e.g., food jars being fitted with tamper-proof vacuum seal button in their lids, or for vehicles to come supplied with a door locking mechanism. If an attempt had been made by SDCs to prevent malicious activity, then they have discharged their duty of care to attempt to prevent harm befalling those affected by AIS use. Those who intentionally circumvent security measures are allocated blame for harms caused as a result.

Having applied theories of justice and moral responsibility to the scenarios which I have identified in this thesis, I shall now introduce a concept that helps the reader picture further why stakeholder's being placed in a position of moral responsibility by others is problematic.

Identifying the 'moral crumple zone'

As earlier noted, clinical decision-making has historically been the domain of clinicians. SDCs are attempting to enter this space by offering their AISs to be used in healthcare. The clinician shall always be needed in the healthcare environment to supervise any AIS adopted unless an AIS is perfectly accurate, designed, and approved to work independently of direct clinical supervision. Until that level of AIS is developed, clinicians shall remain in the decision-making loop. I have already shown how the interwoven interests of clinicians and SDCs affect each other, but one matter highlights itself as an obvious, yet until recently unarticulated, issue in the literature: the moral crumple zone.

Elish (2019) describes humans who absorb the legal and moral penalties when a computerised system fails as a 'moral crumple zone'. She describes this as occurring when "responsibility for an action may be misattributed to a human actor who had limited control over the behavior [sic] of an automated or autonomous system" and compares this to the crumple zone safety measures built into vehicles to protect occupants during a crash (Elish, 2019, p.40).

This analogy may be applied to this thesis's subject matter; in the same way that a car will absorb an impact, a clinician who uses an AIS for a patient would absorb the moral and legal consequences should the patient be harmed as a result of using the AIS's outputs. The SDC knows that there is a risk that their AIS could cause harm to the target patient group and attempts to manipulate stakeholders so that responsibility for that harm is deflected away from them, by simply declaring that the AIS does not make decisions.

In the case of IBM Watson, the AIS has been designed so that IBM is protected by its clinical user because the clinician makes the final decision about any treatment that shall be delivered (Hengstler *et al*, 2016, p.115). This chapter has shown that it is unfair for a clinician to solely carry responsibility for the consequences of using an AIS. Therefore, should a patient be harmed due to use of erroneous AIS outputs by clinicians, the clinicians might be considered as second victims. The effects of this are not insignificant; Wu (2000) describes a "second victim" as a clinician suffering from rumination, fear, and guilt after an erroneous action which could lead to burn out, use of substances to cope and potentially loss of the clinician from their profession.[40]

Deflection of personal moral responsibility is unfair (and therefore unjust) as it might deny the patient an opportunity for a safer human-generated decision. It also it allows the deflector to profit from deploying an AIS which may cause harm whilst deflecting the negative consequences of their AISs actions. Had they not made and released the AIS, that risk of harm would not have existed, but regardless of whether the harm happened or not, neither the SDC nor the organisation is prevented from accruing positive benefits from their work (e.g., financial profit). As previously mentioned *per* Keating (1995), and employing Rawls's veil of ignorance, if the SDC and the clinician were to swap places in this scenario, the SDC may not wish to carry personal moral responsibility in the same way that the clinician is currently being asked to do. To not acknowledge this, and not act to solve the problem, could be considered socially uncooperative as per the second of Rawls's three essential features of social cooperation (2001); that rules are fair when everyone accepts them. Creation of a cooperative environment of mutuality or reciprocity allows participants to benefit when standards are publicly agreed. As the aims and interests of the SDC and the clinician cannot be reconciled when one is using the other as a moral crumple zone, the use of such tactics seems unjust.

How might the legal position be challenged using ethical theory?

I have established in this chapter that there is personal moral responsibility owed by SDCs and clinicians; that it is possible for each actor to be held causally as well as prospectively and retrospectively morally responsible for their part of the consequences of the development, deployment, and use of their AIS's outputs. I have not argued that the clinician should be absolved of all personal moral responsibility when using an AIS's outputs as they have a historic and justified duty

[40] Clarkson et al (2019) claim that allowing the use of the term "second victim" allows for the clinician to become passive and responsibility for actions to not be taken; they fear that this term undermines patient safety initiatives which can reduce incidence of overall harm caused to patients as the clinician is obscured as an agent of harm. This is refuted by Gómez-Durán et al (2019) who argue that the clinician's wellbeing is crucial for patient care and that the psychological distress caused by the "second victim" phenomenon affects both clinicians and their patients. Gómez-Durán et al (2019) agree with Wu (2001) that clinicians, patients, institutions, and advocacy organisations should collaborate to address and improve the approach towards medical errors and patient safety.

of care to their patients, but I have argued that an SDC should not necessarily be able to refuse to take personal moral responsibility should patient harm eventuate.

The circumstances of a patient's harm should be examined carefully for evidence of causal responsibility. But should an actor (SDC or clinician) be found causally responsible, this does not necessarily mean that they are morally responsible; this needs to be ascertained by determining if the actor had a duty of care to the harmed individual, if the actor's actions were voluntary, and if the harmful consequences of the actor's acts were foreseeable (as per Zimmerman's prospective model of responsibility, and Fuscaldo's conditions of retrospective responsibility – both of which were discussed earlier in this chapter).

This thesis has discussed how clinicians and SDCs are both eligible to hold a legal and ethical duty of care for the consequences of their actions which affect a patient. Clinicians are used to carrying responsibility for this duty of care due to the implementation and enforcement of their professional codes of conduct by their regulators. This is contrary to SDCs who are not professionally regulated. The duty of care between the SDC and the patient is recognisable and arguable in ethical theory; the existence of a prospective moral responsibility means that the SDC, as a rational actor, has a duty of care to mitigate against the evaluation of foreseeable risks of harm. As outlined in the preceding chapter, a duty of care arises in negligence law when harm is reasonably foreseeable, that the defendant and the claimant have a proximal relationship, and that it is fair, just, and reasonable for liability to be imposed.

However, no technology companies delivering AISs to the clinical environment have formally claimed to have developed and deployed an AIS which possessed the "standard of the ordinary skilled man exercising and professing to have that special skill." (*Bolam v Friern Hospital Management Committee*); thus, it is understandable that the user of their AISs is told that they cannot rely upon any of the AIS's outputs, and the user is the one that makes the final decision about to what extent that follow, or deviate from, the AIS's recommendation. Yet, it is reasonable to assume that an SDC would be able to envisage the impact of their AIS on the clinical decision-making environment as their AIS has been designed to directly influence the key decision maker - the clinician. Thus, it would be foreseeable that an AIS would influence patient care and that there would be direct practical effects from this voluntarily and intentional influence as that is exactly what the AIS has been designed to do by the SDC. My ethical analysis above supports this, provided that the actor was rational and free to act otherwise (i.e., the SDC could have chosen to not develop and deploy the AIS).

An SDC may attempt to argue that they were not at the bedside when their AIS was used, therefore were too distant from any harms caused due to the use of their AIS and would not have been positioned to intervene to prevent harms. Legal proximity between the technologist and patient harm is interrupted thanks to *novus actus interveniens*. The clinician's intervening act to take an AIS's output and to then employ it in their clinical decision-making might be found by courts to be sufficient to obliterate the act of the SDC supplying an AIS which provided inappropriate outputs for the patient. SDCs are thus legally shielded by clinicians because proximity has been given a practical limitation in legal applications. *Novus actus interveniens* is somewhat sensible as it prevents unmanageable chains of claims; for example: from the patient to the clinician, to the SDC, to the computer company who sold the technologist the laptop which they coded the AIS on, to the mine which provided the materials which the laptop was crafted from, or maybe to the university who taught the technologist to write software.

However, as noted in chapter 5, a legal duty of care needs to be fair, just, and reasonable. This requirement does support the use of *novus actus interveniens* as it would not be fair just or reasonable to hold an agent legally responsible when there was such significant distance between the action and the effect that the action could not have been reasonably associated with the effect. But, regarding clinical use of AIS's it is valid to ethically challenge the legal limitation of proximity employed by *novus actus interveniens*: there is arguably some degree of voluntary and intentional proximity between an SDC and a patient, as the SDC has voluntarily and intentionally designed their AIS to directly influence a clinician's decision-making: i.e., to directly influence the clinician's decision-making resulting in a desired effect on a patient. It is not unfair to position the SDC's actions of providing the AIS as a causal link between harm eventuating to the patient and the SDC as the SDC's acts were foreseeable; both voluntarily and rationally. Yet, the SDC is legally buffered against the risk of a negligence claim by the presence and actions of the clinical user. The clinician is used as a buffer to the SDC in the eventuality of their AIS giving advice which is harmfully inappropriate for the patient in question: this is ethically unfair and unjust to the clinician who then is burdened fully with the patient's negligence claim with no option to share that burden.

Conclusion

At the start of this chapter, I asked: how can ethical responsibility for the consequences of the use of AIS in clinical decision-making be determined and allocated?

Theories of ethics regarding the allocation of responsibility were explored as relevant to this thesis's concerns. Rawls's contractarian approach was noted to be potentially useful in achieving social co-operation between stakeholders. This theory facilitates fairness by having actors consider the effects of actions upon others. Causation was explored and I found that SDCs and clinicians can both be prospectively and retrospectively responsible for their respective deployment and use of AISs. Whilst the SDC is not at the patient's bedside when the outputs of their AIS is used, applying the theory of prospective moral responsibility means that the SDC has a duty of care to patients to mitigate against the evaluation of foreseeable risks of harm.

Ethical analysis can attribute responsibility to both SDCs and clinicians for patient harms which could eventuate due to the use of AISs. This has allowed me to argue that there is an imbalance in legal responsibility when its allocation is considered through an ethical lens. The clinician's legal burden is unjust as recognised under the egalitarian approach of ethical theory because if the clinician is forced to solely carry the legal burden in a negligence claim, they are being treated unequally to the SDC who does not have to carry any of that burden despite their contribution to the harm caused to the patient.

To begin to ethically address the clinician's unjust legal burden, there could be value in using the communitarian approach of stakeholder discussion and collaboration. This would allow the exploration and evaluation of the problem of the allocation burden of ethical and legal responsibility when AISs are used in clinical decision-making. Through stakeholder discussion, the clinician as a disadvantaged stakeholder may be formally identified, a joint and rational plan may be worked upon, and a fair social contract established (*a la* Rawls) which would allow the just allocation of the practical legal and ethical burdens of responsibility.

An opportunity could be taken to incorporate practical discussions of how stakeholders can work more closely together to prospectively prevent harms from eventuating in the first place as well as planning how problems will be addressed retrospectively should they happen, e.g., they could discuss issues such as the calculation and distribution of specific proportions of responsibility to each actor. The calculations of proportions of responsibility are distinctly different from the factual determination of causal and moral responsibility. I have not attempted to create a formula which could be used to solve how much moral responsibility is carried by each stakeholder in a given scenario as the scenarios are numerous and complex, the discussion required places this problem firmly outside of this thesis's limits.

The negotiations between stakeholders could result in any arrangement of the distribution of responsibility. At this point, it is reasonable to ask: what could be a fair balance of responsibility between the clinician and the SDC when AISs are used in clinical decision-making?

The next chapter proposes a possible practical solution to aid the *sharing* of the ethical and legal burden of responsibility for the use of AI in clinical decision-making between clinicians and SDCs.

Chapter 7: Solutions

Before launching into possible solutions, I will briefly recap the last couple of chapters to help re-orientate the reader. Chapter 5 speculated how legal responsibility could be assigned using the tort of negligence should a patient be harmed as a consequence of using a AIS to inform the clinical user's decision-making. This legal analysis indicated that a negligence claim might be successful against an SDC as well as a clinical user, but that the claim against the clinical user appears to be stronger due to *novus actus interveniens*. The last chapter 6 used ethical theories to posit that it would be unfair for a clinician to solely carry responsibility for the consequences of using an AIS. Here, it was argued that, along with clinical users, the SDCs who develop and deploy AISs may be allocated a prospective duty of care for the effects of the use of their technology and that they may be assigned retrospective responsibility for the foreseeable effects of the use of their AISs upon patients. *How* that responsibility could be expressed has yet to be clearly and reasonably identified.

This 7th chapter leads on from chapter 6 by offering a solution to the following question: how could the allocation of responsibility be fairly balanced between the clinician and the SDC when AISs are used in clinical decision-making?

In this chapter, I shall explore potential practical routes which would allow and might encourage SDCs to embrace their prospective and retrospective responsibility to patients. These routes aim to benefit patients by allowing actors to be held responsible for their actions, whilst restoring fairness to clinical users, but without reducing the clinician's own duty of care to their patients. The initial discussion shall touch on why prospective and retrospective solutions applied separately to distinct stakeholder groups would be insufficient; the remaining bulk of the discussion will be committed to exploring an idea identified in this thesis's literature review: Whitby (Whitby, 2015) stated that responsibility could be *shared* between clinicians and SDCs should there be negative consequences to AIS use. To investigate this idea, rather than solely promoting the allocation of responsibility to individual actors or distinct stakeholder groups, potential routes to the fair sharing of responsibility for AIS use between the SDC and the clinical user will be probed. This shared model of responsibility will provide a basis for the practical contractarian-based solution of risk pooling where both clinicians and SDCs might work together to hold personal moral responsibility in a manner which is fair and protective to all stakeholders.

Current practical models of responsibility are inadequate

Zimmerman's (1992) model of personal moral responsibility identifies prospective and retrospective approaches, and it seems that prospective and retrospective approaches for managing the use of AIS are being developed or are currently in place. The following discussion shows how separate elements of prospective and retrospective responsibility are identified when individual actors are regulated, and retrospective responsibility is identified as negligence claims after a harm has occurred.

Regulating stakeholders

Mandatory and enforceable codes of professional conduct are modelled by the clinical professions (the wider question about how effective the current clinical professional regulatory system is will not be examined here). Their respective codes allow members to know what is required of them and promote uniformity of professional approach when they practice.

Chapter 4's literature review noted an absence of authoritative codes of conduct for SDCs and technologists. The question was raised of whether there ought to be a requirement for technologists and/or the SDCs that employ them, who create and deploy AIS to be used in clinical decision-making, to be regulated in a similar fashion to clinicians. Without authoritative codes of conduct, SDCs and technologists are left lacking in guidance in two ways. Firstly, they are without either an implied or specifically formalised recognised duty of care for those whom their AISs affect. Secondly, they lack standardised and enforceable codes of professional conduct (that their clinical counterparts benefit from) by which professional standards may be measured. Thus, if a patient comes to harm due to the use of an AIS in clinical decision-making, the courts considering a subsequent legal claim when referring to a *Bolam* standard of negligence will not have such standards available to them to compare an SDC's or a technologist's actions to (as discussed in chapter 5's legal analysis). For this reason, when presented with a negligence claim, the courts might draw their own conclusions of what the standard is as it relates to this stakeholder group.

From an ethical perspective, whilst encouraging registrants to achieve Zimmerman's prospective personal moral responsibility is not the specified aim of a professional regulated body, their codes of conduct happen to discharge this function. The clinical regulators encourage registrants to look forwards in their practice and consider the consequences of their actions in their daily practice. As identified in chapter 5, a duty of care is owed by actors (all actors, not just clinicians); persons cannot act with impunity, expecting that liability will rest with another party. But, if an assigned duty of care (be it assigned ethically or legally) is an example of prospective moral responsibility, it is interesting to note that the duty of care which a clinician owes to their patient is not explicitly mentioned in any of the clinical regulators' codes of conduct. Despite not being specifically spelt out, the clinical codes of conduct do embody the spirit of the duty of care which a registrant owes to their patient. For example:

- "You must be competent in all aspects of your work" (General Medical Council, 2020, p.6) without competence, a registrant risks harming the patient.

- "Identify and minimise risk" (Health & Care Professions Council, 2016): if sources of potential harm are not actively identified and mitigated then there is potential for harm to befall others.
- "Act without delay if you believe that there is a risk to patient safety or public protection" (Nursing and Midwifery Council, 2018): if a risk is identified then it must be addressed; this makes the care of all patients every clinician's responsibility- even if they are not specifically caring for the patient or patient group who is at risk.

Clinical registrants are not directly instructed to observe their duty of care, but their respective codes of professional conduct guide them to do just that; thus, prospective moral responsibility is enacted by the clinical professions due to their following of their codes of conduct. If a registrant's actions are investigated by their professional regulator and are found to be contrary to their code of conduct, the regulator possesses the aforementioned power to put conditions on a registrant's practice, suspend or strike them from the register; thus preventing them from practicing as a clinical professional in that role. Due to professional regulation, the clinician is forced to confront their prospective personal moral responsibility via their duty of care. If the clinician makes a mistake which is reported to their regulator and a penalty is decided by their regulator, they must face and bear their retrospective personal moral responsibility through complying with the corrective sanctions to restore confidence in their practice or being prevented from further working in their profession. This model rightly encourages introspection of a clinician to evaluate and maintain standards within their own practice.

Clinical codes of conduct instruct registrants to work cooperatively (Nursing and Midwifery Council, 2018) and collaboratively (General Medical Council, 2020) with colleagues, but when other actors outside of the clinical professions (e.g., SDCs) wish to affect care decisions there are no instructions on how clinicians might manage dealing with an unregulated external influence. Given that the clinical professions have historically practiced mainly with other similarly regulated clinical professions, these codes of conduct might encourage unthinking collaborations with new colleagues in the clinical environment which turn out to be problematic. If AISs are to be introduced to the clinical area, clinicians might accept these new and non-traditional actors (i.e. SDCs and technologists) and allow them to start to influence their practice, whilst failing to recognise that these actors are not comparably regulated. If this is not addressed and SDCs/technologists and clinicians continue to not be similarly overseen, the lack of parity in professional regulation could be problematic: as noted in chapter 6, SDCs could be found jointly causally responsible for the outcomes of a clinician using an AIS, but would not face the same consequences that the clinician would. In light of this lack of fairness, clinicians might reject using the AIS completely rather than risking the potential of being used as not only as a "moral crumple zone" (Elish, 2019) by being the only actors who are answerable for the effects of using the technology, but also as a legal crumple zone (as described in chapter 5) should harm occur and a claim be made.

Whitby (Whitby, 2015) confirms that the information technology industry is unfamiliar with the need to adhere to strict professional and ethical standards and codes, but it is not as though these have not been available. There has been a plethora of guidance concerning artificial intelligence released by governmental and non-governmental organisations worldwide (Algorithm Watch, 2020), yet these are neither universal nor binding. Indeed, the ACM's code of ethics (Association for Computing Machinery, 2018) is not entirely dissimilar to that of the clinical professions. Whilst it has been recently updated, it is by no means compulsory for non-members. It could be appropriate to use (Hao, 27 December 2019) description of ethics washing here: "where genuine action gets replaced by superficial promises." Voluntary codes of conduct lack the force of their mandatory counterparts; without authoritative force there is the risk of the desired actors following that code either inconsistently or not at all. Additionally, there is the risk of little or no public involvement in standard setting and the monitoring of an SDC's compliance would be done internally with little if any transparency (WHO 2021). If a 'good' code of conduct exists, but is not widely adopted, the endeavour risks being all talk and no action. If it is desirable for a formalised prospective duty of care to be introduced to SDCs who wish to develop and deploy AISs to affect decision making in the clinical environment, there is value in exploring how codes of conduct which constitute a duty of care and possess authoritative force could be adopted.

This might be achieved by placing individual technologists on more equal professional regulatory footing to the clinical professions. For example, the Health and Care Professions Council (HCPC) could expand their scope to adopt the regulation of technologists who work with healthcare applications alongside their highly varied cohort of fifteen clinical professions (HCPC, 2021), thus cementing this group's duty of care to patients. But this approach might result in 'shoehorning' technologists into the clinical professions; this may be inappropriate as their proposed role in healthcare is not patient-facing. If technologists take ownership of this problem and organise their own regulatory body, they then gain the advantage of becoming involved in the solution. In this way they may present and negotiate with their peers to help determine and set the standards and codes of practice to which this stakeholder group must work, rather than have that standard determined for them and imposed by others, e.g., by an external group such as the HCPC or via court actions if a negligence claim is made.

If authoritative regulation is achieved, an attempt could be made to regulate individual technologists with the aim of a mandatory code of conduct; this would facilitate their conscientious adoption of prospective personal moral responsibility. In this vein, it is welcome news that the Royal Statistical

Society is developing accreditation and preparing industry-wide professional standards for data science (2020). This first step will, hopefully, be followed by other initiatives. Ideally there would be a branch of this work specific to AISs developed for clinical development, deployment, and use; such specification would be useful due to the unique reach and impact which the manipulation of clinical decision-making via AISs would have on patient care. Indeed, developments could even encourage the clinical professions to consider their relationship with AISs and technologists and thus accordingly update and compliment their own standards and codes of conduct.

But, even if such codes of conduct were adopted by individual technologist registrants, it would not change the fact that it is incredibly unlikely that a technologist would create and deploy an AIS alone. This thesis's literature review noted Nissenbaum's (Nissenbaum, 1996) 'problem of many hands'; that the identification of precisely *who* has made *what* contribution to a project is obscured where there are several people involved. Determining the root cause of, and then holding individuals responsible for, specific outcomes for the use of an AIS could be an unfathomably difficult job. For this reason, it might be more realistic to regulate and allocate responsibility to the SDC that employs the technologists rather than individual technologists who have worked in a team to develop an AIS. However, regulation only of SDCs rather than individual technologists would not prevent individual technologists from practicing if their contribution to an AIS made it unsafe to the point of it risking or leading to the harm of patients. It needs to be possible to hold individuals to account where appropriate and to intervene before a foreseeable risk from a practitioner's ineptitude eventuates in harm (for example by the practitioner's regulator either prescribing additional supervision or training or by preventing further practice).

Even if SDCs or technologists were discretely regulated, the key difference between both of these stakeholders and their clinical professional counterparts remains: SDCs are remote from decisions at the patient's bedside rather than actively involved in every patient case which their AIS influences. Without the direct effect of the SDC's AIS upon a patient, there is no specific action for the regulator to intervene on; as per this thesis's earlier legal analysis, the *novus actus interveniens* of the clinician making the final bedside decision strikes again.

SDCs, individual technologists, and clinicians carry out different actions whilst performing different roles; each role permits differing levels of proximity to the patient during that decision-making process. Because of this, these key actors are treated separately and differently. This different treatment is notable because they are subjected to different levels of being held prospectively and retrospectively responsible for their actions, despite their both making contributions to the decision which is made at the patient's bedside. The separate and different treatment of these actors when they have contributed to the same event creates a divided and unsatisfying response to the allocation of responsibility, especially as clinicians are positioned to carry the majority of the burden of the effects of the use of the AIS.

Because clinicians are positioned as moral crumple zones (Elish, 2019) and due to the problem of many hands (Nissenbaum, 1996), the regulation of SDCs and/or technologists in a similar fashion to clinicians might not solve the problem of ensuring that actors are held fairly to account. For this reason, a different approach is needed.

However, rather than regulating the SDCs or technologists, the current work in regulatory practices is concerned with regulating the AISs themselves. This thesis is specifically concerned with how responsibility may be allocated to stakeholders and has purposefully avoided discussion of strict liability and regulation of AISs. However, it is impossible to further the discussion without referring to these current developments. Thus, I shall touch on them only-so-far as is necessary to demonstrate that current practical models of responsibility are inadequate.

Regulating the AIS

A 'multiagency advice service for AI technologies in health and social care' project was formally launched on 24 September 2020 (NHSX AI Lab) and is funded by the Department of Health and Social Care's organisation for digital transformation in the NHS: NHSX (Department of Health and Social Care, 2019). This service's launch outlined the aims and initial construct of the service: it will offer support, information and advice on regulation and health technology assessment for artificial intelligence in health and care. It will be administered by a core team from NICE and involve the MHRA, NICE, Health Research Authority (HRA), and the Care Quality Commission (CQC). Its service will dispense a single point for coordinated advice for AIS innovators to understand how to meet medical device regulatory requirements and generate the evidence requirements needed for an AIS to show that it is effective, safe, and cost effective to be used in the NHS. Whilst the involvement of the clinical professions who will be using the technology proposed appears to have been neglected, this service otherwise appears to be an intelligent approach to unify an otherwise highly fractured regulatory environment.

