# Part I:
# Practical and theoretical foundations

## Chapter 2: Artificial intelligence's use in medicine

An accurate description of artificial intelligence's use in medicine must precede any normative or legal assessment thereof. The following chapter provides this in four stages. Section I. provides a definition of artificial intelligence (AI), a description of the structure and development of relevant techniques and the factors affecting their performance. Sections II. to IV. utilise this understanding to anticipate those aspects of AI functioning that will be problematic for the value of patient autonomy. Specifically Section II. offers case studies on the degrees of clinical AI automation, Section III. outlines the literature on the interpretability of AI and Section IV. explores their relation to established human clinical expertise.

### *I. Artificial intelligence*

#### A. Definition

AI has a long history as a field of multidisciplinary research. It is related to the fields of computer science,[116] statistics/mathematics,[117] engineering and neuroscience.[118] The aims of this research have shifted over time, partly corresponding to the demands of each particular discipline, and they remain contested. It is unclear whether the goal is to emulate a general form of intelligence,[119] or to solve practical problems that had previously

116 Alpaydin, *Machine Learning* (Revised Edition 2021) 19.
117 Morik in Bauer and others, *Applications in Statistical Computing* (2019).
118 Vieira, Pinaya and Mechelli in Mechelli and Vieira, *Machine Learning: Methods and Applications to Brain Disorders* (2019) 1.
119 Pennachin and Goertzel in Goertzel and Pennachin, *Artificial General Intelligence* (2007) 1.

been understood to require a certain degree of human intelligence – i.e. to build 'smart tools'.[120] One may also ask what it even means to speak of intelligence in systems. Is it a matter of structure, of behaviour, of rationality, of particular cognitive functions or of a capacity to solve specific problems?[121] All these are questions thrown up by the goal of developing and/or understanding intelligent machines. It is therefore unsurprising that it has been difficult to arrive at a consensus in the definition of AI. The outlined approaches and the aims underlying them may be related in many ways, but they are not compatible.[122]

Any research touching on AI must grapple with this issue, so that it may at least offer a working of definition of the subject.[123] A definition must be found that is tailored to the goal of the investigation. Presently this is to analyse a certain set of challenges (for autonomy) that arise from technologies – which both exist and are in a continuous, rapid state of development – and that aim to solve relatively narrow challenges in a certain context (health care). The focus is squarely on the *application* of such technologies. This is a different investigation than one that abstractly or theoretically seeks to outline the nature of AI.[124] Therefore, rather than getting hung up on 'abstract notions of intelligence', the goal is to facilitate the analysis of 'useful artifacts'.[125]

Oriented towards these priorities, our concern is with intelligence as 'the ability to solve hard problems'.[126] This approach has been summed up as a 'Capability-AI' definition, which serves as a referent '[f]or people whose

120 Nilsson, *The Quest for Artificial Intelligence* (2009) 508-518.

121 These contrasting approaches are explored in: Wang in Wang, Goertzel and Franklin, *Artificial General Intelligence 2008: Proceedings of the First AGI Conference* (2008).

122 Wang, 'On Defining Artificial Intelligence' (2019) 10(2) Journal of Artificial General Intelligence p. 1, 13-14.

123 ibid 2-6.

124 For similar approaches see Matheny and others, 'Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril' (2019) <https://nam.edu/artificial-intelligence-special-publication/> accessed 5.4.2020 13-14: 'This publication does not address the hypothetical (…) and focuses instead on the current and near-future uses and applications of AI'; Turner, *Robot Rules: Regulating Artificial Intelligence* (2019) 15: 'this book does not seek to lay down a universal, all-purpose definition of AI which can be applied in any context. Its aim is much less ambitious: to arrive at a definition which is suited to the legal regulation of AI'.

125 Russell, 'Rationality and Intelligence' (1997) 94(1-2) Artificial Intelligence p. 57, 57.

126 Minsky, *The Society of Mind* (First Edition 1988) 71.

interest in AI mainly comes from its potential applications'.[127] For these 'the intelligence of a system should be indicated by its problem-solving capability'.[128] Therefore, to offer a useful definition of AI we must relate it to a certain problem-solving context, concurrently restricting the research agenda. Here this context is circumscribed by the healthcare field.

It has been a longstanding objective of the computer science community to develop programmes that can master the kinds of tasks that human medical experts solve by drawing on their intuition, knowledge and skill. In the past there were already some limited advances in this field.[129] But it is only over the past decade or so that this objective has increasingly been realised. Computers are now able to perform complex cognitive functions associated with clinical decision making and especially with diagnostic, prognostic and therapeutic tasks.[130] To a lesser degree they are also assisting with motoric actions and interpersonal aspects of healthcare, which form demanding components of human medical expertise in their own right.[131] It is the technologies demonstrating these capabilities in the healthcare field that constitute 'AI' for the purposes of this work.

## B. Machine learning: the underlying technology

With this definition in mind our research can primarily be fixed on one type of AI technology: machine learning (ML). ML is distinct from previously dominant AI methods, referred to as good old fashioned or symbolic AI, and it has played an indispensable role in recent advances. It is devices with ML components that have begun to exhibit the outlined clinical expertise, most especially the cognitive capabilities involved in diagnoses, prognoses and the proffering of therapeutic advice.[132]

127 Wang, 'On Defining Artificial Intelligence' (2019) 10(2) Journal of Artificial General Intelligence p. 1, 10-11.

128 ibid 10.

129 See: Szolovits, *Artificial Intelligence in Medicine* (1982); Yu, Beam and Kohane, 'Artificial Intelligence in Healthcare' (2018) 2(10) Nature Biomedical Engineering p. 719, 719-722.

130 Braude in Schramme and Edwards, *Handbook of the Philosophy of Medicine* (2017) 702: 'Cognition refers to all mental processes related to knowledge, including but not limited to memory, attention, perception, representational schemas, consciousness, and language'.

131 ibid 706-712.

132 See Section II. below.

ML itself is a complex concept, serving as an umbrella term for many more specific techniques that accomplish these tasks.[133] As Burrell notes:

> popular machine learning models include neural networks, decision trees, Naïve Bayes, and logistic regression. The choice of model depends upon the domain (i.e. loan default prediction vs. image recognition), its demonstrated accuracy in classification, and available computational resources, among other concerns. Models may also be combined into 'model ensembles,'[134]

These techniques share the ability to automatically learn from data and to improve with experience: 'developers in ML program computers to find solutions on their own'.[135] They are able to autonomously adjust various aspects of their structure to find an efficient way to accomplish certain tasks, such as making classifications and making predictions about as-yet unseen data.[136] In addition, in contrast with traditional AI types these structures are also said to operate at a sub-symbolic level[137] – rather than consisting of units that are easily interpretable, with a clear conceptual meaning,[138] ML techniques tend to consist of 'fine grained dynamical features that are below the conceptual level'.[139] Relatedly, rather than attempting to model tasks on the basis of their logical description, making clear logical inferences,[140] the newer ML technologies function largely on the basis of probability theory – enabling these machines to deal with the ambiguities and uncertainties of life.[141] In making decisions ML algorithms are able to settle 'into an equilib-

133 Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective* (Second Edition 2020) 2-4.
134 Burrell, 'How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms' (2016) 3(1) Big Data & Society p. 1, 5.
135 Zednik, 'Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence' (2021) 34(2) Philosophy & Technology p. 265, 267.
136 Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable* (2019) 13-14.
137 Berkeley, 'The Curious Case of Connectionism' (2019) 2(1) Open Philosophy p. 190.