This multiagency advice service is in the embryonic stages of development and stakeholder engagement will give those affected the opportunity to express what they will need out of such a service to enable AISs to reach the clinical environment. It appears to have the beginnings of a framework whereby SDCs would be encouraged to adopt prospective moral responsibility via following regulations forcing them take care in the development of their AISs before their AIS may be permitted for deployment in the clinical environment.

Whilst a product regulatory approach is not a prospective ethical code of conduct, it may effect the same result for those stakeholders who would be affected by an AIS's deployment. The net result would be 1) that SDCs would be required to demonstrate that they had taken care when they had developed the AIS for deployment, and 2) that clinicians would continue to be separately regulated as usual and may be penalised if they failed to take care and minimise risks in clinical practice when utilising AIS. Whilst not all jointly coordinated, each actor within this constellation of regulation would share the same aim as the others: to reduce the risk borne by patients. Whilst this would not ensure that individual actors are held retrospectively responsible for their actions by their respective regulators, it would create an environment whereby actors would be unable to deploy or use an AIS without prospectively demonstrating the safety of the AIS *prior* to its use.

The multiagency advice service would not necessarily be alone in this work. The recent Cumberlege Review (Cumberlege, 2020) reported on the safety of medicines and medical devices. Whilst its focus was on devices and medicines (namely hormone pregnancy tests, sodium valporate use in pregnancy, and pelvic mesh) and is not specifically AIS related, this review recommended an independent Redress Agency for those harmed by medicines and medical devices. AISs for clinical decision making are medical devices, therefore such an agency would find these systems within its remit. This agency would use "a non-adversarial process with determinations based on avoidable harm looking at systemic failings, rather than blaming individuals".

Unfortunately, the Government has no current plans to establish the Redress Agency, the reason being that other redress schemes have been previously established without the need for the creation of another agency (Dorries, 2021), but the new Medicines and Medical Devices Act 2021 has made provision for a Patient Safety Commissioner. This will be a statutory role whereby the commissioner shall "promote the safety of patients and the importance of the views of patients in relation to medicines and medical devices" (Department of Health and Social Care, 2021c). As an independent advocate with powers and functions, it is hoped that the Commissioner will be a beacon for listening to and reflecting patient safety concerns (Dorries, 2021) thus containing the spirit of the Cumberlege Review's recommendations of a voice which speaks and acts from the patient's perspective to hold the system to account (Cumberlege, 2020).

As the new Patient Safety Commissioner will be able to make reports and recommendations to the healthcare sector (both NHS and independent) (Dorries, 2021) it is not impossible that they could potentially work alongside the proposed multiagency advice service. The multiagency advice service would determine what had happened, and the patient's voice in the investigative process could be supported by advocacy from the Patient Safety Commissioner.

If the new multiagency advice service finds that there are issues with an AIS which threatens the fundamental safety of patients, ideally, the body ought to have the ability to arrange with the appropriate regulator (e.g., MHRA if a device issue, CQC or the clinical regulators if a user issue) for an AIS to be suspended from use, temporarily or permanently as appropriate, if it is found to be sub-standard or misused until such a time that the root cause of the problem is rectified.

It is not implausible to suggest that errors will exist in medical systems deployed into service (Whitby, 2015). If it is accepted that errors may result due to using AISs in clinical decision making, then it is unfair on patients for there to be no preparation for the eventuation of that risk. The measured potential for that risk will only be calculated through rigorous testing of the system prior to deployment in the clinical area; a system's risk profile and the potential cost of failure depends upon factors such as what the system is, how it is deployed, and how it is used. However, there is no way to exhaustively test an AIS prior to its deployment at the bedside (Whitby, 2015). This risk may reach and threaten the patient via erroneous AIS outputs mixed with a clinical user's atrophy of vigilance. Because of this risk, it is only fair to patients for the compensation of that risk to be planned for. The addressing and dispensing of patient compensation via the use retrospective approaches of responsibility are discussed next.

Retrospective negligence claims

A retrospective approach to personal moral responsibility would have an actor face the actual (rather than possible or potential) consequences of their actions. This approach may be practically expressed in legal claims and has already been explored in depth in the legal negligence discussion in chapter 5. It is important that a process is available for patients to seek redress should they be injured due to the use of AIS in their clinician's decision-making, however, it appears that a negligence claim is the only option which patients currently have. Yet, prospective and retrospective models of responsibility can be problematic when they are considered in isolation from each other. The issue is visible when considering potential retrospective negligence claims by patients for harms caused to them by the use of AISs.

In these retrospective claims, the activities of identified individuals are scrutinised and penalised rather than the consideration of collective actions which have resulted in the harm caused. This individualistic focus creates a fractured approach to personal moral responsibility: many people may have been involved in the creation, deployment, and use of AISs and, as we have seen, this factor makes it challenging to assign moral and legal responsibility for consequences. The penalisation of a single person in this context would be an inaccurate representation of the events which would have led up to the eventuation of the patient's harm. One could argue that this might be the case with our approach to justice at large - crimes and civil wrongs penalise individuals (usually) and certainly don't penalise the rich matrix of actors (from politicians to personal contacts) who may have contributed to the wayward behaviour. As such an atomistic approach is consistent with society's general approach to retrospective justice. However, regardless of the scenario played out, there is always rich matrix of causal actors, but only a small number are sufficiently proximate to be held accountable. The entire course of events leading up to the harm where an AIS has been used would not simply have been the individual clinician's choice to use an AIS, but also a multitude of other activities; for example, the choice of the SDC to make and release the AIS to be used, the regulatory approval for the AIS to be permitted for clinical use, a senior clinician's choice to approve the purchase the use of an AIS when their junior staff will be using it and potentially be more reliant upon it's outputs than themselves. An AIS is permitted to be used in the clinical environment because of this complex underlying activity with multiple persons involved; all of this activity has led to a single moment which resulted in an AIS affecting a clinician's decision-making for an individual patient. Unless the clinician solely designed, deployed, and used their AIS without outside help, the chances are strong that other parties were involved. It is unfair for one person to be held individually responsible for harms caused when, in fact, others are involved too in a meaningful and proximate way that engenders moral responsibility (as argued above) and that these others are not held proportionally responsible for their part.[41]

The isolation of stakeholders from each other creates no opportunity to collectively work towards preventing and managing problems should they arise other than reactively responding to them after they have already happened. Indeed, this approach could be unnecessarily restrictive of the possibilities of how responsibility could be allocated and managed. Similarly, regulating the development and deployment process of an AIS will serve to improve safety (and that ought always to be striven for), but the development and deployment of an AIS is separate event to the use of an AIS. Regulating the development and deployment process of an AIS does not address the question of who will be responsible for the consequences of its use and how the consequences of that use may be planned for. Bridges between the stakeholders, the SDCs, and the consequences of the use of the proposed technology are needed to be able to prospectively plan for the consequences of AIS use. Rather than a divisive approach which separates AIS development and developers (i.e., SDCs) from its use (i.e., clinicians), there is an opportunity now to explore a united one.

Therefore, I suggest that a more holistic and inclusive method may be fairer, which not only allows for but actively encourages the involvement of all stakeholders, instead of permitting the focus on the

[41] In part this is the function of vicarious liability in law, but obviously this does not operate at a professional regulatory level.

actions of the single clinical actor. Solutions which offer a mixed prospective/retrospective responsibility approach might offer a more practical, collective, and flexible solution whilst potentially avoiding the need for novel statutory mandates are now presented and discussed in depth.

Introducing a shared model of responsibility

Currently, as the use of AIS in clinical decision-making is still a novel, there is an opportunity to consider and plan how to proactively manage the potential risk of harms to patients. Forethought before AIS deployment may be beneficial to stakeholders by 1) increasing the effort of avoiding patient harm and 2) preventing clinicians from becoming "second victims" (Wu, 2000) to a negligence claim or by being used (intentionally or not) as a moral crumple zone by SDCs. Clinicians routinely take professional and therefore personal responsibility for their actions, yet neither SDCs nor their technologists currently have a formalised professional obligation to others. Given that both clinicians and SDCs wish to optimise the clinical decision-making process and that this thesis has argued that they owe a duty of care to the patients affected by their AIS (albeit in different ways), there is scope to explore a shared model of responsibility.

As has been identified in chapter 6, there appears to be an injustice in the benefit-risk ratio because it appears that clinicians carry the weight of responsibility, and therefore professional risk, for clinical decision making. To allow a system to be used in England and Wales under the conditions described in the scenarios explored in this thesis could amount to permissive exploitation of clinical staff as the SDC is taking all the benefit whilst the patient and the clinician are bearing all the risk of harm or being claimed against for that harm. The evolution of this unjust position makes some sense when considering Taddeo and Floridi's comments (2018); they state that existing responsibility frameworks consider the actions of individuals and are unsuited to situations where many actors are involved. Thus, other models of responsibility need to be considered.

A fairer situation would see personal moral responsibility for safe application of system use being shared by clinicians and SDCs together. This was expressed by the WHO (2021) as a model of collective responsibility. Whilst this was a great idea by the WHO, they did not take the opportunity to elaborate. I shall do so now.

A shared model of responsibility would create an opportunity for stakeholders to work together to achieve mutual and wider benefits, whilst not removing the route for holding negligent individuals accountable. Shared models of responsibility were identified in the literature review and will be briefly recapped now. Pouloudi and Magoulas (2000) suggested that responsibility should be shared out between stakeholders and suggested defined obligations and interdisciplinary working. Whitby's (2015) collaborative position also underlined that SDCs must share responsibility for consequences of

AIS use with the clinical users; his focus was on preventing harms rather than allocating blame. As noted in chapter 4’s literature review, no authors have specified how responsibility can be allocated between the two key stakeholders of clinicians and SDCs in the context of England and Wales. Chapter 5’s legal analysis agreed that there are grounds to argue that liability for damages ought to be shared among the stakeholders when AISs are developed and deployed in the clinical setting.

As well as being more just, shared models of responsibility could make using an AIS safer. This is because sharing responsibility between stakeholders would need a shared platform to manage that shared responsibility, thus invoking stakeholder discussions and tackling the foreseeable causes of harm which could eventuate. Such a shared platform may afford the opportunity for meaningful discussions resulting in clarity on issues which might have been otherwise neglected. For example, clinicians could explain that it is foreseeable that their vigilance might atrophy, thus dimming their alertness to potential erroneous AIS recommendations. If SDCs also knew that this foreseeable circumstance might happen, there would be an opportunity to jointly address how to manage the atrophy of user vigilance. SDCs, knowing of the issues of potential atrophy of vigilance, would be incentivised to ensure that their system was optimal prior to dissemination; thus, reducing the overall possible risk of erroneous harmful system recommendations being generated or used. Clinicians might additionally address this issue by recognising their own limitations, undertaking to critically appraise every AIS output prior to use, and conscientiously attempting to not allow their vigilance to wane. However, just because one stakeholder has made a positive effort to make the use of AISs safer does not excuse other stakeholders from making their own efforts. The collective aim ought to be to make AIS use as safe as possible.

A shared approach would reflect the contemporary movement of “learn not blame” (Robinson, 2019) which was initially identified with the 2013 Berwick review into patient safety (Berwick, 2013). Berwick suggests moving away from blaming NHS staff as “in the vast majority of cases it is the systems, procedures, conditions, environment and constraints they face that lead to patient safety problems.” The adoption and use of an AIS within the NHS would certainly qualify as a new system for making decisions. A clinician without a computer science background using an AIS would surely be constrained by their lack of knowledge in this discipline. Rather than looking for actors to allocate blame to, a shared model of responsibility could be designed to accept that AISs could improve clinical decision-making whilst being prepared to jointly manage the consequences for its use. For this to work clinicians and technologists must recognise and accept their relationship; the siloed thinking of ‘a clinician makes the final decision when using an AIS and therefore takes all responsibility for the consequences of that decision’ must be abandoned.

By employing shared responsibility and (inspired by Rawls, as discussed in chapter 6) using a contractarian approach, all stakeholders can communicate, discuss, and openly negotiate to devise intentions, values, and goals they can collectively subscribe to. An opportunity for communication may allow clinicians to voice that the burden of responsibility placed on them when using AISs is currently too great, would be unjust to clinicians if a patient came to harm, and that SDCs must recognise that they ought to shoulder some of this burden. SDCs are likely to want clinicians to have confidence in, and make use of, their technologies – thus SDCs could recognise a shared interest in demonstrating that they themselves are confident in, and committed to the safety of, their technologies. Stakeholder communication may involve patient groups, thus promoting understanding and autonomy if patients are involved in deciding whether to use an AIS's outputs when there is no specialist clinician available. This may promote justice by encouraging the negotiation of stakeholder responsibility allocation rather than that responsibility being assigned via the legal route speculated in chapter 5.

However, there are two circumstances which could arguably be incompatible with the employment of a shared model of responsibility.

Firstly, personal moral responsibility could be shared between the SDC and the clinician for as long as the clinician is required to supervise AIS use in clinical care; but should the system develop to the extent that it no longer needs a clinician's supervision and the clinician is no longer required for the system to be used safely, the clinician ought to be allowed to step away from holding personal moral responsibility for the system's use as the final decision which the AIS reaches no longer involves the clinician. Here, the SDC would then be responsible for an AIS's achievements and harms as their product will be directly interacting with the patient without the clinician's input.

Secondly, we need to ensure that shared models of responsibility are used only when scenarios dictate that sharing responsibility is appropriate. If patient harm happened due to, for example, clearly malicious use of an AIS, then a shared model of responsibility is clearly inappropriate as personal moral responsibility remains with the malicious actor. It would also be inappropriate to share responsibility in circumstances of non-malicious negligence caused by error, misuse or not following the standard operating procedure; again, so long as the design process of the AIS took reasonable steps to proactively prevent such negligence, personal moral responsibility remains with the individual actor.

Whilst this thesis offers solutions to the problem of allocating responsibility, both the technology and the conversation surrounding the use of AIS in clinical decision-making changes and updates daily. As such, it is to be expected that the proposed solution of a shared model of responsibility shall be challenged, changed, and possibly rejected by stakeholders. This model allows the negotiation to continue, and the model to change as the technology changes. It is not fixed, and that is ideal in this context. However, this solution offers benefits to stakeholders and offers discussion in this area where other works have not, and so, whilst imperfect, the shared model of responsibility is hence presented.

Mixed prospective/retrospective approaches incorporating a shared model of responsibility

As identified at the start of this chapter, if it is foreseeable that an AISs could dispense an output which might cause harm, and that harm reaches a patient, then an agent (forewarned due to their knowledge of their prospective responsibility) may ameliorate a potential failure by preparing to simultaneously assume both prospective and retrospective responsibility for their actions. This means that it would be recognised that an actor has a duty of care to the patient and that the actor had made preparations to accept that duty before the AIS was used in clinical decision-making.

Mixed prospective and retrospective approaches could allow for a bridge to be made between the clinician and the SDCs via a shared model of responsibility. This section will discuss and explore such models which have been used internationally and, where suitable, apply them to the context of England and Wales.

If a shared model of responsibility is to be adopted by stakeholders, how could a fair balance be practically achieved? What could a shared model of responsibility look like? Potentially, a united multi-agent front could be created with a holistic approach to the common aim of ensuring that a system is a safe as possible and that stakeholders[42] are involved in the planning, development, and deployment of AISs, as well as having the opportunity to consider the effects of the use of the system in question. Stakeholders could prospectively consider and engage with 1) establishing a prospective consideration of which stakeholders owed a duty of care to whom and what that duty would entail, and 2) planning for the potential for retrospective consideration of the foreseeable risks of the AIS which stakeholders wished to use, i.e., to have a plan to provide compensation for harms which eventuate as a result of using the AIS in clinical decision-making.

The outcomes of the use of AIS in clinical decision-making would be dependent on the actions of key stakeholders who currently do not closely communicate or associate. The discussion regarding the allocation of responsibility for the use of AIS is currently fragmented between stakeholders and external observers (such as legal and ethical commentators). This fragmentation is not conducive to creating a holistic solution which hears and honours the position of every group affected by the use

[42] Whilst this thesis is principally concerned with the clinical stakeholder, in practice all stakeholders could be invited.

of AIS in healthcare. As per Rawls' model of social cooperation, inclusion of all stakeholder views can be achieved by creating a space where all stakeholders involved in and affected by the development and use of AIS in clinical decision making can assemble; in this space each stakeholder's view may be weighed and appraised in union by the congregation.

If, as per the IBM Watson in Mongolia example identified in the literature review (Ross and Swetlitz, 2017), the AIS in question was unproven by third parties and at risk of erroneous outputs, the need for communication and negotiation of responsibility between the SDC, the clinician and the patient is particularly strong. Under a shared responsibility model, the SDC and clinician may have frank discussions as to the risks and values of exposing the patient to the system. Given the potential for risk, the importance of patient involvement increases dramatically; involvement increases the opportunity for comprehension of the risk to which they are exposed. If all stakeholders decided to go ahead and use the system, then all would have the potential to benefit from the system's use: 1) the SDC gets recognition and is paid for the system being used, 2) the clinician is supported in their clinical decision making, and 3) the patient may benefit with potentially improved clinical decision making. But if the patient is harmed due to the use of the AIS's outputs, the possibility of and responsibility for that negative outcome would have already been discussed, allocated and accepted by all stakeholders prior to that harm arising. If potential for harm is discussed and responsibility for potential harms is pre-determined where possible, stakeholders can prepare to avoid or make amends for that harm.

Two mixed prospective/retrospective approaches were identified during the course of researching this thesis and shall now both be examined. These are presented within international contexts and offer plans for compensation in the eventuation of risks. Whilst they are not expected to be definitive solutions (indeed, one is rejected) these approaches may serve as initial examples of shared responsibility from which more refined approaches may be later developed.

Both approaches involve financial restitution to the patient to who has been harmed due to the use of AIS in their clinical decision making. The IEEE's (2017) guidance call governments to make the provision of financial responsibility (e.g., insurance) for those harmed by AISs a requirement. Currently, for England and Wales, the Consumer Protection Act 1987 dictates that manufacturers (and sometimes supply chain providers) are liable for their products[43]. However, other than in the context of a clinical trial using an investigational medicinal product (Health Research Authority, 2021), it

[43] As noted in chapter 5, discussion in this thesis's scope is limited to discussion on 'fault' rather than 'strict' liability which is addressed by the Consumer Protection Act. However, I nod to it here to illustrate the lack of insurance obligation placed on the SDC.

appears that the Medicines and Healthcare Products Regulatory Agency (MHRA) has not specified an obligation that insurance be in place to cover that liability prior to the deployment of a device. Yet, healthcare organisations are advised by the MHRA to ensure that a medical device provider (in this case the SDC offering an AIS) has "adequate insurance or indemnity in place" (Medicines and Healthcare Products Regulatory Agency, 2021b, p.34). This puts responsibility onto the clinical team to ensure that an SDC has its insurance in place prior to device deployment rather than the MHRA to proactively compel the SDC to obtain it.

Additionally, in England and Wales, all clinicians are required to be covered by a professional indemnity arrangement for their clinical work (Nursing and Midwifery Council; General Medical Council, 2020; Health and Care Professions Council, 2020). SDCs do not have a regulator to enforce a professional indemnity arrangement requirement upon either SDCs or technologists. It seems that the burden of establishing that an SDC and its AIS are suitably indemnified falls to the clinical user. The lack of formal structure for such insurance may be off-putting to clinicians when choosing whether to use an AIS or not; additionally, a lack of a transparent insurance arrangement may not be reassuring to patients who hope to benefit from the AIS in question. To address this, the discussion now moves to examine models of cover which might be arranged to ensure that SDCs may also carry their share of prospective/retrospective responsibility.

New Zealand- "No Fault" system.

Rather than looking at the actions of individual AI systems, technologists, SDCs, and clinicians, 'big picture' solutions have been floated in the literature. One model from New Zealand is representative of such a solution: the no-fault accident compensation scheme. Here, in the case of accidents, a governmentally administered taxpayer funded scheme, the Accident Compensation Corporation (ACC), dispenses damages to the victim regardless of who was at fault (Turner, 2019). In principle, this type of system would be ideal for the claimant in the case of AIS use having harmed them; regardless of the mechanism of the accident, the injured party would be directly attended to.

Both Holm *et al* (2021) and Yeung (2019) suggest the use of no-fault approaches, however, Holm *et al* claim that the projected cost of a no-fault scheme is prohibitive in the UK, even in the presence of the successful New Zealand model. This is as a no-fault scheme at a national level places the burden of injuries upon society as a whole via taxation and employer levy (Bismark and Paterson, 2006), rather than limiting the financial burden of the scheme to those whose actions had caused the harm. The benefit of a no-fault scheme is obvious when the use of AISs are universally accepted, encouraged, and utilised by a population within their universal healthcare system. For this reason, it is easy to see how it could be applied to the use of AISs in the NHS, particularly if, overall, the costs of such a scheme are lower than the savings made by using the AIS. Disputes and claims for compensation regarding patient harm due to the clinical actions and operations delivered by the NHS are managed by NHS Resolution (NHS Resolution, 2020a). Given that the UK's healthcare is predominantly delivered by the NHS, the liability management of using AISs in clinical decision making could continue to be centralised using NHS Resolution; thus, a centrally dictated system of restitution might provide a convenient one-size-fits-all approach.

Still, this scheme would not necessarily encourage an SDC to place any increased emphasis on the operational safety of their product if they knew that the nation would be paying for any mishaps rather than the SDC themselves making financial amends for their part in the harm caused. This is especially true when considering that the right to sue for tort is prohibited in New Zealand if the personal injury is provided for through the ACC scheme. Were such a condition placed on a similar scheme for AISs, SDCs would be protected rather than penalised for harms eventuated. This concern could also be extended to the clinical professional using the AIS; although, as the clinical professional would still be professionally regulated, if it were found that their actions were not of the standard of conduct prescribed by their code of practice then the clinician could still face professional sanctions. A no-fault scheme would allow the nation to bear the brunt of compensating for the injury rather than the responsible individual, and the prohibition of tort actions might result in less care being taken, thus leading to the unsatisfactory result of *less* uptake of personal moral responsibility by actors (e.g., technologists, SDCs, or clinicians) rather than more. Thus, the prospective and retrospective considerations in this approach are made by the ACC on behalf of the nation rather than the actors whose actions had resulted in the eventuation of harms.

This scheme relies upon a national solution rather than an individual one. It does not ensure that actors (either individually or in groups e.g., SDCs or hospital organisations) prepare to undertake personal retrospective responsibility for their actions in response to the prospective responsibility which they owe to patients. A national scheme does not proactively incentivise actors (the SDC/technologist or the clinical user) to ensure that an AIS is appropriate and safe for use in clinical decision-making, nor does it ensure that actors are able to express their personal moral responsibility to patients by arranging to compensate for the harms which their actions might cause. For this reason, this thesis rejects the use of a similar no-fault scheme for AIS use in clinical decision-making at a national level. Instead, we shall now examine a possible novel alternative.

Risk pooling

If, as identified earlier in this thesis, it is fairer when personal moral responsibility for safe application of AIS use is shared between clinicians and SDCs, a solution which employs both prospective and retrospective approaches (a mixed prospective/retrospective approach) could be designed and deployed. This would aim to promote fairness when allocating responsibility to actors and to ensure that potential negative effects of the use of the AIS upon the patient is considered and mitigations planned prior to AIS deployment.

Yeung (2021, p.62) suggests the use of “some kind of mandatory insurance scheme” on a no-fault basis to cover the general use of AIS’s in society, but Allain, writing from a US perspective, suggests that the ‘enterprise liability’ approach may satisfy the interests of all parties when AISs are used specifically in healthcare (Allain, 2013). Enterprise liability is broadly identified by Klemme as the imperative that ‘losses to society created or caused by an enterprise or, more simply, by an activity, ought to be borne by that enterprise or activity’ (Klemme, 1975-6, p.158). The fault of each actor within an ‘enterprise’[44] might well be covered by either a singular insurance paid for by both the SDC and the clinician or by separate policies which reduce the impact on each individual in the event of a claim (Allain, 2013).

Allain proposed that the user of devices such as IBM Watson be indemnified with insurance which combines aspects of product liability and vicarious liability as well as medical malpractice and allows the spread of fault between the clinicians using the system, the SDCs who have developed the system and the hospital where the system is being used. The enterprise liability model reduces the burden on claimants as the court will not need to analyse each actor’s role in the claimant’s misfortune, but instead look at the actions of the team as a whole. In this way, ‘insurance acts to better spread the risk of loss throughout society reducing the economic impact of each individual judgment’ (Allain, 2013).

Allain proposes that restitution to the claimant would be shared equally between the SDC who has supplied the system and the clinician or hospital who has adopted the system. She argues that if stakeholders equally share the burden of loss, the risk of loss is shared across all actors resulting in reduced economic disincentives. This would encourage SDCs to ensure that their system is as accurate as possible, and hospital management teams can be reassured that they will not be left to shoulder the full cost without appropriate contributions or reimbursement from their technical partners.

However, it is arguably not enterprise liability if the economic impact of a claim is assigned only to an insurer which shields the actors who caused the harm. Loss-spreading using insurance allows stakeholders to hold prospective and retrospective liability for their actions by sharing the financial

[44] ‘Enterprise’ meaning a group of people who have embarked on a project.
Sadly, ‘enterprise’ in this context does not refer to any of the famous fictional star ships.

cost of the burden of liability. Loss-spreading and enterprise liability are distinct, and the scenario relies upon actors purchasing insurance which covers their activities.

If there is a risk that an AIS will wrongly advise a clinical user and that harm could result to a patient, stakeholders will need to be incentivised to accept that risk. Insurance is beneficial if society accepts that the introduction of AIS in clinical decision-making is, on balance, beneficial to the public and the adoption of this technology is deemed collectively to be in society's interest. If the SDC and the clinician must obtain insurance for use of an AIS to be permissible in the clinical decision-making context, the cost of paying for that insurance will fall upon those who pay for the healthcare provided. Therefore, rather than considering the SDC and the clinician as paying for the required insurance, it is in fact the patient or those who pay for the patient's care who are burdened with that cost (similarly to New Zealand's ACC scheme which is funded by the taxpayer); the community truly does carry the cost of liability when seen this way. Again, it is likely that the use of an AIS in clinical decision-making and the associated costs of insuring its use will only be accepted if the use of AIS is beneficial to society in general and if the costs of the scheme are lower than the savings made by using the AIS.

Enterprise liability needs to be fit for purpose else it may lead to parties other than the insurer being liable. For example, if the clinician uses a recommendation which leads to patient injury and the insurance policy's wording does not cover this tort, the insurance might not benefit the injured patient; thus, negating the enterprise's intention of obtaining insurance.

As a solution to these issues, this thesis proposes the use of 'risk pooling'. Risk Pooling is a prospective arrangement between all stakeholders who have a duty of care to commit to retrospectively addressing possible harms which could befall a patient due to the intervention in question- in this case, the consequences of the use of an AIS in clinical-decision making. A risk pooling arrangement may be constructed to reflect the multiple stakeholders' collective and intentional prospective acceptance of their part of responsibility should there be an eventuation of risk when AISs are used. It encourages actors' preparation to assume retrospective responsibility if harms arise when AISs are used in clinical decision-making by providing a platform for proactively and prospectively planning how potential harms could be financially recompensed. The intention of this system is to safeguard not just the patient, but also enterprise and innovation itself - which aims to aid the patient both in the present and in future AIS developments. Where a risk is foreseeable, e.g., if a system is only permitted to be used under the supervision of a specialist clinician (e.g., IBM's Watson for Oncology as identified in chapter 4's literature review) it is reasonable that stakeholders ought to discuss how to manage that potential risk before it eventuates.

Song defines risk pooling as:

'…if X performs an action which imposes an unreasonable risk of harm on Y, then X is liable to Y, and therefore obliged to make an ex ante compensation into a social pool that is roughly equivalent to the cost of expected harm (i.e., the probability of actual harm multiplied by the amount of the cost incurred by the harm).'

Song, 2019

Merkin and Steele (Merkin and Steele, 2013) speak of insurance operating under an actuarial model to spread risk across a discrete pool. A risk pool in this thesis's context could be a single insurance policy which charges each participant as according to their own risk. Speculating the calculations of individual stakeholder's contributions will be a highly complex task which is outside of the scope of this thesis; however, some comments may be made here.

Merkin and Steele (Merkin and Steele, 2013) describe fairness to pool members as the principle of stakeholders receiving what they have paid in. Yet, they warn that, whilst risk pools may be well defined, they are rarely homogenous and that there is no requirement for premiums to be allocated as per the class of risk, and so premiums could be unfairly distributed between stakeholders. For example, the clinician is closer to the patient and is in the position of being the final decision-maker in clinical decision-making. Therefore, they might hold a larger proportion of immediate responsibility than the SDC. If the clinician has not used the system in the way intended, it is not just or reasonable for the SDC to subsidise the clinician's wrongs. Nevertheless, if the clinician has used the AIS appropriately and was unable to detect a system error, the SDC's contribution ought to be commensurate with their involvement in creating the risk of harm in the first place.

Risk pooling in England and Wales: proposals for reform

The attraction of risk pooling is that the regulatory disconnect between the clinician, the SDCs, and the patient is addressed. Risk pooling can be engineered so that it relieves a large portion of the patient's burden of making a negligence claim. After a claim is made, a court may engage in detailed consideration of the legal issues as they apply to the defendants, for instance establishing whether the duty of care exists or dealing with problems of causation. The standard court process can be costly and prolonged, thus delaying resolution. With an agreed risk pooling arrangement in place, this process is not needed; consequently, the patient can commence their recovery journey sooner and additional distress is avoided by the injured party. A risk pooling insurance scheme would cover harms when they arise, but unlike Allain's proposal, the defendants only contribute in accordance with the extent of their liability.

1. Insurance schemes for clinical malpractice should include coverage for AI-related damage

As already noted above, clinicians have a professional obligation to carry adequate insurance against injuries arising from their activities (Nursing and Midwifery Council; General Medical Council, 2020; Health and Care Professions Council, 2020). They may additionally be subject to vicarious liability rules, whereby the employer is deemed to adopt the liability of the conduct of its employees.

As mentioned earlier, NHS bodies operate under a regime of self-insurance which is administered through NHS Resolution schemes (NHS Resolution, 2020a). Risk pooling would be a novel approach for NHS resolution to take. Currently, the liability cover provided by NHS Resolution covers only clinical liability (NHS Resolution, 2020b). If risk pooling were to be adopted for AISs being used for clinical decision making in the NHS it would mean a step change; here, NHS Resolution would be co-organising cover with SDCs who seek to influence clinician's actions without being clinical actors themselves. Risk pooling might be a difficult concept to sell to NHS Resolution as it would involve prospective negotiation with external actors rather than keeping liability issues 'inhouse'. However, if clinicians are to be directly influenced by the SDC's product, and the recommendations made by that product will directly affect the actions of the clinician with the possibility this will result in harm, then NHS Resolution might be interested in using risk pooling as a way of spreading this novel risk with SDCs.

To summarise, the advantage would be twofold: 1) that risk pooling could be structured to allow faster and easier access to restitution for patients whilst 2) allowing the reflection of that risk to be adjusted to each actor's subscription to the risk pool. The reflected risk would be calculated as per the initially predicted (and then later the actual) risk of using the AIS.

If risk pooling is to be widely adopted, ideally it would need buy-in from large authoritative bodies rather than from individual actors at individual clinical institutions. Approval of a risk pooling scheme, not just by NHS Resolution, but, additionally by other larger organisations, e.g., NICE, the Royal Colleges, or the new multiagency advice service for AI technologies in health and social care might help to create the (aforementioned in chapter 5) *Bolam-Bolitho* calibration of the standard of care in this area. As such, if an AIS for clinical decision making has not had risk pooling arranged for its use then a lack of prospective preparation may be considered illogical (as per *Bolitho*), and that to use that AIS might not be in accordance with a responsible body of opinion (as per *Bolam*) if that body insists that risk pooling is in place. It is not inconceivable that, if successful, professional regulators may extend their indemnity requirement to make membership of a risk pool a condition of professional registration if the registrant is using AIS in their clinical decision making.

As noted in the thesis introduction, not all healthcare provision in the UK is provided by the NHS. Private healthcare institutions working outside of NHS frameworks may optionally take out public liability insurance which may supplement or replace the insurance coverage of the individual clinicians who work for them. SDCs can purchase public liability insurance which can indemnify their liability in negligence. A fractured coverage approach such as this could lead to inefficiencies compared to a more organised and considered risk pooling model.

In order for claimants to receive the protection of the risk pooling scheme, it might be necessary for clinicians and SDCs wishing to deploy/use an AIS in a clinical area to be jointly under a statutory obligation to hold compulsory insurance for harms related to AIS use, thereby sidestepping the issues of patchy cover through voluntary schemes. Insurers for clinical malpractice might be best placed to process claims directly with patients. Rules directing that clinical insurers are liable to pay claims in the first instance would mean that the patient would not need to identify multiple potential defendants in order to pursue a claim.

Conditions for the deployment and adoption of AISs can be stipulated in a risk pooling agreement. For example, thanks to the Medical Devices Regulations 2002, the SDC has a duty to ensure that their system is safe. If they do not, they may be obliged to recall it or make it safe under the Medical Devices Regulations 2002. If the SDC fails to make a system safe, for example, via an update, they have breached this statutory duty. Once this update is available, the clinician cannot reasonably claim against the insurance scheme if they have not updated their AIS. Additional conditions could include the requirement for approval from the multiagency advice service for AI technologies in health and social care for the AIS to be deployed.

2. Insurance schemes for clinical malpractice should have powers to recover costs

Although the insurer is intended to be the sole point of contact for a claimant, risk pooling does not require that the insurer fund the full amount of compensation from the insurer's own reserves. The notion of risk pooling recognises that the totality of the damage is the result of numerous, sometimes undetectable, errors and mistakes. The actions or omissions of one party may indeed be the most proximate cause to the damage in time and space, yet, as found in *Hughes v Williams*,[45] it is still legitimate to seek a contribution from co-defendants whose negligence also led to the occurrence of the damage.

This entitlement exists in statute via the Civil Liability (Contribution) Act 1978; therefore, in risk pooling, the insurer acts as a centralised claims administrator but can also investigate specific incidents in more depth. An individual claimant may only be able to amass enough evidence to merit

[45] The defendant hit a car in which the claimant was a passenger. The case questioned if the mother of the claimant had incorrectly chosen an appropriate child seat for the claimant. The court found that the claimant's injuries would have been largely avoided had the correct child seat been used. A contribution of 25% was ordered.

a claim against some defendants but not others. Meanwhile, insurers can mobilise their institutional resources in preparing a claim. Moreover, they are incentivised to seek contributions from other parties as it can eliminate or reduce the burden on the insured community, thereby keeping premiums at competitive rates. The risk pool is this way protected, but without inconveniencing the injured patient further.

Whilst not healthcare, there are some parallels to be drawn from recent legislation regarding autonomous vehicles. In part I, section 3 of the Automated and Electric Vehicles Act 2018, if a person allows a vehicle to begin driving itself when to do so was inappropriate, that person may find themselves liable for a contribution towards a claim from an injured party as per the Law Reform (Contributory Negligence) Act 1945. Again, in the Automated and Electric Vehicles Act 2018, part I, section 4, it is outlined that an insurance policy will be limited for a person who fails to install software for their vehicle that they ought to have known was safety-critical. It can be drawn generally that it is in the interests of actors to ensure that any AIS system that they choose to use is fit and appropriate for use prior to its deployment.

Allowing insurance companies to recover costs from individual parties makes it easier for stakeholders to be held financially responsible for their negligent actions without the stakeholder being able to shift that responsibility to others. An agreed risk pool which allows this promotes and enforces the adoption of personal moral responsibility for others (even if only expressed financially) by individual subscribers whilst providing a platform to express collective responsibility to answer an injured patient's claim.

3. Models of joint liability should be used

Where more than one tortfeasor has been identified, the power of a claimant to recover costs or seek financial contribution to an award of damages from a tortfeasor rests on the model of joint liability, (as expressed in *Fairchild v Glenhaven Funeral Services Ltd*).[46] Joint liability means that a claimant may claim for the entirety of their injury from one tortfeasor, even if others were also involved in the negligent act. By eliminating the need for a claimant to bring several cases, joint liability acknowledges that smoothing the pathway for patient claims is a key concern; early and full compensation is crucial for the claimant to have the stability to start their journey of recovery as soon as practicable. Thus, risk pooling is grounded in the idea that the stakeholders who intend to benefit from their

[46] Where the claimant had worked for several employers who had negligently exposed them to asbestos. This case considered how a claim from a single individual exposed to the same tort by multiple parties could be awarded.

participation in a risky activity ought to accept moral and legal responsibilities to the individuals and communities that may suffer the consequences of when things go awry.

The instatement of joint liability could be considered a legal expression of a shared model of responsibly, however this does risk a single tortfeasor carrying the burden of a claim when others are also liable. There is a key case, notable in the area of industrial liability and causation, which is concerned with the damage caused by asbestos to employees. The court in *Barker v Corus UK Ltd*[47] erred in making liability for 'material increase in the risk' 'several' or proportionate to the defendant's share of the risk exposure. This means that when more than one actor has acted negligently, a single actor's liability is not lessened just because another actor has performed the same harm; instead, liability should be divided up between tortfeasors proportionally to reflect the (probable) share of harm which they had caused. Reflecting *Barker*, any party which negligently exposes a claimant to a risk of harm via an AIS used in healthcare should bear full responsibility for the damage if that risk materialises; when more than one stakeholder has caused that harm (e.g., the SDC and the clinician) liability should be fairly and proportionately shared between them as per their contribution to that harm.

However, per *Barker*, this means that individual claims would need to be made to every responsible party where an AIS is used (defendants are severally, but not jointly liable). Parliament recognised this principle and, in section 3 of the Compensation Act 2006, they reversed *Barker* so that a claimant may claim the full loss from an individual tortfeasor. From here, a tortfeasor may pursue a contribution towards that claim from their fellow tortfeasors who also contributed to the initial claimant's harm. Thus, an injured patient has only one case to contend, and the burden of further cases is left with the tortfeasor.

A risk pool's design could represent and cover the activities of its members in a manner which reflects joint liability. The full loss could be claimed from the risk pool which would avoid the need for a potentially protracted legal claims process. This could be more attractive to the pool's members if they jointly knew that their activities were already reasonably covered prior to AIS deployment. This could also be encouraging to patients if they knew that coverage of the risk of AIS being used in their care was already in place, easily accessible to them, and that the coverage fairly reflected the current state of tort law for claimants.

4. Patient consent should be restricted

[47] Similarly to *Fairchild*, claimants in *Barker* had been exposed to asbestos by multiple employers. Here though, the House of Lords held that liabilities should be proportionately liable rather than joint and severally in *Fairchild.* The Compensation Act 2006 reversed this, but only for mesothelioma.

The example of IBM Watson for Oncology's use in Mongolia in chapter 4's literature review showed that AISs to aid clinical decision-making may be desirable in places where specialist clinicians are not available. Where no other options are accessible for a patient, it is understandable that they would consider the use of technology which is reasonably made available to them. However, it is not really a choice if a patient is faced with the use of an AIS or nothing; if the healthcare need is sufficiently pressing, the patient may feel obligated to choose to accept the use of the AIS in their clinical decision-making.

For some activities, the nature of the risk involved is so great that the law has seen fit to restrain unwitting claimants from granting effective consent as this would bar them from compensation to which they might otherwise be entitled. For instance, it has been held that passengers are not able to form the necessary acceptance of risk required for the defence of *volenti non fit injuria* (where someone has willingly placed themselves in a position of potential harm) where the driver of the vehicle is intoxicated as per *Dann v Hamilton*. Statute now excludes this defence in its entirety in road traffic accidents, because all drivers are required to have insurance in order for the speedy administration of justice to claimants (Road Traffic Act 1988 s.149).

A similar contention can be made in favour of patients who have been harmed as a result of defective AISs. As laypersons, patients cannot be expected to take the steps to safeguard their own interests, especially if an AIS is presented to them in a healthcare resource-depleted environment and particularly when explanations of how some AISs operate are not readily forthcoming for commercial and technical reasons. Restricting the patient's ability to consent to treatment by defective AISs would provide enhanced protection for vulnerable claimants. Whilst the notion of restricting a patient's consent may appear overly paternal,[48] in practice this happens frequently. Clinicians make decisions every day based on their understanding of the standard of care and the evidence base of their clinical practice and do not necessarily discuss with the patient the choice of research used to inform care. Whilst the patient can be aware (and maybe arguably ought to be actively involved in) the selection of the use of an AIS to inform their care, it is the clinician who will rationalise and will have to face the consequences for choosing to use it, much the same as they would for choosing another tool (e.g., selecting an appropriate drug from available options).

Advantages of risk pooling

As aforementioned, for risk pooling to be utilised in the context of AI deployment for clinical decision making, responsibility must be embraced for both its prospective and retrospective qualities. Risk

[48] Especially as this author recognises that they are writing from the privileged position of a healthcare resource rich environment.

pooling is prospective as it encourages actors to consider their responsibility to patients and reduce risks before their AIS reaches the bedside, and it is retrospective by planning for restitution for patients long before harm has the opportunity to occur. Therefore, the prime benefit of risk pooling is the adoption of personal moral responsibility in a very practical sense by SDCs wishing to deploy AISs in the clinical environment. The additional benefit is that that SDCs may be able to collectively share that responsibility with other stakeholders without any individual agent losing the personal moral responsibility which they have adopted, and no-one is used as a moral or legal crumple zone. Indeed, through cooperation with other stakeholders, the responsibility held both personally and by the group becomes only stronger.

When considering Rawls's three essential features of social cooperation (Rawls, 2001) it can be argued that risk pooling has the potential to fit all three of these features. The following explores how.

Rawls's first essential feature recognised that rules aid the guidance of social cooperation which in turn regulate the conduct of those who cooperate in the societal structure. Attempts to deploy Rawls's 'veil of ignorance' could be limited in practice as parties may be unable to put aside their own interests but, if the aim is for stakeholders to agree on the fairest way to pool risk, then stakeholders should try look at arguments made by all stakeholders from the original position. As such, the utilisation of open communication may aid every stakeholder to hear and be empathetic to the voice of the other. By sharing concerns and considering their counterpart's positions, they may jointly recognise and value that patient safety is paramount and that all efforts should focus on ensuring that stakeholders adopt a duty of care for the patient to achieve the aim of ensuring patient safety. An open line of communication between the SDC and the clinician and a willingness to communicate is required before any social cooperation can be achieved so that development and deployment of AISs for clinical decision making may be discussed.

Rawls's second essential feature recognised that rules are fair when everyone accepts them; creation of a cooperative environment of mutuality or reciprocity allows participants to benefit when standards are publicly agreed. Once a line of communication has been established and stakeholders have had an opportunity to evaluate and express how the system's deployment will affect them, propositions and negotiations may take place between all parties. Such negotiations serve to help parties collectively decide what rules are required for actors to follow for the AIS to be deployed safely. The rules could encompass establishment of who owes a duty to whom and how that duty is honoured, i.e., in a given scenario, which combination of actors take moral responsibility for harm eventuating due to the use of an AIS for clinical decision making. Such planning ought to ensure that the rationale for any awarding of responsibility is clearly set out to ensure fairness. Should harms later eventuate, there ought to be a mechanism to re-evaluate how responsibility has been awarded if a stakeholder believes that they have been unfairly awarded that responsibility. Openly discussed and rationally allocated responsibility to actors with iterative evaluation will aid fairness and acceptance.

Rawls's third essential feature recognised that participants shall wish to somehow advance their own position. It is understandable and far from unreasonable that stakeholders are motivated to improve their own standing; an SDC may wish to help others and get financial compensation for that help, the clinician may have the same goals, the patient may wish to simply restore their health. But advancing one's own position may threaten the position of another stakeholder. For example:

- the SDC may desire that their system is used but they have an active interest in not taking this responsibility, and pass it on to others. This may impact upon clinicians who might carry the eventuating legal burden of AIS use. Whilst clinicians may not want to carry the burden of the responsibility alone, they will have an active interest in reassuring other clinicians and their patients that they are confident enough in the AIS that informs their practice to take *some* responsibility – their involvement could thus be an assurance standard, but that does not excuse the SDC from carrying their own portion of responsibility.
- the clinician may wish to benefit by handing over the cognitive burden of clinical decision making responsibility to an AIS yet may not recognise that a system is not advanced enough to make that decision without clinical supervision. Following an AIS's recommendation could be a risk to the patient.
- the patient needing specialist clinical help will be interested in having access to the best of both worlds; they might wish their health to be positively affected by the use of an AIS, and for others being entirely liable if AIS use causes harm.

Ultimately, all parties have an interest in reaping the benefits of AIS use, but also in passing as much risk/cost onto others as possible.

The idea of a shared model of responsibility as outlined here is fair as it does not allow any stakeholder (SDCs, lawmakers, or regulators for that matter) to impose responsibility upon another. Instead, it utilises collaboration and allows all stakeholders to input their values, negotiate their position and, having made that negotiation, take conscientious and active ownership of their role as responsible agents when an AIS is deployed for clinical decision making. Communication is key though, and no stakeholder may be an island to the others to which their actions affect. Reluctant stakeholders could be encouraged to participate if they recognise that the potential cost to not engaging in this process could be to being excluded entirely and thereby losing out on ensuring that their needs are heard and met via agreement with the members of the risk pool.

Depending upon how it is negotiated and subsequently organised, risk pooling could be designed as a communitarian/contractarian solution; i.e., the design of the risk pool could potentially be initially guided/mandated by government, but, as long as it places the care of patients first, could also be managed by stakeholders who have come together to create a risk pool and have all collectively and constructively argued their positions (as per Cox's nexus of contracts discussed in chapter 6, (Cox, 1997). Skilled communication and negotiation between all stakeholders would need to take place for such a construct to be successful. The resultant nexus of contracts, which would create the risk pool between stakeholders, and the chosen insurance institution would include agreement about who is liable for what action in each foreseeable situation when the AIS is used, as well as negotiating how each stakeholder would contribute to the risk pool. A nexus of contracts which has been well-argued and agreed by all involved would embody Rawls' first two features of cooperation via the negotiation, development, and adoption of rules which have been agreed by all stakeholders.

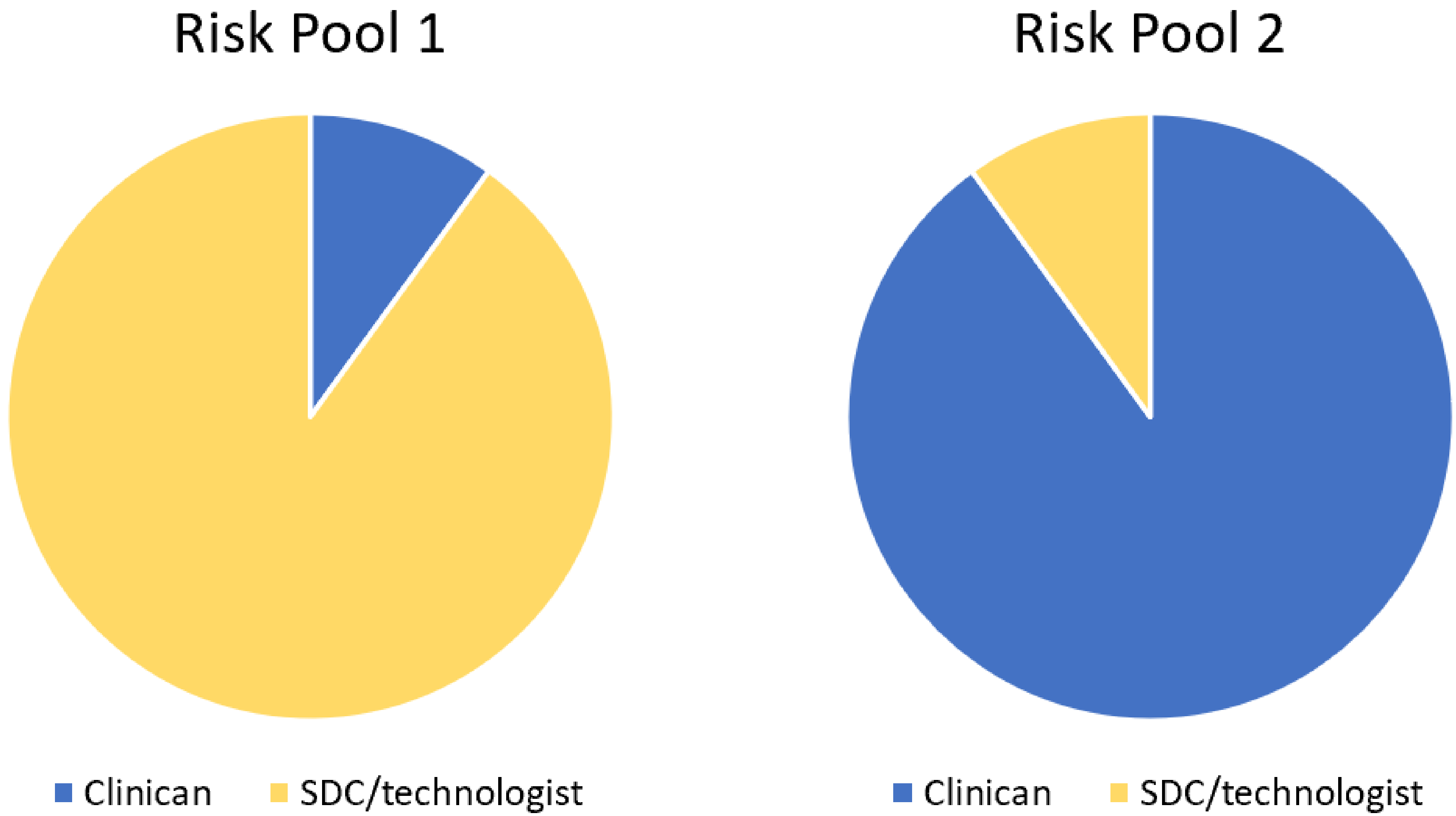
Employing the nexus of contracts approach would aim to open the conversation about responsibility between stakeholders with the ultimate aim of reducing unfairness in the calculation of contributions of each actor to the risk pool, whilst ensuring that the potential claims of an injured party are met. These negotiations and the resultant risk pool establishment would allow stakeholders to satisfy the wish to advance their position (Rawls' third feature) by allowing SDCs to deploy their systems, clinicians to use that system, and for patients to reap the rewards of better decision making as a result of using that system along with financial protection should an erroneous AIS recommendation be used.

Risk pooling would likely be individualised to specified locations or incidences of AIS use, however care must be taken. Individualisation may encourage large variations in risk pooling approaches leading to inconsistencies, such as contributions to risk pool costs, for stakeholders. To arrest this problem from the outset, it would be valuable for stakeholders to explore agreed minimum standards for all risk pool arrangements involving AIS in clinical decision-making to ensure fairness to stakeholders and quality of coverage for their interests. However, it would need to be made clear that any minimum standard would be a place for risk pooling negotiations to build from, rather than to provide a target for the least possible coverage to be aimed for.

Any nexus of contracts created needs not to be either static or permanent. Risk pooling could be iterative and specific to the AIS and the stakeholders to whom its use applies. Responsibilities between stakeholders should be re-negotiated over time as each individual AIS proves itself; this is to allow titration of the amount of responsibility held by each stakeholder reflecting their actual (rather than predicted) risks. Re-negotiation of the nexus of contracts could take place using examples of incidents which resulted from using the AIS; each iteration of re-negotiation could re-evaluate the risk of the use of an AIS application for each stakeholder using it. If a system were found to have developed to be so good that a clinician was no longer needed to oversee its utilisation, the clinician could in theory hand responsibility for the use of the system to the SDC and exit the risk pool. At this point, the SDC and the patient may negotiate with each other directly. But if the system were to perform worse than expected, and the AIS were to remain in service, the clinician's responsibility would increase due to the need for closer supervision of the AIS to ensure patient safety. In this negotiation the clinician would have a voice to express whether this is a responsibility they wish to hold, or if using the AIS is a permissible risk to expose their patients too. Similarly, patient stakeholder groups should be consulted to determine if they are happy to accept the increased potential for exposure to that risk. The cost to the SDC to remain in this risk pool would increase to reflect that they had provided an AIS which may bring a higher risk of harm occurrence to the patient.

This exemplifies how one nexus of contracts might very well not be universal; agreements and payments which may suit one application of AI use may not reflect the use in another scenario. For example, low contributions of a clinician into a risk pool if they have specialist training and vast experience of the conditions for which the AIS will be used (as per figure 7, risk pool 1), versus higher contributions if the clinician has less experience (as per figure 7, risk pool 2). The same could be true depending on the characteristics of the AIS in use, e.g., the degree of opacity, the confirmation of third party validation of the AIS (e.g., by the multiagency advice service), or the track record of an AIS may affect the contribution which a SDC may make to that same risk pool. A contract may vary still according to who was pushing for its use. For example, if the SDC was pressing for inexperienced doctors to use their AIS, the SDC might be expected to bear more responsibility as they were insisting that the risk of using inexperienced staff was taken. If one stakeholder group had had more incidents using or deploying their AIS than another stakeholder group, then that may be taken into consideration too. If there is more than one risk pool for the use of a single AIS in healthcare (e.g., the AIS used at different NHS Trusts) and each risk pool is individually crafted, then fairness could be specifically sculpted to fit the skill and risk which was present in each clinical use situation. However, common features could apply to every risk pool which would ensure universally high-quality attention to harmed patients; for example, minimum awards to patients as dependent on the severity of the actual harm, maximum wait for a patient from submitting a claim to resolution.

Figure 7: Risk Pooling



Along with regulating clinicians and SDCs together, risk pooling might address Nissenbaum's (Nissenbaum, 1996) 'problem of many hands'; not just the many hands of the SDC, but also of the clinical user. For example: for the purposes of a patient claim, ideally, all that would be necessary to establish would be that the AIS's recommendation had been followed and that the patient had come to harm as a result of the clinician following that recommendation. Incident analysis activities could be used to establish if the clinician ought to have spotted that the recommendation was faulty and the contribution of the SDC via their development of the AIS. Such causal stories can be argued separately and away from the injured patient's claim. The needs of the patient are made primary and are immediately addressed by the risk pool; who-did-what is a secondary concern.

Risk pooling would only be ethically and professionally sound if schemes are fit for use. Justice for patients would not be achieved if the compulsory insurance component of risk pooling were used to rationalise the rapid deployment of AISs just because insurance exists to cover the risks. Nevertheless, if the risk pooling approach were to be adopted, insurance companies would drive up safety standards to protect their business. For example, in the UK, insurers would presumably be reluctant to enter commercial partnerships where proof of compliance with standards set by organisations such as NICE, MHRA and NHSX is absent. In cases where the benefits of the AIS may be apparent, but the insurance companies decided risks were not profit worthy, the state may need to be involved to ensure clinicians may access an AIS which would be otherwise un-insured.

Multi-stakeholder ethical governance

Organisations such as NICE, MHRA and NHSX represent national standard setting in healthcare; to compliment this, there could be additional multi-stakeholder ethical governance at a local level.

Winfield and Jirotka (2018) define ethical governance "as a set of processes, procedures, cultures and values designed to ensure the highest standards of behaviour." They recognise ethical governance as part of the practice of responsible research and innovation (Winfield and Jirotka, 2018). As such, ethical governance allows stakeholders to identify the risks of deploying an AIS, track its operations, and be assured that it operates as intended (Falco *et al*, 2021); subsequently, ethical governance is necessary to reduce negative incidents and promote trust in systems (Theodorou and Dignum, 2020). Within the context of this thesis, ethical governance would amount to the prospective recognition of issues before they arise rather than relying on a retrospective approach of solving problems after they have arisen. It is important that ethical governance structures are in place prior to the selection and deployment of an AIS in clinical practice to aid its safe adoption into healthcare; without this, public and clinical confidence in AISs generally could be undermined by even only one major mishap involving AIS use (Reddy *et al*, 2019).

Chapter four’s literature review noted the large volume of recently published ethical guidance for AI. This body of literature creates the problem of identifying and then transforming relevant ethical principles (i.e., good intentions) to good practice; on their own, principles are but signposts: they are only effective if they are followed (Eitel-Porter, 2020). As such, we need to move on from high-level statements about AI guidelines and towards more practical approaches (Theodorou and Dignum, 2020).

To achieve this, the risk pool agreement could mandate that the use of an AIS is only permissible in clinical practice if it is demonstrably compliant with relevant ethical principles as per a model of multi-stakeholder ethical governance. Theodorou and Dignum (2020) recommend the use of ethics boards when AI technologies are considered for use; these would work comparably to other similar structures such as university ethics boards which approve or veto research projects. Reddy *et al* (2019) take this one step further for healthcare applications by suggesting governance committees need to be populated by clinicians, managers, patient group representatives, and technical and ethics experts. The ‘multi’ characteristic of a multi-stakeholder ethical governance committee is extremely important; inclusive and comprehensive stakeholder involvement promotes an environment where an effective 360° assessment can be performed to identify and mitigate issues when considering AIS deployment. Arguably, to ensure relevancy, the establishment of a multi-stakeholder ethical governance committee needs to be specific to the proposed site of deployment to the AIS, not only geographically to the institution in which the AIS is proposed for use, but also in terms of clinical and technological speciality and representative of the patient population which it is aimed to serve.

As aforementioned, the noting of principles alone is not enough to guarantee that the deployment of AISs is ethical (Mittelstadt, 2019), however principles can provide a common language between stakeholder groups upon which ethical issues can be identified and subsequently addressed (Mittelstadt, 2019). A multi-stakeholder ethical governance committee could determine which ethical principles need to be applied to a given clinical use of an AIS; for example, selecting which of Jobin *et al*’s (2019) identified principles of transparency, justice and fairness, non-maleficence, responsibility, and privacy is relevant to that particular AIS use. Jobin *et al*’s (2019) five principles are not conclusive though, a committee may choose to supplement these with additional principles, e.g., autonomy, as per the group’s determinations. The principles which are chosen to be addressed also need to speak to the values of those who use the system (Falco *et al*, 2021) and those who are affected by its use. An example value here could be borrowed from aviation; rather than ‘save the lives of passengers’ (Falco *et al*, 2021, p.570) the required value of the AIS would be to ‘save the life of the patient’. This is no small task as there is no unified approach of implementation for the practical translation of ethical principles into practices (Mittelstadt, 2019). This means that, currently, each committee is free to make these translations of principles into practice differently; instead of encouraging a unified approach, the application of ethical principles would be individualised to the locality of the AIS proposed for use. If the committee is not satisfied that issues related to key ethical principles have not been adequately addressed, then they are in a position to safeguard society's interests, i.e., clinical practice and patient safety. This would be by either denying the AIS's deployment in the healthcare setting or by recalling it from practice if its use is found to later to not meet the requirements of a key ethical principle to the satisfaction of the committee. Eitel-Porter, (2020) notes that reusing governance structures within single enterprises (in the context of this thesis, the 'enterprise' would refer to the NHS) stops the need of duplication of effort or the introduction of competing forms of governance within an organisation, but Mittelstadt (2019) argues that the use of 'bottom-up' approaches has value as new challenges for AI ethics are revealed by novel cases which can be to the educational benefit of all involved in the field. Given the massive size of the NHS and the enormous variety of potential AIS applications within the enterprise, the initial committee creations may benefit from being set up to reflect local requirements. The advantage of ethical governance model being adopted locally is that it would enable the bespoke fitting of an AIS to an individual clinical service and the patients it serves. This would be appropriate as, for example, the use of AISs in psychiatric care would have different requirements to that being used in oncology services. To meet those needs, input would be needed from different stakeholder groups such as affected clinical professionals and patients as well as technologists with expertise of the proposed AIS model. These committees could be subject to planned future general reviews to determine which committee structures and processes work best overall, which would then inform further committee design iterations. Such reviews may then result in a unified ethics governance model and streamline governance practices – this would be particularly important for parity in the instance of SDCs/clinicians finding it harder to satisfy the requirements of one ethics governance committee to allow AIS deployment in one location than another.

It is outside of the scope of this thesis to comprehensively identify and demonstrate how all ethical principles could be applied to AIS development, deployment, and use, but, to offer a brief single example, one central ethical principle which most committees may determine as needing to be satisfied might be the principle of transparency. This would involve the adoption of appropriate practices which enable actors to value the principle of transparency throughout a system's lifecycle: e.g., the transparency of process of an AIS's development by the SDC, the transparency of product itself (Winfield and Jirotka, 2018) (thereby beginning to address the problem of system opacity), and an account of the final AISs use by clinicians in their decision-making would need to be available for inspection. An ethical governance committee may choose to follow Falco *et al*'s (2021) AAA governance model of Assessments, Audits, and Adherence. This three step model would enable the committee to evaluate how the principle of transparency and its associated issues had been considered and managed throughout its use lifecycle: during development, at the point of deployment, and through its time in the clinical environment. The advantage of an ethical governance committee insisting on transparency would be the benefit of assessing the AIS for suitability prior to deployment for issues such as ensuring the minimisation of bias (Reddy *et al*, 2019). Transparency of the system and the activities of all causally related stakeholders would also make the task of performing retrospective analysis of incidents easier if harm were to eventuate from AIS use. Whilst an instant criticism of a multi-stakeholder ethical governance committee approach could be that, again (as noted in chapter 5), ethics has no teeth, the teeth can come from the risk pool itself - that the AIS can't be deployed without the local multi-stakeholder ethical governance committee's approval. Therefore, in the example of the principle of transparency, if an AIS were deployed and it were found that the AIS itself, it's developer's practices or user's practices were not transparent enough to ensure that it's use was ethically safe, then the committee could reserve the right to recall it from practice.

A multi-stakeholder ethical governance committee whose members included clinicians and technologists would be a strong practical example of the prospective sharing of responsibility prior to the deployment of an AIS. The prospective nature of this arrangement would be that the committee would ensure that the AIS satisfies the ethical conditions that they have jointly chosen to safeguard patients and patient care. To ensure strong ethical governance, Winfield and Jirotka (2018) suggest that those party to this process:

1. have available to them an ethical code of conduct which they can refer to,
2. have responsible research and innovation training,
3. practices that training by performing an appropriate ethical risk assessment on each AIS prior to deployment,
4. be transparent about the nature of their ethical governance via publication of their activities,
5. value ethical governance as a core value throughout the organisation.

Theodorou and Dignum (2020) offer that computer science practitioners need to be trained in areas such as ethics, safety, system transparency. Other non-technologically specialised members of a multi-stakeholder ethical governance committee would also arguably benefit from training of all these issues too so that they can collectively appraise proposed AISs.

Whilst Winfield and Jirotka (2018) note that valuing ethical governance would be difficult to evidence, yet this may well be observable through the transparency of a governance committee's publications. I also would venture that valuing ethical governance would be something that the clinical professions would especially choose to champion, as its goals would be in alignment with their established professional codes of conduct. Enthusiastically valuing ethical governance as a core value would be in the interests of SDCs to value ethical governance too if the deployment of an AIS is contingent on the successful evaluation of a multi-stakeholder ethical governance committee.

Whilst the above has discussed ethical governance within the structure of a committee, to be effective, ethical governance would need to be in everyone's remit. Eitel-Porter (2020) speaks of ethical fire wardens who are trained to raise the alarm within organisations, but, given the huge number of people who AISs in healthcare could affect, anyone needs to be able to push the alarm button to summon help. It ought not be only for committee members to raise alarms if they have reason to believe that an AIS is not fit for purpose; indeed, any person with concerns about the practical development, safe use, or ethical implications of an AIS in healthcare ought to be able to raise those with the committee.

As standards and governance models become more detailed and based on empirical evidence, the premiums for the risk pool could conceivably reduce (though this is essentially a question for actuaries). If the cost of insurance is prohibitive because the risks are very high, this financial inconvenience could be welcomed as it will prevent risky and/or harmful AISs from being used in clinical practice at a time when those risks cannot be effectively mitigated. Yet, this could result in a gap between risk of non-profitability and risks of harm resulting in patients missing losing the opportunity to benefit from AIS.

Collectively, complex negotiations will be required to achieve a nexus of contracts which are fair to all stakeholders. Such discussions could be guided by those who have experience in wrangling issues of moral justice. Allow me to touch here on the potential for a role here for an AI ethicist.

AI Ethicists

Away from the context of insurance considerations, Gambin (2020) describes the difficult contemporary role of ethicists who specialise in advising in AI development and deployment in industry. In Gambin's (2020) definition, an AI ethicist applies abstract concepts to concrete situations; they use this skill to determine right and wrong in the development and use of AIS. An industry AI ethicist could advise either an SDC or a clinician to ensure that an AIS is both legally compliant and ethically satisfactory prior to deployment, but the impartiality of a single AI ethicist's counsel could be called into question if they are employed by only one stakeholder. If the ethicist were employed to advise only one stakeholder group there would be an obvious conflict of interest; this would interfere with their ability to facilitate fairness across all interested and concerned parties. Gambin (2020) speaks of the need for the AI ethicist to be brave in their role when advising SDCs. Whilst one may agree that this may be necessary if an employee must tell their employer that the AIS which they had developed was not fit for deployment, braveness ought not be a necessary quality in a shared responsibility model. Rather, an AI ethicist should simply be free to express key concerns about the allocation of responsibility (or any other ethical matter) which may affect any stakeholder group without fear of reprisal, either for themselves or for the other stakeholder groups which they represent. Freedom and ability to speak truth to power is critically necessary in this role so that the ethicist might fairly represent all stakeholder groups as they would be impartial in the weight given to a stakeholder's interests. For this reason, I believe that, whilst SDCs (or any stakeholder group for that matter) may employ and be advised by AI Ethicists, there also ought to be impartial AI Ethicists available as one of the services (alongside actuarial) which could be delivered by a risk pool. If the risk pool benefitted from an AI ethicist's advice, the positions of all stakeholders could be represented and expertly considered, rather than adroit arguments being posed by only those who could afford to employ ethicists.

Risk pooling would complement professional regulation

As a final benefit to highlight, risk pooling would not detract from the professional regulation which governs some of the actors. Clinical users would still be required to maintain membership of their profession to be able to practice; SDCs and technologists may also separately pursue their own regulation. The work of the multiagency advice service would augment that of the risk pool by helping gather information about the status and appropriateness of the use of an AIS in the clinical environment when the risk pool is being negotiated. Risk pooling could work alongside and compliment regulatory structures rather than replace them. As a result, the clinician's and the SDC's duty of care to patients will not be degraded, and instead enhanced through the additional safeguarding of the interests of patients which the risk pool could offer.

Whilst there are several benefits to risk pooling, there are also accompanying disadvantages.

Problems with risk pooling

'Moral hazards' exist when actors behave differently because they are protected in some way from the costs of that behaviour (Rowell and Connelly, 2012). The problem of moral hazard would need to be strongly addressed prior to the adoption of AISs in the clinical area. Just because a risk pool is in place does not mean that actions are free of responsibility (Merkin and Steele, 2013). Thus, if an AIS risk pool for all concerned stakeholders is in place and there is still a potential risk of injuring a patient through using the AIS, the provision of insurance cover which benefits a harmed patient does not morally excuse the deployment of an AIS when unjustified and disproportionate foreseeable harm could take place. AISs are tools and human actors must not pass responsibility for harms to their tools; intentionally shifting personal responsibility for harms from the use of an AIS to a risk pool could be seen as similarly responsibility-avoidant behaviour. But whilst the use of AISs in clinical decision-making is currently novel, the use of insurance to offset non-intentional risk has long been accepted by society (e.g., vehicle insurance). If society accepts the use and risks of using AI in clinical decision-making, it is conceivable that risk pooling would be accepted here too for non-intentional harms; the incentive for an actor to comply as far as possible to the terms to the insurance would be to avoid higher premiums.

There is no reason for the provision of a risk pool to prevent an injured patient from pursuing a legal route to claim from an individual stakeholder as there could be any number of reasons why a risk pool claim may be inappropriate for a patient (something could have been unforeseen when the risk pool was constructed and agreed). There is nothing to prevent a patient making a claim via the courts for harms caused, unless the right to sue for tort is prohibited when a risk pool is in place (e.g., similarly to New Zealand's ACC scheme). However, it would be in the interests of all stakeholders to ensure that the threshold for an injured patient to successfully claim from a risk pool would be low enough to achieve a convenient, swift, reasonable, and fair claim for an injured patient whilst rigorous enough to prevent spurious claims from threatening the financial integrity of the scheme.

Similar to the criticism noted of enterprise liability, if the wording in a risk pool's insurance policy does not cover what it intends to, such as the tort of negligence, the insurance might not benefit the injured patient; thus, negating the purpose of insurance. Great care would need to be taken to ensure that the risk pool's arrangements serve the injured patient rather than excluding them and that an appeal process is available should a patient's claim be unsuccessful.

The contribution to the risk pool by NHS clinical users would be met by taxpayer funding via UK national funding sources. Additionally, the cost of the SDC's contribution to the risk pool would likely also be passed on to the NHS (and consequently the taxpayer) through price increases. These costs would need to be ascertained prior to engagement; if costs can be recovered by insurance schemes, then the financial impact on the NHS could potentially be minimised.

Risk pooling may need to be mandated or else risk being yet another patchy answer to what may become a universal problem if AISs are used throughout healthcare in the future. However, if risk pooling were mandated for adoption in England and Wales, the multiagency advice service would be in a perfect position to signpost SDCS and clinicians to resources for developing their risk pool.

Finally, risk pooling would not be able to prompt the adoption of personal moral responsibility in the same way that professional regulation of individual clinicians can. However, the other members of a risk pool and the insurer who is underwriting it may refuse to enter a risk pool agreement if the multiagency advice service does not provide the appropriate signoff for the AIS's use in clinical decision-making. In this way, again, the SDC would be forced to demonstrate prospective moral responsibility to the multiagency advice service and the members of the risk pool before their AIS could possibly reach the patient with a clinician's involvement. A lack of multiagency advice service sign-off could amount to the NHS not adopting the AIS, therefore keeping patients safe from potential harms. Whilst this is a high hurdle for an SDC to clear for their AIS to be adopted, it is not unreasonable if their system is to enter the sensitive and high-stakes environment of clinical decision making.

Next steps: consultation before AIS adoption

The use of AISs in clinical decision-making is a shift away from the traditional methods which clinicians currently employ. The ethical and legal allocation of responsibility explored in this thesis is but one facet of multiple issues which need to be addressed before the implementation of this change in the core methods of how clinical decisions are made (others include issues such as privacy and bias (Reddy *et al.*, 2019)). Change is not necessarily bad, but it needs to be carefully governed (Reddy *et al.*, 2019). Whilst this thesis is chiefly concerned with how ethical and legal responsibility is considered with the adoption of AISs which influence clinical decision making, there is value in comprehensively seeking and answering to the views of affected stakeholder groups before that adoption and before the appropriate governance structures are formed. Gaining and incorporating societal views will increase the chance of an AIS being accepted; without acceptance from key stakeholders that an AIS is beneficial, the chance of AIS adoption is low.

Indeed, this is a strategy which the NHS AI Lab have taken whilst setting up their multiagency advice service for AI technologies in health and social care (NHSX AI Lab, 24 September 2020). They are currently in the process of performing "audience research" to "identify and address areas where regulation is unclear or ineffective". (NHSX AI Lab, 24 September 2020). They have already identified that SDCs are concerned about which party is accountable for what when AIS devices are implemented, and that there is a need for clarity over arrangements in place between developers of AIS and providers of health and social care (NHSX AI Lab, 24 September 2020), thus there is scope for stakeholders to explore this thesis's suggestion of utilising a shared model of responsibility.

Consultation is a highly necessary first step to determine stakeholder views on issues such as the acceptability of risk in using AIS in clinical decision making, how that risk should be managed and how responsibility for that risk could be shared. Dialogue should not be focussed on optimising the journey for AIS deployment to benefit SDCs. Instead, consultation should be an opportunity to investigate how deployment of effective AISs can be achieved in the most responsible manner which respects all affected stakeholders and protects patients.

As I noted in chapter 6, Edwards and Deans (2017, p.61) describe the Rawlsian account of ethics as one where an ethicist would “play a substantial role in specifying and applying the content of public reason.” To only use this approach would be prescriptive to rather than inclusive of stakeholders. To avoid excluding stakeholders, consultation could be used to reflect Edwards and Deans’s (2017) suggestion of a Habermasian (1990) approach. Consultation can be created to be a structured environment which encourages direct stakeholder inclusion by promoting self-representation, communication, and negotiation between stakeholders.

Consultation

For risk pooling (or any other contractarian solution) to be a desirable and potentially acceptable way forward for stakeholders, detailed investigation into the potential contents of the necessary contracts will be needed. To achieve this, there would need to be consultation with bodies such as clinicians and the organisations which they work for, lawyers, actuarians, SDCs, patient groups, clinical professional and products regulatory bodies. This consultation could be designed to determine the aims and principles which a scheme could adopt and begin to recognise the values and needs of affected stakeholders.

Consultation of clinicians and the organisation which employ them

Taking into account all aspects of this thesis’s work, I suggest that it is very much in the interests of clinicians to become keenly interested and very involved in any consultation in this area. This interest ought to be at all levels, e.g., individual, representative professional bodies such as trade unions, and within the organisations in which they are employed. To not do so leaves clinicians and/or the organisations in which they are employed voiceless within a consultative process, thus at risk of being unfairly allocated responsibility for using an AIS when it might be possible to share that responsibility with the SDC whose AIS has influenced their clinical decision-making.

Kalluri (2020) questions how applications of AISs shift power in society, and that those affected by AIS projects should be involved in its creation. Clinicians might not initially realise this, but *they are in a position of power*. Whilst it may appear that SDCs are able to dictate to clinicians the terms of AIS use, clinicians do not have to give up their power by agreeing to those terms. If clinicians refuse to adopt AISs *en masse* the AISs cannot reach patients. However, such an action can only fall in clinicians’ favour if they act together to reject an unsafe AIS, for example through a united front which incorporates organisations which represent clinical practice, such as the trade unions and Royal Colleges.

Consultation of SDCs and technologists

SDCs and technologists might not understand the benefits of a contractarian approach at the outset of a consultation if it has been historically perceived that clinicians carry the burden of responsibility for all clinical decision-making. Consulting SDCs and technologists as to their understanding of their responsibility towards the patients whose lives they wish to influence, and how they wish to discharge that responsibility, may help them to understand that the burden of responsibility is also theirs. If they come to appreciate that clinicians and patients may reject the use of AISs if responsibility is not carried by causally responsible actors, then SDCs and technologists may choose to then engage with the contractarian process on offer.

Consultation of patient groups

Consultation of patients who represent those who will be directly affected by clinicians using AIS in their care is vital to ensure that their needs and opinions are noted and later represented in negotiations. There is no guarantee that even a well negotiated risk pool will be accessed by patients if they needed to make a claim. If it is desirable that patients appropriately access risk pools, there is an essential need to ensure that the construction and processes of the risk pool is attractive to the patients who will use it. If the arrangements provided by a risk pool fail to gain patient support by meeting foreseeable and potential needs, there is scope for patients to decline to allow AISs to be used in their care decisions, or, if they do permit the use of AISs and are harmed as a result, the patient may turn to the traditional and non-tested legal routes for restitution (as speculated in chapter 5). Both outcomes would disappoint clinicians who wish to gain the aid of AISs and SDCS who wish their AISs to be adopted.

Legal advice should be sought

This thesis's legal analysis is vastly incomplete for this scenario. Negligence is but one legal area of consideration; strict product liability and contract law as well as regulatory considerations are non-exhaustive examples of other areas which shall require analysis prior to the adoption of AISs in clinical decision-making.

Every stakeholder who might be affected by the introduction of AISs in clinical decision-making (especially patients and clinicians) would be wise to seek competent legal advice to ensure that their needs are met by the risk pool and not neglected.

Consultation of relevant specialist bodies

Whilst actuaries will not be involved in the deployment or use of the AIS in question, their skills will be required to assess the potential risk of using the AIS and to help cost the risk pool. Regulatory bodies such as NHSX AI Lab's multiagency advice service could advise the minimum standards of products and practice which actors must achieve to be able to benefit from the protective coverage which a risk pool might offer.

Negotiation

With the findings of consultation work, an initial risk pool proposal may be drafted to provide a starting point for further negotiations before an AIS is deployed into practice. Subsequent examination and explicit discussion of the prospective allocation of responsibility for the outcomes of using AISs in clinical decision-making will allow the determination of the sharing or insuring of costs of potential harms to be fairly shared between stakeholders who are causally responsible for its effects. The provision for patients affected by AIS use is the first concern and ought to be the prime aim of any risk pool, but this ought not be to the detriment of other stakeholders. Where responsibility is negotiated conscientiously, purposefully, deliberately, and unanimously, there is more opportunity for that process to result in fairer allocation of its burden to all involved stakeholders. As earlier identified, whilst Rawls's veil of ignorance cannot be literally deployed to negotiations, parties may be encouraged to consider the viewpoint of every other group involved. Additionally, this is in the spirit of the Habermasian approach, noted earlier, where negotiations will be structured so that participants may mutually recognise one another's positions.

Review

To ensure ongoing fairness to its members, a risk pool could be arranged so that any agreement which has been negotiated can be intermittently revisited, reviewed, revised, and updated according to its effectiveness. For example, the contribution of actors to the risk pool might need adjusting: SDCs may find that their AIS is more or less accurate than initially realised, clinicians may find that they make more or fewer mistakes than originally estimated, the harms which patients experience may be more or less expensive than originally projected. SDCs may eventually develop their AISs to be usable directly by patients without the need for clinical supervision, and this development may be viewed by clinicians as grounds to leave the risk pool altogether.

Now that this thesis has set out a framework by which responsibility for the use of AI in clinical decision-making may be shared, the next step would be to test that idea for the feasibility of its resolutions. An empirical bioethics approach may provide a good platform upon which to proceed.

Empirical work

Much contemporary ethical research is performed using empirical methods. A significant criticism of this thesis could be that I have consulted no stakeholders during the formation of ethical theory. Yet, rather than excluding the value of empirical approaches, this work provides a theoretical foundation from which empirical work might be later launched. Whilst clinicians and SDCs/technologists have been central to this thesis, consulting *all* stakeholders who would likely be affected by the adoption of AISs in clinical decision-making is necessary before attempting to develop and propose any kind of scheme; this is especially so when considering a scheme which depends on a contractarian approach.

Empirical work is valuable here as it would capture the voices of affected stakeholders; their views could inform of the development of a risk pool approach that reflects their needs. Stakeholder input could challenge, validate, or reject the ethical reasoning which underpins the approach that has been set out in this thesis. Importantly, empirical work would detect if there were any desire to proactively recognise and allocate prospective and retrospective responsibility to actors using a contractarian approach, and, if risk pooling were not desirable, empirical work would provide a space to stakeholders to what solutions would be preferred.

Moving forward, I would like next to develop a qualitative research project to test this thesis's recommendations. I would be tempted to use *reflective equilibrium* and employ stakeholder's contributions to reconcile "(a) a set of considered moral judgments, (b) a set of moral principles, and (c) a set of relevant background theories" (Daniels, 1979, p.258). To achieve this, I would design a pilot study with a small, selected sample of expert interviews to test this thesis's ideas with. I would look for participants with clinical experience in using AISs, SDC/technologist who had developed AISs for clinical decision-making, those with policy backgrounds, or persons who might span all three of these groups (e.g., employees of NHSX). I would present the analysis and recommendations which this thesis has set out and I would ask these key stakeholders if they felt its contents were plausible and if risk pooling had the potential to be a feasible scheme to consider in the future. I would expect to find that stakeholders may find some of this thesis's contents acceptable, but some might be found unpalatable, especially if a recommendation resulted in a stakeholder having to accept responsibility when they previously had not. By working backwards and forwards through the stakeholder's comments, I would be able to consider and reach an equilibrium between judgments, moral principles, and background theories (Daniels, 1979).

If this pilot project indicated that this thesis's analysis and recommendations were plausible, then a larger project may be considered where the participants might be less specialised specifically to AI, but more representative of the affected stakeholder populations. Here, their input may allow the ethical theory to be further developed and refined. From this point, if still plausible and feasible, translational ethics may be considered; how would this refined idea be introduced into clinical practice?

Conclusion

The ethics chapter 6 ended by stating that there is an opportunity to incorporate practical discussions of how stakeholders can work more closely together to prospectively prevent harms from eventuating, as well as planning how problems will be addressed retrospectively should they happen. I have not attempted to create a formula which could be used to solve how much moral responsibility is carried by each stakeholder in any given scenario, as the potential scenarios are numerous and complex. However, now is the time to start to consult, negotiate, and fairly balance responsibility between stakeholders when considering the use of AISs in clinical decision making. Rather than sleepwalking into uncharted legal claims territory, if carefully considered and negotiated, stakeholders may calmly and intentionally plan how patient injuries could be handled should they eventuate. This would allow stakeholders to start as they mean to go on by making good of the opportunity to take ownership of their actions and to accept, plan, and act upon their personal moral responsibility for consequences arising from their actions.

The discussion in this chapter has explored the possible paths which could be taken when stakeholders embrace the prospective and retrospective responsibility which they owe to the patients who would be affected by the use of AIS in clinical decision-making. Solutions which assign responsibility to individual stakeholders have been rejected in favour of a holistic approach of shared decision-making. The solution of risk pooling has been introduced as a model of the united adoption of responsibility by stakeholders who have developed, deployed, used, or are finally directly affected by the proposed AIS. Communication between stakeholders has been proposed to achieve the nexus of contracts (*à la* Rawls) required to create a functioning risk pool. The risk pool method would aim to address the needs of patients harmed due to the use of AIS in their care, whilst permitting the perceived overall larger benefit of using AIS in healthcare. The shared responsibility model has in no way diminished the professional responsibility carried by the clinical decision-maker, but instead has increased the visible involvement of the SDC at the bedside via the proposed deployment of their AIS. This model allows the SDC to step forward with clinicians in partnership and plan how they can carry their retrospective responsibility for the prospective duty of care which they owe the patients to which their AIS will affect. Risk pooling offers the opportunity to create a space for stakeholders such as clinicians, SDCs, and patients to come together and discuss the risk and benefits of adopting AIS in the clinical environment prior to its deployment.

Chapter 8: Conclusion

Due to the imperfect nature of human clinical decision-making, there may be a place in the future for artificially intelligent systems to aid healthcare professionals. Although there has been much discussion and preparation at both the level of Government and within NHS in England and Wales regarding the introduction of artificially intelligent systems into the clinical environment, little addresses the potential aftermath of any potential negative consequences of its use. This thesis has contributed to the body of knowledge by offering analyses of the determination of the allocation of ethical and legal responsibility in the scenario of a clinician using an AIS to aid their decision-making and patient harm eventuating as a result of that use. Both analyses inform the proposed shared model of responsibility of risk pooling which is designed to facilitate a fair distribution of responsibility to all causally responsible stakeholders.

I started by describing how clinicians currently make decisions, that human powered clinical decision-making is problematic, and that AI might be employed to help. I described what AI is and that there were associated issues with its use. Specific issues were purposefully chosen and expanded upon; these were based on the problem of opacity in an AIS's processes impacting its user's ability to account for their choice to use it. Subsequently, if a patient came to harm as a result of a clinician using an AIS's recommendation, the question arose of who would be responsible for that harm and the associated legal liability.

My first step to investigate these issues was to perform a narrative literature review supported by a systematic approach. This work found that a clinician's use of opaque AISs in decision-making could affect their ability to practice accountably. There was a consensus that clinicians ought to hold responsibility for using AISs in their decision-making, but there were indications that this responsibility could be shared with the SDCs who designed and deployed them. The literature also identified that there is a lack of case law or legislation specific to negligence when AISs are used in the clinical environment, and that waiting for courts to make such a ruling would require a patient to be harmed through AIS use first.

These findings gave me two avenues of novel analysis to explore. Firstly, to examine the actions of SDCs and clinical users, where my legal analysis explored fault-based liability, i.e., how the tort of negligence might be applied. Secondly, ethical analysis to determine how ethical theory could inform the allocation of responsibility and to use that analysis to ethically challenge the legal analysis's findings.

My legal analysis found that, when considering the tort of negligence in the context of England and Wales, both SDCs and clinicians might be awarded a duty of care for patients affected by the use of AIS in clinical decision-making. However, clinicians would most likely be found causally responsible for harms for two main reasons. Firstly, because the action of a clinician using an AIS recommendation would be noted as the new intervening act which allows harm to reach the patient. Secondly, because the use of an unsuitable AIS recommendation might be judged a failure of the clinician to reach the standards of other professionals. The potential of *novus actus interveniens* seems to protect the SDC while leaving the clinician vulnerable to negligence claims, and this seemed unfair to clinicians. This apparent unfairness underlined the need for ethical analysis.

My ethical analysis applied respected theoretical ethical frameworks to the scenario of AIS use resulting in patient harm. I found that personal moral responsibility can be awarded both prospectively and retrospectively to both clinicians *and* SDCs, and that there was scope in certain situations for this responsibility to be shared in the case of harm eventuating due to the use of AISs in clinical decision-making. It argued that SDCs are unfairly treating clinical users as moral crumple zones if they allow the clinical user to be singly allocated and burdened with responsibility for harms when responsibility could be fairly shared between clinical users and SDCs. These findings indicate that there is an imbalance in legal responsibility when its allocation is considered through an ethical lens.

Whilst the literature review suggested, and my ethical analysis confirmed, that responsibility *could* be shared fairly, a clearly defined model which allowed stakeholders to be treated fairly in this context was yet to be suggested. My final substantive chapter presented the strengths and weaknesses of different solutions which currently exist before identifying a contractarian-based approach which allows actors to embrace both prospective and retrospective responsibility for their actions. I have suggested regulating SDCs and technologists to enable them to formally adopt a duty of care for the patients whose care they seek to influence. I have also built on Allain's (2013) suggestion of enterprise liability and developed it further into the fairer contractarian-based approach of risk pooling; my approach proactively employs stakeholder discussion and collaboration to create risk pools which are constructed to provide for patients who have experienced harm due to the use of AISs. Risk pooling is an example of a contractarian solution which might be applied to the problem of clinicians being used as moral and legal crumple zones. When risk pools are devised, stakeholder consultation will be essential to achieve the spirit of a contractarian approach; this is to ensure that all groups' needs have been met and that concerns have been addressed prior to the adoption of AISs in clinical decision-making.

To be achievable, a contractarian approach such as risk pooling would require a great deal of consideration, consultation, and communication between several stakeholders. I have outlined that such a consultation must not be limited to SDCs, clinicians and the organisations for which they work for; it should also involve stakeholders such as patient groups, regulators, policy makers, and the actuarians who would calculate the pool as well as those who may manage any claims which are made, for example practicing lawyers.

To test the plausibility of the shared model of responsibility and the suggestion of risk pooling, I have outlined an initial plan for empirical and consultation work to establish if stakeholders find my analysis and suggestions for solution to be potentially favourable or not.

Finally

This thesis has served to raise the problem of the allocation of ethical and legal responsibility and I hope that its suggestion of the employment of a contractarian approach along with stakeholder regulation serves to create a starting point for further analysis, discussion, and empirical work which identifies a fair way forward for all stakeholders should AISs ultimately be adopted to aid clinical decision-making.

When I commenced this body of work it appeared that the introduction of AI would be inevitable and that there is a strong potential for clinicians to unfairly shoulder the burden of responsibility for its use, but this does not have to be. Rather than passively allowing AISs to be introduced and proliferate in healthcare, the consequences of its adoption ought to be examined and the terms of that adoption deliberately and consciously determined and agreed via fair negotiation between the stakeholders affected. This more considered route is arguably desirable as a lack of planning for foreseeable potential harms leaves patients who might be later affected without a clear path to restitution.

If clinicians insist on the employment of a contractarian approach, they will be well situated to use their newly recognised position of power to deny AIS adoption until an arrangement of the fair allocation of ethical and legal responsibility is agreed. The provision of a risk pool might ultimately benefit the patients who clinicians and the SDCs aim to serve whist preventing clinicians from becoming legal crumple zones in a negligence claim.

References

Cases

A v. Ministry of Defence [2005] QB 183, 203 (CA)

ABC v. St George's Healthcare NHS Trust [2017] EWCA Civ 336; 160 BMLR 19

ABC v. St George's Healthcare NHS Trust [2020] EWHC 455

Alcock v. Chief Constable of South Yorkshire [1992] 1 AC 310, 410

Bailey v. Ministry of Defence [2008] EWCA Civ 1052

Barker v. Corus UK Ltd [2006] 2 AC 572

Barnett v. Chelsea and Kensington Hospital Management Committee [1969] 1 QB 428

Barry v Cardiff and Vale University Local Health Board [2018] 12 WLUK 723

Bazley v. Curry [1999] 2 S.C.R. 534

Bevan Investments v. Blackhall & Struthers [1979] 11 BLR 78

Bolam v. Friern Hospital Management Committee [1957] 2 All ER 118

Bonnington Castings v. Wardlaw [1956] AC 613

Bolitho v. City and Hackney Health Authority [1998] AC 232

C v North Cumbria University Hospitals NHS Trust [2014] EWHC 61 (QB)

Caparo Industries v. Dickman [1990] 2 AC 605

Cassidy v. Ministry of Health [1951] 2 KB 343

Chester v. Afshar [2005] 1 AC 134

Cowley v Cheshire and Merseyside Strategic Health Authority [2007] EWHC 48 (QB)

Dann v. Hamilton [1939] 1 KB 509.

Darnley v. Croydon Health Services NHS Trust [2019] AC 831

Darnley v. Croydon Health Services NHS Trust [2018] UKSC 50

Daubert v. Merrell Dow Pharmaceuticals, Inc. [1993] 509 U.S. 579

De Freitas v. O'Brien [1993] 4 Med LR 281

Donoghue v. Stevenson [1932] AC 562

Fairchild v. Glenhaven Funeral Services Ltd [2003] 1 AC 32

Farraj v. King's Healthcare NHS Trust [2010] 1 WLR 2139 (CA)

FB v. Princess Alexandra Hospital NHS Trust [2017] EWCA Civ 334

Glasgow Corporation v. Muir [1943] AC 448, 457

Gold v. Essex County Council [1942] 2 KB 293 (CA)

Gregg v Scott [2005] UKHL 2

Hedley Byrne & Co Ltd v. Heller & Partners Ltd [1963] 2 All ER 575

Heil v. Rankin [2001] PIQR Q3

Horton v. Evans [2006] EWHC 2808 (QB); [2007] P.N.L.R. 17

Howmet Ltd v. Economy Devices Ltd [2016] EWCA Civ 847

Hughes v. Williams. [2012] EWHC 1078 (QB)

John v. Central Manchester and Manchester Children's Hospital University Hospitals NHS Trust [2016] EWHC 407

Jones v Conwy and Denbighshire NHS Trust [2008] EWHC 3172 (QB)

Kirkham v. Chief Constable of Greater Manchester [1990] 2 QB 283

Lister v. Romford Ice and Cold Storage Co. [1957] A.C. 555

Luxmore-May v. Messenger May Baverstock [1990] 1 All ER 1067

M v. Calderdale & Kirklees HA [1998] Lloyd's Rep Med 157

Maynard v. West Midlands Regional Health Authority [1985] 1 All ER 635

McFarlane v. Tayside Health Board [2000] 2 AC 59

McLoughlin v. O'Brian [1983] 1 A.C. 410

Michael v. Chief Constable of South Wales [2015] AC 1732, 156-158

Montgomery v. Lanarkshire Health Board [2015] UKSC 11

Nettleship v. Weston [1971] 2 QB 691

Page v. Smith [1996] 1 AC 155

Performance Cars v. Abraham [1962] QB 33

Perret v. Collins [1998] 2 Lloyd's Rep 255, 263

Philips v. William Whiteley Ltd [1938] 1 All ER 566

Price v Cwm Taf University Health Board [2019] EWHC 938

Rahman v Arearose Ltd [2001] Q.B. 351

Re A (Children) (Conjoined Twins: Surgical Separation) [2000] EWCA Civ 254

Rich v Hull and East Yorkshire Hospitals NHS Trust [2015] EWHC 3395

Robinson v. Chief Constable of West Yorkshire Police [2018] AC 736, [26]-[27]

Sienkiewicz v Grief [2011] UKSC 10 70

Simms v. Simms and An NHS Trust [2002] EWHC 2734

Smith v. Charles Baker & Sons [1891] AC 325

Smith v. Eric S Bush [1990] UKHL 1

Smith v. Littlewoods Organisation Ltd [1987] AC 241

Spencer v. Wincanton [2009] EWCA Civ 1404

Stovin v Wise [1996] AC 923

Webb v Barclays Bank plc and Portsmouth Hospitals [2001] EWCA Civ 1141

Wilkes v DePuy International Ltd [2018] QB 627

Wilsher v. Essex Area Health Authority [1987] QB 730 (CA)

Woodland v Swimming Teachers Association [2014] AC 537

All other references

Algorithm Watch, 2020. *AI Ethics Guidelines Global Inventory.* Last updated April 2020. Available: https://inventory.algorithmwatch.org/ Accessed: 4 November 2021

Allain, J. S., 2013. From Jeopardy! to Jaundice: The Medical Liability Implications of Dr. Watson and Other Artificial Intelligence Systems. *Louisiana Law Review* 73(4): 1049-1079

Allen, D., Harkins, K.J., 2005. Too much guidance? *Lancet*. May 21-27;365(9473): p.1768

Alper, B.S., Hand, J.A., Susan E.G., Kinkade, S., Hauan, M.J., Onion, D.K., Sklar, B.M., 2004. How much effort is needed to keep up with the literature relevant for primary care? *Journal of the Medical Library Association*. 92(4): p.429–437

Anderson, E.S., 1999. What is the Point of Equality? *Ethics*. 109(2): 287-337

Armstrong, K., 2018. If You Can't Beat It, Join It: Uncertainty and Trust in Medicine. *Annals of Internal Medicine.* 168(11)818-819

Association for Computing Machinery, 2018. *ACM code of ethics and professional conduct.* Available: https://www.acm.org/code-of-ethics Accessed: 4 November 2021

Automated and Electric Vehicles Act 2018, c.18

Ayres, I., 2008. *Supercrunchers: How anything can be predicted.* London: John Murray

Bainbridge, D. I., 1991. Computer-Aided Diagnosis and Negligence. *Medicine, Science and the Law*. 31:127-136.

Barrett, W., 2004. Responsibility, Accountability and Corporate Activity. *ON LINE Opinion*. Available: http://www.onlineopinion.com.au/print.asp?article=2480 Accessed: 4 November 2021

Bate, L., Hutchinson, A., Underhill, J., Maskrey, N., 2012. How clinical decisions are made. *British Journal or Clinical Pharmacology.* 74(4) p.614-620

Beauchamp, T. & Childress, J., 2013. *Principles of Biomedical Ethics*. 7th ed. New York: Oxford University Press

Bell, D., 2020. Communitarianism. *Stanford Encyclopedia of Philosophy.* Revised 15th May 2020. Available: https://plato.stanford.edu/entries/communitarianism/ Accessed: 4 November 2021

Bentham, J., 1983. *Deontology : together with A table of the springs of action and Article on utilitarianism.* Goldworth, A. (ed). Oxford: Clarendon Press.

Berlin, I., 1955-6. *Equality.* Proceedings of the Aristotelian Society LVI. p.301–326

Berner, E., Graber, M., 2008. Overconfidence as a cause of diagnostic error in medicine. *American Journal of Medicine*. 121(5): p.S2-23

Berwick, D., 2013. *A promise to learn– a commitment to act: Improving the Safety of Patients in England. National Advisory Group on the Safety of Patients in England*. Available: https://www.gov.uk/government/publications/berwick-review-into-patient-safety Accessed: 4 November 2021.

Beuermann, C., 2017. Do Hospitals Owe A So-Called 'Non-Delegable' Duty of Care to their Patients? *Medical Law Review.* 26(1): p.1-26

Bismark, M., Paterson, R., 2006. No-Fault Compensation in New Zealand: Harmonizing Injury Compensation, Provider Accountability, And Patient Safety. *Health Affairs.* 25(1): 278-283

Boden M., Bryson J., Caldwell D., Dautenhahn K., Edwards L., Kember S., Newman, P., Parry, V., Pegman, G., Rodden, T., Sorrell, T., Wallis, M., Whitby, B., Winfield A. (2011). Principles of robotics. Swindon, UK: Engineering and Physical Sciences Research Council. *Archived at The National Archives* Available: https://webarchive.nationalarchives.gov.uk/ukgwa/20210701125353/https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/ Accessed 9 May 2022

Boenink, M., Swierstra, T., & Stenmerding, D., 2010. Anticipating the interaction between technology and morality: A scenario study of experimenting with humans in Bionanotechnology. *Studies in Ethics, Law and Technology* 4(2)1-38

Bogost, I., 2017. 'Artificial Intelligence' Has Become Meaningless. *The Atlantic.* Available: https://www.theatlantic.com/technology/archive/2017/03/what-is-artificial-intelligence/518547/ Accessed: 4 November 2021

Bourne, T., Vanderhaeegen, J., Vranken, R., Wynants, L., De Cock, Bavo., Mike, Peters., Timmerman, R., Van Calster, B., Jalmbrant, M., and Van Audenhove, C., 2016. Doctors' experiences and their perception of the most stressful aspects of complaints processes in the UK: an analysis of qualitative survey data. *BMJ Open*. 6:e011711. Available: https://bmjopen.bmj.com/content/6/7/e011711 Accessed: 4 November 2021

Bourne, T., Shah, H., Falconieri, N., Timmerman, D., Lees, C., Wright, A., Lumsden, M.A., Regan, L., Van Calster, B., 2019., Burnout, well-being and defensive medical practice among obstetricians and gynaecologists in the UK: cross-sectional survey study. *BMJ Open*. 9:e030968. doi: 10.1136/bmjopen-2019-030968

Box, G. E. P., 1976. Science and statistics. *Journal of the American Statistical Association.* 71 (356):791–799

Box, G. E. P., Draper, N. R., 1987. *Empirical Model Building and Response Surfaces.* John Wiley & Sons: New York, NY.

Braun, V., Clarke, V., 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3(2):77–101

Brey, P. A. E., 2012. Anticipatory ethics for emerging technologies. *NanoEthics* 6(1),1–13.

Brito, J.P, Gionfriddo, M., Morris, J.C., Montori V.M., 2014. Overdiagnosis of Thyroid Cancer and Graves' Disease. *Thyroid*. 24(2): 402-403

Brownsword, R., 2008. Bioethics: Bridging from Morality to Law? In: Freeman, M. (ed.), 2008. *Law and Bioethics: Current Legal Issues*. Oxford: Oxford University Press

Bryson, J. J., 2019. New Job: Professor of Ethics and Technology. *Adventures in NI* Available: https://joanna-bryson.blogspot.com/2019/09/new-job-professor-of-ethics-and.html?m=1 Accessed: 4 November 2021

Bryson, J., 2020. “I do not understand why people think that talking about "automated decision making" or "autonomy" gets them around the conundrums of defining "AI". (I just take simple, well-defined definitions of "intelligent" and run from there...). *Twitter*. 13 October 2020. Available: https://twitter.com/j2bryson/status/1316065871505297408 Accessed: 4 November 2021

Bryson, J.J., Diamantis, M.E., Grant, T.D., 2017. ‘Of, for, and by the people: the legal lacuna of synthetic persons.’ *Artificial Intelligence and Law.* 25: 273–291

Bryson, J., Winfield, A., 2017. Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems. *Computer*. 50(5):116-119

Buchanan, B.G., Shortliffe, E.H., 1984. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA: Addison Wesley

Bucknall, T.K., 2000. Critical care nurses’ decision-making activities in the natural clinical setting. *Journal of Clinical Nursing*. 9: p.25-36

Cane, P., Goudkamp, J., 2018. *Atiyah’s Accidents, Compensation and the Law*. Cambridge: Cambridge University Press

Care Quality Commission, 2017. *Taking action*. Available: https://www.cqc.org.uk/what-we-do/how-we-do-our-job/taking-action Accessed: 4 November 2021

Caruana, R., Lou, Y., Gehrke, J., Kock, P., Sturm, M., Elhadad, N., 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. *KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* August 2015. 1721–1730 Available: https://doi.org/10.1145/2783258.2788613 Accessed: 4 November 2021

Castelvecchi, D., 2016. Can we open the black box of AI? *Nature.* 538(7623) p.20–23 6 October

Castelvecchi, D., 2016. Can We Open the Black Box of AI? *Scientific American*. Available: https://www.scientificamerican.com/article/can-we-open-the-black-box-of-ai/ Accessed: 4 November 2021

Champagne, D., Hung, A., & Leclerc, O., 2015. *The road to digital success in pharma.* Available: https://www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/the-road-to-digital-success-in-pharma Accessed: 4 November 2021

Chapman, E.N., Kaatz, A., Carnes, M., 2013. Physicians and Implicit Bias: How Doctors May Unwittingly Perpetuate Healthcare Disparities. *Journal of General Internal Medicine*. 28(11) p.1504–1510

Char, D.S., Magnus, D., Shah, N.H., 2018. Implementing machine learning in health care — addressing ethical challenges. *New Eng J Med* 378(11):981–983

Cheng-TekTai, M., 2013. Western or Eastern principles in globalized bioethics? An Asian perspective view. *Tzu Chi Medical Journal.* 25 (1): p.64-67

Civil Liability (Contribution) Act 1978 c.47

Clarke, A.C., 1973. Hazards of Prophecy: The Failure of the Imagination. In: *Profiles of The Future: An Inquiry Into The Limits Of The Possible* 12, (21) 1 (originally 1962, rev. ed. 1973)

Cochrane Library, 2021. *About the Cochrane Library*. Available: https://www.cochranelibrary.com/about/about-cochrane-library Accessed: 4 November 2021

Committee on Legal Affairs, 2017. *Report with recommendations to the Commission on Civil Law Rules on Robotics* (2015/2103(INL)). Strasbourg: European Parliament. Available: https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html?redirect Accessed: 4 November 2021

Compensation Act 2006. c.29

Consumer Protection Act 1987, c.43

Cox, P. N., 1997. The public, the private, and the corporation. *Marquette Law Review*. 80:2 393-530

Croskerry, P., 2002. Achieving quality in clinical decision making: cognitive strategies and detection of bias. *Academic Emergency Medicine*. 9(11): 1184-1204

Cudd, A., 2021. Contractarianism. *Stanford Encyclopedia of Philosophy*. Updated September 2021. Available: https://plato.stanford.edu/entries/contractarianism/ Accessed: 4 November 2021

Cumberlege, J., 2020. *First do no harm: the report of the Independent Medicines and Medical Devices Safety Review.* Available: https://www.immdsreview.org.uk/Report.html Accessed: 4 November 2021

Daniels N., 1979. Wide reflective equilibrium and theory acceptance in ethics. *Journal of Philosophy.* 76:256–82

Delvaux, M., 2017. *Report 27 January 2017; with recommendations to the Commission on Civil Law Rules on Robotics* (2015/2103(INL)) Available: https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html Accessed: 4 November 2021

Department for Transport, 2022. *The Highway Code.* Accessed 12 May 2022. Available: https://www.gov.uk/guidance/the-highway-code

Department of Health and Social Care, 2019. *NHSX: New Joint Organisation for Digital, Data and Technology*. Available: https://www.gov.uk/government/news/nhsx-new-joint-organisation-for-digital-data-and-technology Accessed: 4 November 2021

Department of Health and Social Care, 2021a. *£36 million boost for AI technologies to revolutionise NHS care*. Available: https://www.gov.uk/government/news/36-million-boost-for-ai-technologies-to-revolutionise-nhs-care Accessed: 4 November 2021

Department of Health and Social Care, 2021b. *A guide to good practice for digital and data-driven health technologies*. 21 January 2021. Available: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology Accessed: 4 November 2021

Department of Health and Social Care, 2021c. *Factsheet: Patient Safety Commissioner* Available: https://www.gov.uk/government/publications/medicines-and-medical-devices-bill-overarching-documents/medicines-and-medical-devices-bill-patient-safety-commissioner Accessed: 4 November 2021

De Vries R. & Gordijn B., 2009. Empirical Ethics and its alleged meta-ethical fallacies. *Bioethics*. 23(4) 193-201.

Dignum, V., 2019. *Responsible artificial intelligence: how to develop and use AI in a responsible way.* Switzerland: Springer

Dignum, V., 2018. The ART of AI — Accountability, Responsibility, Transparency. *Medium*. Available: https://medium.com/@virginiadignum/the-art-of-ai-accountability-responsibility-transparency-48666ec92ea5 Accessed: 4 November 2021

Doroszewski, J., 1988. Ethical and methodological aspects of medical computer data bases and knowledge bases. *Theoretical Medicine.* 9(2): p.117-128

Dorries, N., 2021. *Update on the Government's response to the Independent Medicines and Medical Devices Safety Review.* 11 January 2021. Statement UIN HCWS692. Available: https://questions-statements.parliament.uk/written-statements/detail/2021-01-11/hcws692 Accessed: 4 November 2021

Downey, A., 2019. Government pledges £250m for National AI Lab to improve diagnostics. *Digital Health.* Available: https://www.digitalhealth.net/2019/08/government-250-million-artificial-intelligence-lab-diagnostics/ Accessed: 4 November 2021

Downey, A., 2020. Two Midlands trusts sign deal with Babylon to provide Covid-19 app. *Digital Health.* Available: https://www.digitalhealth.net/2020/04/two-midlands-trusts-sign-deal-with-babylon-to-provide-covid-19-app/ Accessed: 4 November 2021

Dworkin, R., 1981. *What Is Equality? II. Equality of Resources.* Philosophy and Public Affairs 10 (1981):283–345 p.285.

Edwards, K.T., Deans, Z., 2017. Empirical Bioethics and the Role of the Professional Ethicist in Policy-Making: Politics, Authority and Expertise. In: Ives, J., Dunn, M., Cribb, A., 2017. *Empirical Bioethics: Theoretical and Practical Perspectives*. Cambridge: Cambridge University Press

Elish, M.C., 2019. Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology and Society.* 5: p.40-60

Elliott, R.A., Camacho, E., Jankovic, D., Sculpher, M.J., Faria, R., 2021. Economic analysis of the prevalence and clinical and economic burden of medication error in England. *BMJ Quality and Safety*. 30: p.96–105

Eitel-Porter, R., 2021. Beyond the promise: implementing ethical AI. AI and Ethics. 1:73–80 Accessed 10 May 2022. Available: https://link.springer.com/article/10.1007/s43681-020-00011-6

European Court of Human Rights, 1998. *European Convention on Human Rights.* Council of Europe: Strasbourg.

European Group on Ethics in Science and New Technologies, 2018. Statement on artificial intelligence, robotics and 'autonomous' systems. Brussels: European Commission. Available: https://op.europa.eu/en/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d-01aa75ed71a1/language-en/format-PDF/source-78120382 Accessed: 4 November 2021

Evans, J.B.T., 2003. In two minds: dual-process accounts of reasoning. *Trends in Cognitive Science*. 7(10) p.454-459

Evans, J.B.T., 2011. Dual-process theories or reasoning: Contemporary issues and developmental applications. *Developmental Review*. 31(2-3): p.86-102

Everett, J., 2021. From A-levels to pensions, algorithms make easy targets – but they aren't to blame. The Guardian. Available: https://amp.theguardian.com/commentisfree/2021/aug/17/a-levels-pensions-algorithms-easy-targets-blame-mutant-maths? Accessed: 4 November 2021

Evidence-Based Medicine Working Group, 1992. Evidence-Based Medicine; A New Approach to Teaching the Practice of Medicine. *JAMA*. 268(17): p.2420-2425

Expert Group on Liability and New Technologies, 2019. *New Technologies Formation, Liability for Artificial Intelligence and Other Emerging Digital Technologies.* European Union: European Commission. Available: https://op.europa.eu/en/publication-detail/-/publication/1c5e30be-1197-11ea-8c1f-01aa75ed71a1/language-en/format-PDF Accessed: 4 November 2021

Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., Danks, D., Eling, M., Goodloe, A., Gupta, J., Hart, C., Jirotka, M., Jonhson, H., LaPointe, C., Llorens, A.J., Mackworth, A.K., Maple, C., Pálsson, S.E., Pasquale, F., Winfield, A., Yeong, Z.K., 2021. Governing AI safety through independent audits. Nature Machine Intelligence. Accessed 11 May 2022. Available: https://doi.org/10.1038/s42256-021-00370-7

Farnan, J., Johnson, J., Meltzer, D., Humphrey H., Arora, V., 2008. Resident uncertainty in clinical decision making and impact on patient care: a qualitative study. *Quality & Safety in Health Care*. 17:(2) p.122–126

Faure, M., Partain, R.A., 2017. *Carbon Capture and Storage.* London: The MIT Press

Fenech, M., Strukelj, N., Buston, O., 2018. Ethical, social and political challenges of artificial intelligence in health. *Future Advocacy report for the Wellcome Trust*. Available: https://wellcome.org/sites/default/files/ai-in-health-ethical-social-political-challenges.pdf Accessed: 4 November 2021

Ferrira, A.P.R.B., Ferreira, R.F., Rajgor, D., Shah, J., Menezes, A., Pietrobon, R., 2010. Clinical Reasoning in the Real World Is Mediated by Bounded Rationality: Implications for Diagnostic Clinical Practice Guidelines. *PLoS ONE*. 5(4): e10265

Finnis, J., 2001. *Natural Law and Natural Rights.* Oxford: Oxford University Press

Floridi, L., Sanders, J.W., 2004. On the Morality of Artificial Agents. *Minds and Machines*. 14:349–379

Foot, P., 2002. *Moral Dilemmas: and Other Topics in Moral Philosophy.* Oxford: Oxford Scholarship Online. Available: https://oxford.universitypressscholarship.com/view/10.1093/019925284X.001.0001/acprof-9780199252848-chapter-7 Accessed: 4 November 2021

Frankena, W.K., 1973. *Ethics.* Englewood Cliffs, N.J., USA: Prentice Hall, Inc.

Freudenberg, W.R., 1992. Nothing Recedes Like Success? Risk Analysis and the Organizational Amplification of Risks. *Risk: Issues in Health and Safety* 3(1):1-35

Fuscaldo, G., 2006. Genetic ties: are they morally binding? *Bioethics*, 20(2): p.64-76

Future of Life Institute, 2017. *Asilomar AI Principles*. Available: https://futureoflife.org/ai-principles/?cn-reloaded=1 Accessed: 4 November 2021

Gambin, O. 2020. Brave: what it means to be an AI Ethicist. *AI and Ethics* 1: 87–91

Garba, S., Ahmed, A., Mai, A., Makama, G. and Odigie, V., 2010. Proliferations of Scientific Medical Journals: A Burden or A Blessing. *Oman Medical Journal*. 25(4): p.311–314

General Medical Council, 2013. *The Good Medical Practice Framework for appraisal and revalidation*. Available: https://www.gmc-uk.org/-/media/documents/The_Good_medical_practice_framework_for_appraisal_and_revalidation___DC5707.pdf_56235089.pdf Accessed: 4 November 2021

General Medical Council, 2020. *Good medical practice: General Medical Council*. Available: https://www.gmc-uk.org/ethical-guidance/ethical-guidance-for-doctors/good-medical-practice Accessed: 4 November 2021

General Medical Council, 2021a. *Investigating and acting on concerns about doctors.* Available: https://www.gmc-uk.org/about/what-we-do-and-why/investigating-and-acting-on-concerns-about-doctors Accessed: 4 November 2021

General Medical Council, 2021b. *Insurance indemnity and medico-legal support*. Available: https://www.gmc-uk.org/registration-and-licensing/managing-your-registration/information-for-doctors-on-the-register/insurance-indemnity-and-medico-legal-support Accessed: 4 November 2021

Gert, J. 2020. The Definition of Morality. *Stanford Encyclopedia of Philosophy*. Available: https://plato.stanford.edu/entries/morality-definition/ Accessed: 4 November 2021

Giliker, P., 2017, *Tort*. London: Sweet and Maxwell

Godfrey-Smith, P., 2003. *Theory and Reality: an introduction to the philosophy of science*. Chicago and London: The University of Chicago Press

Gosepath, S., 2021. *Equality*. Stanford Encyclopedia of Philosophy. Available: https://plato.stanford.edu/entries/equality/ Accessed: 4 November 2021

Gould, M., 2020. Regulating AI in health and care. 11 February. Available: https://digital.nhs.uk/blog/transformation-blog/2020/regulating-ai-in-health-and-care Accessed: 4 November 2021

Government Digital Service, 2019. *Understanding artificial intelligence ethics and safety*. 10 June 2019. Available: https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety Accessed: 4 November 2021

Government Digital Service, 2019. *Assessing if artificial intelligence is the right solution*. 10 June 2019. https://www.gov.uk/guidance/assessing-if-artificial-intelligence-is-the-right-solution Accessed: 4 November 2021

Government Digital Service, 2020. *Data Ethics Framework*. Updated 16 September 2020. Available: https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework-2020 Accessed: 4 November 2021

Government Equalities Office and Equality and Human Rights Commission, 2015. *Equality Act 2010: guidance.* Available: https://www.gov.uk/guidance/equality-act-2010-guidance Accessed: 4 November 2021

Government Office for Science, 2016. *Artificial Intelligence: Opportunities and Implications for the Future of Decision Making*. Available:

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/566075/gs-16-19-artificial-intelligence-ai-report.pdf Accessed: 4 November 2021

Green, S., 2017. Fairchild and the single agent criterion. *Law Quarterly Review*. 133: p.25–31

Greenawalt, K. 1998. Too thin and too rich: Distinguishing Features of Legal Positivism. In: George, R.P., (ed) 1999. *The Autonomy of Law: Essays on Legal Positivism*. Oxford: Oxford University Press

Greenhalgh, T., 2019. Tweet of 7 November: "For this reason, candidates should be taught to make the 'red thread' (central argument) of the thesis crystal clear as well as periodically summarising where the red thread's got to, and making extensive use of cross-referencing." *Twitter*. Available: https://twitter.com/trishgreenhalgh/status/1192532247355809792 Accessed: 4 November 2021

Groopman, J., 2007. Interviews: "How Doctors Think." *NPR Books*. Available: http://www.npr.org/templates/story/story.php?storyId=8892053 Accessed: 4 November 2021

Griffin, A., 2017. Saudi Arabia grants citizenship to a robot for the first time ever. *Independent*. 26 October. Available: https://www.independent.co.uk/life-style/gadgets-and-tech/news/saudi-arabia-robot-sophia-citizenship-android-riyadh-citizen-passport-future-a8021601.html Accessed: 4 November 2021

Gupta, H., 2020. Rapid response to: The doctors navigating covid-19 with no internet. *BMJ* 369 Available: https://www.bmj.com/content/369/bmj.m1417/rr-0 Accessed 4 November 2021

Gutenstein, M., 2014. Psychological factors in emergency medicine. *Emergency Medicine Australasia*. 26 p.295-299

Habermas, J. 1990. Moral consciousness and Communicative Action, trans. Lenhardt, C. and Nicholsen S.W. Cambridge, MA: MIT Press

Hambury, D., 2018. Health secretary Matt Hancock: 'AI can augment the human factor' of medicine. *Evening Standard.* 27 November. Available: https://www.standard.co.uk/futurelondon/health/matt-hancock-on-ai-and-the-nhs-a3998006.html Accessed: 4 November 2021

Hao, K., 2019. In 2020, let's stop AI ethics-washing and actually do something. *MIT Technology Review.* 27 December. Available: https://www.technologyreview.com/2019/12/27/57/ai-ethics-washing-time-to-act/ Accessed: 4 November 2021

Hartman, D. E., 1986. On the Use of Clinical Psychology Software. Practical, Legal, and Ethical Concerns. *Professional Psychology: Research and Practice*. 17(5):462-465

Harwich E, Laycock K., 2018. Thinking on its own: AI in the NHS. *Reform*. Available: https://reform.uk/sites/default/files/2018-11/AI%20in%20Healthcare%20report_WEB.pdf Accessed: 4 November 2021

Health and Care Professions Council, 2016. *Standards of Conduct, Performance and Ethics*. Available at: https://www.hcpc-uk.org/standards/standards-of-conduct-performance-and-ethics/ Accessed: 4 November 2021

Health & Care Professions Council, 2018. *Fitness to practise - Raising concerns.* Available: https://www.hcpc-uk.org/concerns/raising-concerns/ Accessed: 4 November 2021

Health and Care Professions Council, 2021. *Who we regulate.* Available: https://www.hcpc-uk.org/about-us/who-we-regulate/ Accessed: 4 November 2021

Health and Care Professions Council, 2018. *Professional indemnity*. Available: https://www.hcpc-uk.org/registration/your-registration/legal-guidelines/professional-indemnity/ Accessed: 4 November 2021

Health and Social Act 2008. c.14

Health Research Authority, 2021. *Roles and responsibilities*. Available: https://www.hra.nhs.uk/planning-and-improving-research/research-planning/roles-and-responsibilities/ Accessed: 4 November 2021

Hengstler M., Enkel E., Duelli S., 2016. Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices. *Technological Forecasting & Social Change* 105: p.105–120

Heywood, R. 2021. *Reliance on clinical guidelines in contemporary negligence litigation in the UK: influential or determinative?* In: Samanta, J., Samanta, A., 2021. *Clinical guidelines and the law of medical negligence : multidisciplinary and international perspectives*. Cheltenham, UK: Edward Elgar Publishing Limited

High-Level Expert Group on Artificial Intelligence, 2019. *Ethics Guidelines for Trustworthy AI.* Brussels: European Commission. Available: https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1 Accessed: 4 November 2021

Hobbes, T., 2018. *Leviathan*. Minneapolis: First Avenue Editions

Hodgson, D., 2008. *The Law of Intervening Causation.* Aldershot: Ashgate

Hoffmaster, B., 2018. From applied ethics to empirical ethics to contextual ethics. *Bioethics* 32:119-125

Hogan, H., Healey, F., Neale, G., Thomson, R., Vincent, C., Black, N., 2012. Preventable deaths due to problems in care in English acute hospitals: a retrospective case record review study. *BMJ Quality and Safety*. 21:737-745

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H., 2019. Causability and explainability of artificial intelligence in medicine. *Data Mining and Knowledge Discovery*. 9(4): e1312

Holm, L., 2017. Analyzing the Theme of Equality in Thomas Hobbes' "Leviathan". Owlcation. Available: https://owlcation.com/humanities/Analyzing-the-Theme-of-Equality-in-Thomas-Hobbes-Leviathan Accessed: 4 November 2021

Holm, S., Stanton, C., Bartlett, B., 2021. A New Argument for No-Fault Compensation in Health Care: The Introduction of Artificial Intelligence Systems. *Health Care Analysis*. 29:171–188

Holmes, J. G., Rempel, J. K., 1989. Trust in close relationships. In: Hendrick, C. (ed), *Close Relationships: Review of personality and social psychology.* Volume 10, p.187-220. Newbury Park, California: Sage Publications, Inc.

Holmes, W., 1918. Natural Law. *Harvard Law Review.* 32(1):40-44

Horsfall, S., 2014. Doctors who commit suicide while under GMC fitness to practise investigation. *General Medical Council.* Available: http://www.gmc-uk.org/Internal_review_into_suicide_in_FTP_processes.pdf_59088696.pdf Accessed: 4 November 2021

House of Commons: Science and Technology Committee, 2016. *Robotics and artificial intelligence Fifth Report of Session 2016–17.* London: House of Commons. Available: https://publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf Accessed: 4 November 2021

House of Commons: Science and Technology Committee, 2018. *Algorithms in decision-making*. London: House of Commons. Available: https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/351/351.pdf Accessed: 4 November 2021

House of Lords: Select Committee on Artificial Intelligence, 2018. *AI in the UK: ready, willing and able?* London: House of Lords. Available:

https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf Accessed: 4 November 2021

Hume, D., 2014. Nidditch, P.H. (ed). Enquiries Concerning Human Understanding and Concerning the Principles of Morals Oxford: Oxford University Press

Huxtable, R., 2020. COVID-19: where is the national ethical guidance? *BMC Medical Ethics*. 21:32 Available: https://doi.org/10.1186/s12910-020-00478-2 Accessed: 4 November 2021

Huxtable, R. & Ost, S., 2017. Voices carry? The voice of bioethics in the courtroom and the voice of law in bioethics. In: Huxtable, R., Ter Meulen, R. (eds), *The Voices and Rooms of European Bioethics*. London: Routledge

Huxtable, R., 2016*.* Friends, Foes, Flatmates: On the Relationship between Law and (Empirical) Bioethics. In: Ive, J., Dunn, M., Cribb, A., 2016. *Empirical Bioethics Theoretical and Practical Perspectives.* Cambridge: Cambridge: University Press

Huxtable, R., 2012. *Law, Ethics and Compromise at the Limits of Life: To Treat or Not to Treat?* London: Routledge.

Iannello, P., Perucca, V., Riva, Silva., Alessandro, A., Pravettoni, G., 2015. What Do Physicians Believe About the Way Decisions Are Made? A Pilot Study on Metacognitive Knowledge in the Medical Context. *Europe's Journal of Psychology*. 11:(4) p.691-706

IBM, (undated). *IBM Watson for Oncology: What Watson for Oncology can do for your organization.* Available: https://www.ibm.com/products/clinical-decision-supportoncology Accessed: 8 March 2020. *Webpage had been removed when this reference list's links were checked in November 2021.*

IBM, 2020. *Watson Health: Go beyond human capabilities*. Available: https://www.ibm.com/watson-health/oncology-and-genomics?loc=uk-en Accessed: 21 October 2020 *Webpage had been removed when this reference list's links were checked in November 2021.*

IEEE, 2017. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.* Available: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf?utm_medium=undefined&utm_source=undefined&utm_campaign=undefined&utm_content=undefined&utm_term=undefined Accessed: 4 November 2021

IEEE, 2021. *IEEE 7000-2021 Standard Model Process for Addressing Ethical Concerns during System Design.* Available: https://standards.ieee.org/standard/7000-2021.html Accessed: 4 November 2021.

Information Commissioner’s Office and the Alan Turing Institute, 2020. *Explaining decisions made with AI.* 20 May 2020. Available: https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/ Accessed: 4 November 2021

Inter-American Commission on Human Rights (IACHR), 1948. *American Declaration of the Rights and Duties of Man.* Available: http://www.oas.org/en/iachr/mandate/Basics/declaration.asp Accessed: 4 November 2021

Inthorn, J., Tabacchi, M.E., Seising, R., 2015. Having the final say: Machine support of ethical decisions of doctors. In: van Rysewyk, S.P., Pontier, M. (eds) *Machine Medical Ethics: Intelligent Systems, Control and Automation: Science and Engineering*. Switzerland: Springer p.181-206

Ives, J., 2014. A method of reflexive balancing in a pragmatic, interdisciplinary and reflexive bioethics. *Bioethics*. 28(6): p.302–312

Izzard, E., 1998. *Dressed to Kill.* Vision Video

Jobin, A., Ienca, M., Vayena, E., 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*. 1: p.389–399

Jones, M., 2006. Proving causation – Beyond the “but for” test. *Professional Negligence.* 22(4):p.251-269

Jones, M., 2021. *Medical Negligence*. 6th edition. London: Sweet & Maxwell

Kalluri, P., 2020. Don’t ask if artificial intelligence is good or fair, ask how it shifts power. Nature. 7 July 2020. Available: https://www.nature.com/articles/d41586-020-02003-2 Accessed: 4 November 2021

Kant, I., 1998. *Groundwork of the Metaphysics of Morals.* Translated and edited by Gregor, M. Cambridge: Cambridge University Press

Kant, I., 1793. *Theory and Practice.* Volume VIII: 297 (Reiss, 79).

Kant, I., 2004. *Prolegomena to Any Future Metaphysics, translated and edited by Gary Hatfield, revised edition*. Cambridge: Cambridge University Press

Kant, I., 2011. *Groundwork of the Metaphysics of Morals: A German-English edition.* Cambridge: Cambridge University Press

Kaul, V., Enslin, S., Gross, S.A., 2020. History of artificial intelligence in medicine. *Gastrointestinal Endoscopy*. 92(4): p.807-812

Keating, G. C. 1995. Reasonableness and rationality in negligence theory. *Stanford Law Review.* 48: p.311-384

Keeton, W.P., Dobbs, D.B., Keeton, R.E., Owen, D.G., 1984. *Prosser and Keeton on Torts.* USA: West Publishing Company

Keikes, L., Medlock, S., van de Berg, D., Zhang, S., Guicherit, O.R., Punt, C.J.A., van Oijen, M.G.H., 2017. The first steps in the evaluation of a "black-box" decision support tool: a protocol and feasibility study for the evaluation of Watson for Oncology. *Journal of Clinical and Translational Research*. 3(S3): p.411-423

Kellmeyer, P., Cochrane, T., Müller, O., Mitchell, C., Ball, T., Biller-Andorno, J. J., Fins, N., 2016. The Effects of Closed-Loop Medical Devices on the Autonomy and Accountability of Persons and Systems. *Cambridge Quarterly of Healthcare Ethics*. 25(4): p.623-631

Khan O.F., Bebb, G., Alimohamed, N., 2017. Artificial intelligence in medicine What oncologists need to know about its potential —and its limitations. *Oncology Exchange*. 16(4)8-13

Khan, K.S., Kunz, R., Kleijnen J, Antes, G., 2003. Five steps to conducting a systematic review. *Journal of the Royal Society of Medicine* 96: p.118–121

Kim, L.M., Looser, P., Swaminathan, R.V., Horowitz, J., Friedman, O., Shin, J.H., Minutello, R.M., Bergman, G., Sing, H., Wong, C., Feldman, D.N., 2016. Sex-Based Disparities in Incidence, Treatment, and Outcomes of Cardiac Arrest in the United States, 2003–2012. *Journal of the American Heart Association.* 5(6) e003704. Available: https://www.ahajournals.org/doi/10.1161/JAHA.116.003704 Accessed: 4 November 2021

Klemme, H. C., 1975-6. The Enterprise Liability Theory of Torts. *University of Colorado Law Review*, 47, p.153-232

Kyte, D.G., Draper, H., Ives, J., Liles, C., Gheorghe, A., Calvert, M., Timmer, A., 2013. Patient reported outcomes (PROs) in clinical trials: is 'in-trial' guidance lacking? A systematic review. *PLoS ONE* 8(4):e60684

Lagnado, D. A. & Gerstenberg, T., 2017. In: Waldmann, M.R. (ed), 2017. *The Oxford Handbook of Causal Reasoning.* Oxford: Oxford University Press

Laing, J., McHale, J. (eds), 2017. *Principles of Medical Law.* Oxford: Oxford University Press

Latifi, R., 2016. *Surgical Decision Making: Beyond the Evidence Based Surgery*. Switzerland: Springer

Law Reform (Contributory Negligence) Act 1945. c.28

Lenzer, J., 2016. Rita Redberg: an unwavering campaigner against the harms of too much medicine. *The British Medical Journal*. 354:i4390 Available: https://www.bmj.com/content/354/bmj.i4390 Accessed: 4 November 2021

Lewis, D., 1973. Causation. *The Journal of Philosophy.* 70(1): p.556-567

Lighthall, G.K., and Vazquez-Guillamet, C, 2015. Understanding Decision Making in Critical Care. *Clinical Medicine and Research*. 13(3-4): p.156-168

Lloyd, S.A., 2018. *Hobbes's Moral and Political Philosophy.* Stanford Encyclopedia of Philosophy. Available: https://plato.stanford.edu/entries/hobbes-moral/ Accessed: 4 November 2021

Luxton, D.D., 2014. Recommendations for the ethical use and design of artificial intelligent care providers. *Artificial Intelligence in Medicine*. 62(1): p.1-10

Mackenzie, J., 2017. *The Medical Duty of Care to a Third Party*. Available: https://www.anthonygold.co.uk/latest/blog/medical-duty-care-third-party/ Accessed: 4 November 2021

Makary, M.A., Daniel, M., 2016. Medical error—the third leading cause of death in the US. *The British Medical Journal*. Available: https://doi.org/10.1136/bmj.i2139 Accessed: 4 November 2021

Marr, B., 2016. What Is The Difference Between Artificial Intelligence And Machine Learning? *Forbes.* 6 December. Available: https://www.forbes.com/sites/bernardmarr/2016/12/06/what-is-the-difference-between-artificial-intelligence-and-machine-learning/#5cd79e7a2742 Accessed: 4 November 2021

Maslow, A.H., 1966. *The Psychology of Science.* New York: Harper & Row p. 15.

Masnick, M., 2016. Activists Cheer On EU's 'Right To An Explanation' For Algorithmic Decisions, But How Will It Work When There's Nothing To Explain? *Techdirt*. 8 July. Available: https://www.techdirt.com/articles/20160708/11040034922/activists-cheer-eus-right-to-explanation-algorithmic-decisions-how-will-it-work-when-theres-nothing-to-explain. Accessed: 4 November 2021

Matthias, A., 2004. The responsibility gap: Ascribing responsibility for the actions of learning Automata. *Ethics and Information Technology* 6: 175–183

McHale, J., Laing, J., (eds), 2010. *Principles of Medical Law.* Oxford: Oxford University Press

McLoughlin v O'Brian [1983] 1 AC 410, 425

Medical Devices Regulations 2002 No.618.

Medicines and Medical Devices Act 2021 c.3

Medicines and Healthcare Products Regulatory Agency, 2021a. Good Machine Learning Practice for Medical Device Development: Guiding Principles. 27 October 2021. Available: https://www.gov.uk/government/publications/good-machine-learning-practice-for-medical-device-development-guiding-principles Accessed 4 November 2021.

Medicines and Healthcare Products Regulatory Agency, 2021b. *Managing Medical Devices: Guidance for health and social care organisations.* January 2021. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/965010/Managing_medical_devices022021.pdf Accessed: 4 November 2021

Medicines and Healthcare Products Regulatory Agency, 2019. *Medical devices: the regulations and how we enforce them.* 26 February. Available: https://www.gov.uk/government/publications/report-a-non-compliant-medical-device-enforcement-process/how-mhra-ensures-the-safety-and-quality-of-medical-devices Accessed: 4 November 2021

Menzies, P., Beebee, H., 2019. *Counterfactual Theories of Causation.* Stanford Encyclopedia of Philosophy. Available: https://plato.stanford.edu/entries/causation-counterfactual/ Accessed: 4 November 2021

Merkin, R., Steele, J., 2013. *Insurance and the law of obligations.* Oxford: Oxford University Press.

Metz, C., 2016a. What the AI behind AlphaGo can teach us about being human. 19 May 2016. *Wired*. Available: https://www.wired.com/2016/05/google-alpha-go-ai/ Accessed 4 November 2021.

Metz, C., 2016b. In Two Moves, AlphaGo and Lee Sedol Redefined the Future. 16 March 2016 *Wired*. Available: https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/. Accessed: 4 November 2021.

Microsoft, 2021. *Project InnerEye – Democratizing Medical Imaging AI*. Available: https://www.microsoft.com/en-us/research/project/medical-image-analysis/ Accessed: 4 November 2021

Miles, A., 2007. 'Science: A limited source of knowledge and authority in the care of patients. A Review and Analysis of: 'How Doctors Think. Clinical Judgement and the Practice of Medicine.' Montgomery, K', *Journal of Evaluation in Clinical Practice*, 13(4): p.545-563

Miller, D. 2021. Contractarianism and Justice. *Stanford Encyclopedia of Philosophy*. Available: https://plato.stanford.edu/entries/justice/#ContJust Accessed: 4 November 2021

Mittelstadt, B., 2019. Principles alone cannot guarantee ethical AI. Nature Machine Intelligence. 1:p.501–507. Accessed 10 May 2022. Available: https://www.nature.com/articles/s42256-019-0114-4

Monett, D., Lewis, C.W.P., Thorisson, K.R., 2020. Introduction to the JAGI Special Issue "On Defining Artificial Intelligence"—Commentaries and Author's Response. *Journal of Artificial General Intelligence* 11(2) 1-4

Morris, Z.S.M., Wooding, S., Grant, G., 2011. The answer is 17 years, what is the question: understanding time lags in translational research. *Journal of the Royal Society of Medicine.* 104(12): p.510-520

Mukherjee, S., 2017 A.I. versus M.D.: What happens when diagnosis is automated? *The New Yorker.* 3 April. Available: http://www.newyorker.com/magazine/2017/04/03/ai-versus-md Accessed: 4 November 2021

Mumford, S., Anjum, R., 2013. With Great Power Comes Great Responsibility. In: Kahmen, B., Stepanians, M., (eds) 2013. *Causation and Responsibility: Critical Essays*. Berlin: de Gruyter, p.219-37

National Institute for Health and Care Excellence, 2020. *Artificial intelligence for analysing CT brain scans: Medtech innovation briefing [MIB207]*. Available: https://www.nice.org.uk/advice/mib207 Accessed: 4 November 2021

National Institute for Health and Care Excellence, 2021. *Evidence Standards Framework for Digital Health Technologies*. Available: https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies Accessed: 4 November 2021

National Institute of Health and Care Excellence, 2020. *NICE guidance.* Available: www.nice.org.uk/guidance Accessed: 4 November 2021

National Institute of Health and Care Excellence, 2021. *What we do*. Available: https://www.nice.org.uk/about/what-we-do Accessed: 4 November 2021

NHSX, undated. *The NHS AI Lab: Accelerating the safe adoption of Artificial Intelligence in health and care*. Available: https://www.nhsx.nhs.uk/ai-lab/ Accessed: 4 November 2021

NHS England, undated a. *C the Signs – How artificial intelligence (AI) is supporting referrals*. Available: https://www.england.nhs.uk/cancer/case-studies/c-the-signs-how-artificial-intelligence-ai-is-supporting-referrals/ Accessed: 4 November 2021

NHS England, undated b. *NHS Digital Academy*. Available: https://www.england.nhs.uk/digitaltechnology/nhs-digital-academy/ Accessed: 4 November 2021

NHS Resolution, 2020a. *What we do.* Available: https://resolution.nhs.uk/about/our-work/ Accessed: 4 November 2021

NHS Resolution, 2020b. Clinical schemes. Available: https://resolution.nhs.uk/services/claims-management/clinical-schemes/ Accessed: 4 November 2021

NHS Resolution, 2021a. NHS Resolution Annual report and accounts 2020/21. Accessed 13 May 2022. Available: https://resolution.nhs.uk/wp-content/uploads/2021/07/NHS_Resolution_Annual-Report-2021.pdf

NHS Resolution, 2021b. Claims Management. Available: https://resolution.nhs.uk/services/claims-management/ Accessed: 4 November 2021

NHS Resolution, 2021a. *About NHS Resolution.* Available: https://resolution.nhs.uk/about/ Accessed: 4 November 2021

NHSX, undated. What we do. *About Us*. Available at: https://www.nhsx.nhs.uk/about-us/what-we-do/ Accessed: 4 November 2021

NHSX, 2019. Artificial Intelligence: How to get it right. Available: https://www.nhsx.nhs.uk/media/documents/NHSX_AI_report.pdf Accessed: 4 November 2021

NHSX, 2020. *A Buyer's Guide to AI in Health and Care*. 8 September. Available: https://www.nhsx.nhs.uk/ai-lab/explore-all-resources/adopt-ai/a-buyers-guide-to-ai-in-health-and-care/ Accessed: 4 November 2021

NHSX AI Lab, 2020. *The NHS AI Lab Virtual Event, 24 September*: *Multiagency Advice Service for AI Technologies in Health and Social Care*. Available: https://www.youtube.com/watch?v=rAWRAXFPd6E&feature=youtu.be&t=7467 Accessed: 4 November 2021

Nissenbaum, H. (1996) 'Accountability in a computerized society' *Science and Engineering Ethics*, 2: 25-42.

Noee, E, Noee, M., Mehrpouyan, A., 2016. Attribution of Liability among Multiple Tortfeasors under Negligence Law: Causation in Iran and England. *Journal of Politics and Law* 9(7):219–229

Nozick, R. 1974. *Anarchy, State, and Utopia.* Oxford: Blackwell Publishers Ltd

Nursing and Midwifery Council, 2018. *The Code for Nurses and Midwives*. Available: https://www.nmc.org.uk/standards/code/read-the-code-online/ Accessed: 4 November 2021

Nursing and Midwifery Council. 2018a. *Our investigations.* Available: https://www.nmc.org.uk/ftp-library/understanding-fitness-to-practise/investigations/ Accessed: 4 November 2021

Nursing and Midwifery Council, 2021. *Revalidation / What You Need To Do: Written reflective accounts*. Available: http://revalidation.nmc.org.uk/what-you-need-to-do/written-reflective-accounts.html Accessed: 4 November 2021

Nursing and Midwifery Council, 2020. *Professional indemnity arrangement*. Available: https://www.nmc.org.uk/registration/staying-on-the-register/professional-indemnity-arrangement/ Accessed: 4 November 2021

Ockenden, D., 2022. *Final report of the Ockenden review: Findings, conclusions and essential actions from the independent review of maternity services at the Shrewsbury and Telford Hospital NHS Trust.* Department of Health and Social Care. Accessed 12 May 2022. Available: https://www.gov.uk/government/publications/final-report-of-the-ockenden-review

O'Dowd, A., 2015. Doctors increasingly practise "defensive" medicine for fear of litigation, says regulator. *The British Medical Journal*. 350:h87 Available: http://www.bmj.com/content/350/bmj.h87 Accessed: 4 November 2021

OECD, 2018. OECD AI Principles. Accessed 10 May 2022. Available: https://oecd.ai/en/ai-principles

Ortashi, O., Virdee, J., Hassan, R., Mutrynowski, T., Abu-Zidan, F., 2013. The practice of defensive medicine among hospital doctors in the United Kingdom. *BMC Medical Ethics.* 14:42

Oshana, M., 2004. Moral accountability. *Philos Top* 32(1–2):255–274

Open Access Government, 2019. NHS uses AI technology to fight coronary heart disease. *Open Access News*. Available: https://www.openaccessgovernment.org/ai-technology-nhs-coronary-heart-disease/61166/ Accessed: 4 November 2021

Panel for the Future of Science and Technology, 2020. *The ethics of artificial intelligence: Issues and initiatives*. Brussels: European Parliament. PE 634.452 Available:

https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf Accessed: 4 November 2021

Parliament of the United Kingdom, 2020. *Medicines and Medical Devices Bill 2019-21.* Available: https://services.parliament.uk/bills/2019-21/medicinesandmedicaldevices.html Accessed: 4 November 2021

Pearce, M.S., Salotti, J.A., Little, M.P., McHugh, K., Lee, C., Kim, K.P., Howe, N.L., Ronckers, C.M., Rajaraman, P., Craft, A.W., Parker, L., Berrington de Gonzalez, A., 2012. Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study. *Lancet*. 380(9840): p.499–505

Petricek, T., 2017. Miscomputation in Software: Learning to Live with Errors. *The Art, Science and Engineering of Programming* 1(2)14: p.1-24

Pojman, L., 1995. Theories of Equality: A Critical Analysis. *Behavior and Philosophy*. 23(2): p.1-27

Poorman, E., 2016. Staying Current in Medicine: Advice for New Doctors. *NEJM Knowledge+.* Available: https://knowledgeplus.nejm.org/blog/staying-current-in-medicine-advice-for-new-doctors/ Accessed: 4 November 2021

Pouloudi, A., Magoulas, G.D., 2000. Neural expert systems in medical image interpretation: Development, use, and ethical issues. *Journal of Intelligent Systems.* 10(5-6): 451-472

Professional Standards Authority, undated. *About regulators.* Available: https://www.professionalstandards.org.uk/what-we-do/our-work-with-regulators/about-regulators Accessed: 4 November 2021

Rawls, J., 1999. A Theory of Justice: Revised Edition. Cambridge, USA: Belknap Press of Harvard University Press

Rawls, J., 2001. *Justice as Fairness; a restatement.* London: Belknap Press of Harvard University Press

Rawls, J., 2007. *Lectures on the History of Political Philosophy*. Cambridge, USA: Belknap Press of Harvard University Press

Reddy, S., Allan, S., Coghlan, S., Cooper, P., 2019. A governance model for the application of AI in health care. *JAMIA*, 27(3): 491–497

Reschovsky, J.D. and Saiontz-Martinez, C.B., 2017. Malpractice Claim Fears and the Costs of Treating Medicare Patients: A New Approach to Estimating the Costs of Defensive Medicine. *Health Services Research.* 53(3): p.1498-1516

Resuscitation Council UK, 2021. *2021 Resuscitation Guidelines.* Available: https://www.resus.org.uk/library/2021-resuscitation-guidelines Accessed: 4 November 2021

Reyna, V. F. & Rivers, S. E., 2008. Current theories of risk and rational decision making. *Developmental Review*. 28: p.1-11

Richardson WS, Glasziou P, Polashenski WA, Wilson, M.C., 2000. A new arrival: evidence about differential diagnosis. *BMJ Evidence-Based Medicine* 5:164-165

Road Traffic Act 1988 c.52

Roberts, I., Ker, K., Edwards, P., Beecher, D., Manno, D., Sydenham, E., 2015. The knowledge system underpinning healthcare is not fit for purpose and must change. *The British Medical Journal*. 350 h2463

Robinson, F., 2019. Learn Not Blame: how a grassroots campaign struck a chord. BMJ. 365:l4232. Available: https://www.bmj.com/content/365/bmj.l4232 Accessed: 4 November 2021

Rogers, E. M., 2003. Diffusion of innovations (5th ed.). New York, NY: Free Press Rogers, E. M. (2003). Diffusion of innovations (5th ed.). New York, NY: Free Press

Rosa, M., 2020. On Defining Artificial Intelligence—Commentary. *Journal of Artificial General Intelligence* 11(2) 60-62

Ross C., Swetlitz I., 2017. IBM pitched Watson as a revolution in cancer care. It is nowhere close. *STAT News*. https://www.statnews.com/2017/09/05/watson-ibm-cancer/ Accessed: 4 November 2021

Rousseau, J.J., 2002. *The Social Contract and The First and Second Discourses.* Dunn, S. (ed). New Haven: Yale University Press

Rowell, D., Connelly, L.B., 2012 "A History of the Term Moral Hazard" *The Journal of Risk Insurance* 19 (4): 1051-1075

Rowland, D., Rowland, J.J., 1993. Competence and Legal Liability in the Development of Software for Safety-Related Applications. *Information & Communications Technology Law* 2: p.229-243

Royal Statistical Society, 2020. *Professional standards to be set for data science*. 23 July. Available: https://rss.org.uk/news-publication/news-publications/2020/general-news/professional-standards-to-be-set-for-data-science/ Accessed: 4 November 2021

Russell, S. Norvig, P., 2016. *Artificial Intelligence: A Modern Approach (Global Edition*). 3rd ed. London: Pearson

Samanta, A., Mello, M.M., Foster, C., Tingle, J., Samanta, J., 2006. The Role of Clinical Guidelines in Medical Negligence Litigation: A Shift from the *Bolam* standard? Medical Law Review, 14:p.321-366

Sarre-Lazcano, C., Alonso, A.A., Melendez, F.D.H., Arrieta, O., Norden, A.D., Urman, A., Perroni, M., Landis-Mcgrath, A., Medina-Franco, H., 2017. Cognitive computing in oncology: A qualitative assessment of IBM Watson for Oncology in Mexico. *Journal of Clinical Oncology*. 35:15_suppl, e18166-e18166

Schäfer, G., Prkachin, K.M., Kaseweter, K.A., Williams, A.C. de C., 2016. Health care providers' judgments in chronic pain: the influence of gender and trustworthiness. *Pain*. 157(8): p.1618–1625

Schedlbauer, A., Prasad, V., Mulvaney, C., Phansalkar, S., Stanton, W., Bates, D., Avery, A.J., 2009. What Evidence Supports the Use of Computerized Alerts and Prompts to Improve Clinicians' Prescribing Behavior? *Journal of the American Informatics Association*. 16(4): p.531-8

Schwarcz D., Siegelman, P., 2015. *Research Handbook on the Economics of Insurance Law* Cheltenham, UK: Edward Elgar

Sidgwick, H., 1907. *The Methods of Ethics*. 7th ed. London: Hackett Publishing Co, Inc.

Singer, P., 2011. *Practical Ethics.* 3rd edition. Cambridge: Cambridge University Press

Smart, J.J.C., Williams, B., 1973. *Utilitarianism: For and against.* Cambridge: Cambridge University Press.

Smith, H. & Fotheringham, K., 2022. Exploring Remedies for Defective Artificial Intelligence Aids in Clinical Decision Making in post-Brexit England and Wales. *Medical Law International.* 22(1):p. 33-51 27 DOI: 10.1177/09685332221076124

Somashekhar, S.P., Sepúlveda, M.J., Norden, A.D., Rauthan, A., Arun, K., Patil, P., Ethadka, R.Y., Kumar, R.C., 2017. Early experience with IBM Watson for Oncology (WFO) cognitive computing system for lung and colorectal cancer treatment. *Journal of Clinical Oncology.* 2017 35:15_suppl, p.8527-8527

Song, F., 2019. Regarding a Risk-Pooling System of compensation. *Ratio.* 32: 139–149

Staats, C., 2014. State of the Science: Implicit Bias Review 2014. *Kirwan Institute*. Available: http://kirwaninstitute.osu.edu/my-product/2014-state-of-science-implicit-bias-review/ Accessed: 4 November 2021

Stapleton, J. 2008. Choosing what we mean by “causation” in the law. *Missouri Law Review* 73(2): p.433–480

Stapleton, J., 2009. Causation in the law. In: Beebee, H., Hitchcock, C., Menzies, P. (2009) *The Oxford handbook of causation.* Oxford: Oxford University Press

Stiegler, M.P.S., Tung, A. 2014. Cognitive Processes in Anesthesiology Decision Making. *Anesthesiology*. 120(1): p.204-217

Strech, D., Sofaer, N., 2012. How to write a systematic review of reasons. *Journal of Medical Ethics* 38: p.121–126

Sugden, R., 1990. Contractarianism and Norms. *Ethics*. 100(4): p.768-786

Sukel, K., 2017a. With a little help from AI friends. *New Scientist*. 235(3134): p.36-39

Sukel, K., 2017b. Artificial Intelligence ushers in the era of superhuman doctors. *New Scientist*. Available: https://www.newscientist.com/article/mg23531340-800-artificial-intelligence-ushers-in-the-era-of-superhuman-doctors/ Accessed: 4 November 2021

Sullivan, M. & Reynolds, D., 1998. Where Law and Bioethics Meet ... and Where They Don’t!! *University of Detroit Mercy Law Review.* 75: p.607–620

Swierstra, T., & Rip, A., 2007. Nano-ethics as NEST-ethics: Patterns of Moral Argumentation About New and Emerging Science and Technology. *Nanoethics* 1:3-20

Taddeo, M., Floridi, L., 2018. How AI can be a force for good. *Science*. 361 6404: 751-752

Takala, T., 2015. Get to the Point! Philosophical Bioethics and the Struggle to Remain Relevant. *Cambridge Quarterly of Healthcare Ethics*. 24:149-153

Taylor, C., 1979. Atomism. In: Hill, c., Mansfield, H.C., Taylor, C., Pateman, C., Shklar, J., Hobsbawm., Avineri, S., Leiss, W., Lukes, S., (eds) 1979. *Powers, Possessions and Freedom: Essays in Honour of C.B. Macpherson.* Toronto: University of Toronto Press

Temkin, L., 1986. *Inequality.* Philosophy and Public Affairs, 15: 99–121

Theodorou, A., Dignum, V., 2020. Towards ethical and socio-legal governance in AI. *Nature Machine Intelligence.* 2:p.10–12 Accessed 10 May 2022. Available: https://www.nature.com/articles/s42256-019-0136-y

The General Product Safety Regulations 2005. No. 1803

The Health Protection (Coronavirus, Wearing of Face Coverings in a Relevant Place) (England) Regulations 2020. No. 791

The Medical Devices (Amendment etc.) (EU Exit) Regulations 2020. UK Statutory Instruments, 2020, No. 1478

The Medical Devices Regulations 2002. No. 618

Topol, E. 2019. *The Topol Review: Preparing the healthcare workforce to deliver the digital future; An independent report on behalf of the Secretary of State for Health and Social Care*. February 2019. Available: https://topol.hee.nhs.uk/ Accessed: 4 November 2021

Tupasela, A., Di Nucci, E., 2020. Concordance as evidence in the Watson for Oncology decision-support system. *AI & Society.* 35:811–818

Turner, J., 2019. *Robot Rules: Regulating Artificial Intelligence.* Switzerland: Palgrave Macmillan

Ucapher, H., Areán, P.A., 2000. Physicians Are Less Willing to Treat Suicidal Ideation in Older Patients. *Journal of the American Geriatrics Society*. 48(2): p.188–192

Unfair Contract Terms Act 1977 c.50

UK Statistics Authority, 2021. *Ethical considerations in the use of Machine Learning for research and statistics*. 26 October 2021. Available: https://uksa.statisticsauthority.gov.uk/publication/ethical-considerations-in-the-use-of-machine-learning-for-research-and-statistics/pages/7/#pid-accountability Accessed: 4 November 2021

United Nations, 1948. *Universal Declaration of Human Rights.* Available: https://www.un.org/en/about-us/universal-declaration-of-human-rights Accessed: 4 November 2021

UNESCO, 2021. Draft Recommendation on the Ethics of Artificial Intelligence. General Conference, 41st, [764] 41 C/23 Available: https://unesdoc.unesco.org/ark:/48223/pf0000378931 Accessed: 19 November 2021.

Uzonwanne, F., 2016. *Global Encyclopedia of Public Administration, Public Policy, and Governance.* London: Springer

Vallverdú, J., Casacuberta, D., 2015. Ethical and technical aspects of emotions to create empathy in medical machines. In: van Rysewyk, S.P., Pontier, M. (eds) *Machine Medical Ethics (Intelligent Systems, Control and Automation: Science and Engineering).* Switzerland: Springer p. 341-362

Van Wynsberghe, A., 2014. To delegate or not delegate: Care robots, moral agency and moral responsibility. AISB 20-14 – 50th Annual Convention of the AISB, Goldsmiths, University of London, London, United Kingdom.

Walzer, M., 1983. *Spheres of Justice.* USA: Basic Books

Watcher, R., 2015. *The Digital Doctor.* USA: McGraw-Hill Education

Weizenbaum, J., 1976. *Computer power and human reason: from judgement to calculation.* Harmondsworth, England: Penguin Books Ltd

Wenar, L., 2021, *John Rawls*. Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/entries/rawls/ Accessed on 4 November 2021

Whitby, B., 2015. Automating medicine the ethical way. In: van Rysewyk, S.P., Pontier, M. (eds) *Machine Medical Ethics (Intelligent Systems, Control and Automation: Science and Engineering).* Switzerland: Springer p.223-232

WHO, 2021. *Ethics and governance of artificial intelligence for health: WHO guidance.* Geneva: World Health Organization. Available: https://www.who.int/publications/i/item/9789240029200 Accessed: 4 November 2021

Winfield, A., 2019. Ethical standards in robotics and AI. *Nature*. 2:46-48.

Winfield A.F.T., Jirotka M., 2018. Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philosophical Transactions of the Royal Society. A376: 20180085. http://dx.doi.org/10.1098/rsta.2018.0085

Wittgenstein, L., 1992. *Tractatus Logico-Philosophicus.* C. K. Ogden (trans.), London: Routledge & Kegan Paul. Available: https://www.gutenberg.org/files/5740/5740-pdf.pdf Accessed: 4 November 2021

Witting, C., 2005 Duty of Care: An analytical approach. *Oxford Journal of Legal Studies.* 25(1):p.33-63

Woodhall, B., 2018. Uber avoids legal battle with family of autonomous vehicle victim. *Reuters.* Available: https://www.reuters.com/article/us-autos-selfdriving-uber-settlement/uber-avoids-legal-battle-with-family-of-autonomous-vehicle-victim-idUSKBN1H5092 Accessed: 4 November 2021

Wu, A. W., 2000. Medical error: the second victim. *BMJ*. 320: p.726-727

Yeung, K., 2019. *Responsibility and AI.* Council of Europe study: DGI(2019)05. Strasbourg: Council of Europe

Zimmerman, M. J. (1992) Responsibility. In: Becker, L., Becker, C. (eds) 1992. *Encyclopaedia of Ethics Volume II:L-Index*. London: Garland Publishing Inc.

Appendix A: A compilation of biases, failed heuristics, and cognitive dispositions to respond

This table provides a visual demonstration of some of the various issues which can hamper decision-making. This collection has been compiled from the works of Bate *et al* (2012), Chapman *et al* (2013), Crosskerry (2002), Gutenstein (2014), and Stiegler and Tung (2014).

Cognitive problem	**Description**	**Example of potential outcome**
Affect. (Crosskerry, 2002; Stiegler and Tung, 2014)	Emotional influences on decision behaviour.	Clinician does not offer a treatment for fear of creating anger- for example offending a Jehovah's witness by offering blood transfusion even though it would be very helpful to their recovery from illness.
Aggregate bias. (Crosskerry, 2002)	Taking inferences from one population and applying it to another incorrectly.	Prescribing antibiotics for a viral illness (e.g., a cold) which has similar symptoms to a bacterial illness.
Anchoring bias. (Bate *et al*, 2012; Crosskerry, 2002; Stiegler and Tung, 2014)	Insufficient adjustment from initial assessment of a value or a state. Focussing on a single feature of the case at the expense of other details.	Focussing on cardiac origins of chest pain when the issue could be in the lungs rather than the heart.
Attribution bias. (Gutenstein, 2014)	Blames an individual rather than a situation.	Complaints from colleagues who have been unfairly blamed for negative patient outcomes.
Availability heuristic. (Bate *et al*, 2012; Crosskerry, 2002; Gutenstein, 2014; Stiegler and Tung, 2014)	Identifying by resemblance to previous, highly memorable events.	Clinician thinks that every new presentation of headache is meningitis as they had misdiagnosed it in another patient the week before.
***Bandwagon effect*.** (Bate *et al*, 2012)	"We do it this way here" regardless of other people's input or evidence to the contrary.	Offering patients a bedpan at night rather than assisting them to the toilet because of the ward's culture to not mobilise people at nighttime.
Bias blind spot.	A flawed sense of invulnerability to bias. Bias continues to affect	Individuals from patient groups are offered or denied aspects of care which

(Bate *et al*, 2012; Stiegler and Tung, 2014)	decision making and is not addressed by the individual.	would have been routinely offered to others.
Commission bias. (Crosskerry, 2002; Stiegler and Tung, 2014)	Tendency towards action rather than inaction. Better safe than sorry.	Patient requires a blood transfusion to replace the haemoglobin drop caused by excessive blood testing.
Confirmation/Ascertainment bias. (Crosskerry, 2002; Stiegler and Tung, 2014)	Looking for information which confirms a judgement rather than that which may refute it.	Pain reported by IV drug users considered to be drug seeking behaviour rather than thoroughly investigating the cause of the pain.
Default bias. (Gutenstein, 2014)	Avoiding making a choice rather than choosing.	Low organ donor rates.
Diagnosis momentum/creep. (Crosskerry, 2002)	To diagnose without adequate evidence. Another person's faulty thinking has been inherited and applied to the patient.	Asthma being mistaken for hayfever by a patient in the community and the diagnosis not being challenged when treatment sought from clinicians.
Feedback bias. (Stiegler and Tung, 2014)	Time-lapse between actions and consequences or absence or feedback subconsciously processed as positive feedback.	Clinician fails to control patient's pain as they failed to reassess effect after administering analgesia.
Framing. (Stiegler and Tung, 2014)	Interpretation of a situation which changes the perception and not the facts.	Stating 33% of people will die from the treatment (negative framing) rather than that 66% of people will be saved (positive framing).
Gambler's fallacy. (Bate *et al*, 2012; Crosskerry, 2002)	That a run of diagnoses cannot continue rather than taking each case on its merits.	"I've seen three people with this disease today, this cannot be a fourth."
Gender bias. (Crosskerry, 2002)	Gender of the patient exerting influence on clinical decision making.	Domestic violence being overlooked as a cause of trauma as the victim is male.
Hindsight bias. (Crosskerry, 2002)	Seeing events as having been predictable after the event than when the event was unfolding.	Seeing clouds in the sky and saying, "I knew it would rain" when it did later in the day.

Identifiable victim effect. (Gutenstein, 2014)	Individual outcomes more striking than a group's outcomes.	Expensive treatment offered to an individual which may not be offered to other members of the same group with the same need.
Implicit bias. (Chapman *et al*, 2013)	Stereotypes and prejudice which occurs without conscious awareness.	Female patients considered to be hysterical rather than having valid symptoms which require investigation.
Loss aversion. (Gutenstein, 2014; Stiegler and Tung, 2014)	Losses are more salient than gains.	Refusal of treatment.
Memory shifting. (Stiegler and Tung, 2014)	Failure to accurately recall information.	Clinician confuses one patient's history with another.
Multiple alternative bias. (Crosskerry, 2002)	Multiple potential alternatives can lead to irrational decision making.	Clinician always chooses to use one drug which they have always favoured when there are several new alternatives on the market now which may be better.
Omission bias. (Bate *et al*, 2012; Crosskerry, 2002; Stiegler and Tung, 2014)	Inaction rather than action.	Avoiding child vaccinations due to autism reports and neglecting the risk of harm from the preventable illness.
Order effects. (Crosskerry, 2002)	Tendency to remember the first and last items in an exchange, but not the middle. Information becomes lost as a result when communicating.	A nurse not remembering all the instructions given to her by the doctor.
Outcome bias. (Crosskerry, 2002)	Judges outcome of decision regardless of the actual decision quality.	A drunk driver arriving home safely thinks he made a good choice.
Overconfidence. (Crosskerry, 2002; Stiegler and Tung, 2014)	Inaccurately high self-assessment with regard to positive traits.	Clinician feeling that they could not possibly be wrong.

Playing the odds / Frequency gambling. (Crosskerry, 2002)	The clinician's opinion of the relative chances of a patient having a particular disease.	Acute aortic dissection being mistaken for early stages of constipation as benign conditions outnumber serious ones.
Posterior probability error. (Crosskerry, 2002)	Wrong diagnosis being perpetrated or new diagnosis being missed.	A patient who has presented with migraine six times in the last year presents with headache. This time there is actually has a new/different diagnosis which is being missed as the clinician assumes it is migraine again.
Preference for certainty. (Stiegler and Tung, 2014)	Human preference for certainty over risk at the expense of sacrificing a greater expected value.	Offering a treatment with a lower chance of success but a lower chance of complication than a treatment with a higher chance of success but carries a higher chance of complication.
Premature closure. (Crosskerry, 2002)	A diagnosis is accepted before it has been fully verified.	Patient assumed to be suffering with and treated for a narcotics overdose (as pill bottles found nearby) when the issue was actually low blood sugar levels from the insulin bottle that had not been found.
Representativeness heuristic. (Crosskerry, 2002; Stiegler and Tung, 2014)	Identifying by the degree or resemblance to preexisting or classic models.	Misdiagnosis of dementia when an infection could be causing the confusion in an elderly patient.
Search satisficing. (Bate *et al*, 2012; Crosskerry, 2002)	Having found one diagnosis, other co-existing conditions are not detected.	Missing a second facture in a trauma patient when one fracture has already been identified.
Sutton's slip. (Bate *et al*, 2012; Crosskerry, 2002)	Going for the obvious diagnosis.	Diagnosing musculoskeletal pain in a 28-year-old lady with chest pain and not considering cardiac origin.
Triage cueing. (Crosskerry, 2002)	Under/over appreciation of the acuity of the presenting patient's condition.	Patient with stomach ache thought to be low acuity left to wait in waiting room when the cause was a gastrointestinal bleed rather than food poisoning.
Unpacking principle. (Crosskerry, 2002)	The more specific a description (unpacking of information from the patient) we receive the	Failure to take a fully history from patient lead to the clinician not knowing about their recent long-haul flight which caused the pulmonary embolus, so the

	more likely an event is judged to be.	patient was treated for a presumed chest infection instead.
Vertical line failure. (Bate *et al*, 2012; Crosskerry, 2002)	Routine tasks lead to thinking in silos.	Missing a meningitis patient in an influenza outbreak.
Visceral bias. (Crosskerry, 2002; Stiegler and Tung, 2014)	To allow feelings to affect patient care.	A VIP patient getting priority treatment based on celebrity status when others have to wait their turn.
Yin-yang out / Serum rhubarb. (Crosskerry, 2002)	The patient has already had thorough work-up prior to the clinician seeing them, so low enthusiasm to investigate further as it is felt that nothing new or relevant shall be found.	A patient with Lupus misdiagnosed with Chronic Fatigue Syndrome which has as similar presentation.

References

Bate, L., Hutchinson, A., Underhill, J., Maskrey, N., 2012. How clinical decisions are made. *British Journal or Clinical Pharmacology.* 74(4) p.614-620.

Chapman, E.N., Kaatz, A., Carnes, M., 2013. Physicians and Implicit Bias: How Doctors May Unwittingly Perpetuate Healthcare Disparities. *Journal of General Internal Medicine*. 28(11) p.1504–1510.

Croskerry, P., 2002. Achieving quality in clinical decision making: cognitive strategies and detection of bias. *Academic Emergency Medicine.* 9 (11) 1184-1204

Gutenstein, M., 2014. Psychological factors in emergency medicine. *Emergency Medicine Australasia*. 26 p.295-299.

Stiegler, M.P.S. and Tung, A. 2014. Cognitive Processes in Anesthesiology Decision Making. *Anesthesiology.* 120(1) p.204-217

Appendix B: The 26 items of literature included in chapter 4’s literature review

In order of appearance in the review:

1. Nursing and Midwifery Council, 2018. *The Code for Nurses and Midwives*. Available: https://www.nmc.org.uk/standards/code/read-the-code-online/
2. General Medical Council, 2020. *Good medical practice: General Medical Council*. Available: https://www.gmc-uk.org/ethical-guidance/ethical-guidance-for-doctors/good-medical-practice
3. Health and Care Professions Council, 2016. *Standards of Conduct, Performance and Ethics*. Available at: https://www.hcpc-uk.org/standards/standards-of-conduct-performance-and-ethics/
4. Hengstler M., Enkel E., Duelli S., 2016. Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices. *Technological Forecasting & Social Change* 105: p.105–120
5. Armstrong, K., 2018. If You Can't Beat It, Join It: Uncertainty and Trust in Medicine. *Annals of Internal Medicine.* 168(11)818-819
6. Whitby, B., 2015. Automating medicine the ethical way. In: van Rysewyk, S.P., Pontier, M. (eds) *Machine Medical Ethics (Intelligent Systems, Control and Automation: Science and Engineering).* Switzerland: Springer p.223-232
7. House of Lords: Select Committee on Artificial Intelligence, 2018. *AI in the UK: ready, willing and able?* London: House of Lords. Available: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf Accessed: 4 November 2021
8. Association for Computing Machinery, 2018. ACM code of ethics and professional conduct. Available: https://www.acm.org/code-of-ethics
9. Mukherjee, S., 2017 A.I. versus M.D.: What happens when diagnosis is automated? *The New Yorker.* 3 April. Available: http://www.newyorker.com/magazine/2017/04/03/ai-versus-md Accessed: 4 November 2021
10. House of Commons: Science and Technology Committee, 2018. *Algorithms in decision-making*. London: House of Commons. Available: https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/351/351.pdf

11. Miles, A., 2007. 'Science: A limited source of knowledge and authority in the care of patients. A Review and Analysis of: 'How Doctors Think. Clinical Judgement and the Practice of Medicine.' Montgomery, K', *Journal of Evaluation in Clinical Practice*, 13(4): p.545-563
12. Sukel, K., 2017a. With a little help from AI friends. *New Scientist*. 235(3134): p.36-39
13. Hartman, D. E., 1986. On the Use of Clinical Psychology Software. Practical, Legal, and Ethical Concerns. *Professional Psychology: Research and Practice*. 17(5):462-465
14. Ross C., Swetlitz I., 2017. IBM pitched Watson as a revolution in cancer care. It is nowhere close. *STAT News*. https://www.statnews.com/2017/09/05/watson-ibm-cancer/ Accessed: 4 November 2021
15. Doroszewski, J., 1988. Ethical and methodological aspects of medical computer data bases and knowledge bases. *Theoretical Medicine.* 9(2): p.117-128
16. Van Wynsberghe, A., 2014. To delegate or not delegate: Care robots, moral agency and moral responsibility. AISB 20-14 – 50th Annual Convention of the AISB, Goldsmiths, University of London, London, United Kingdom.
17. Luxton, D.D., 2014. Recommendations for the ethical use and design of artificial intelligent care providers. *Artificial Intelligence in Medicine*. 62(1): p.1-10
18. Delvaux, M., 2017. *Report 27 January 2017; with recommendations to the Commission on Civil Law Rules on Robotics* (2015/2103(INL)) Available: https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html Accessed: 4 November 2021
19. Pouloudi, A., Magoulas, G.D., 2000. Neural expert systems in medical image interpretation: Development, use, and ethical issues. *Journal of Intelligent Systems.* 10(5-6): 451-472
20. Sukel, K., 2017b. Artificial Intelligence ushers in the era of superhuman doctors. *New Scientist*. Available: https://www.newscientist.com/article/mg23531340-800-artificial-intelligence-ushers-in-the-era-of-superhuman-doctors/ Accessed: 4 November 2021
21. Vallverdú, J., Casacuberta, D., 2015. Ethical and technical aspects of emotions to create empathy in medical machines. In: van Rysewyk, S.P., Pontier, M. (eds) *Machine Medical Ethics (Intelligent Systems, Control and Automation: Science and Engineering).* Switzerland: Springer p. 341-362
22. Fenech, M., Strukelj, N., Buston, O., 2018. Ethical, social and political challenges of artificial intelligence in health. *Future Advocacy report for the Wellcome Trust*. Available: https://wellcome.org/sites/default/files/ai-in-health-ethical-social-political-challenges.pdf Accessed: 4 November 2021

23. Inthorn, J., Tabacchi, M.E., Seising, R., 2015. Having the final say: Machine support of ethical decisions of doctors. In: van Rysewyk, S.P., Pontier, M. (eds) *Machine Medical Ethics: Intelligent Systems, Control and Automation: Science and Engineering*. Switzerland: Springer p.181-206
24. Kellmeyer, P., Cochrane, T., Müller, O., Mitchell, C., Ball, T., Biller-Andorno, J. J., Fins, N., 2016. The Effects of Closed-Loop Medical Devices on the Autonomy and Accountability of Persons and Systems. *Cambridge Quarterly of Healthcare Ethics*. 25(4): p.623-631
25. Government Office for Science, 2016. *Artificial Intelligence: Opportunities and Implications for the Future of Decision Making*. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/566075/gs-16-19-artificial-intelligence-ai-report.pdf Accessed: 4 November 2021
26. Bainbridge, D. I., 1991. Computer-Aided Diagnosis and Negligence. *Medicine, Science and the Law*. 31:127-136.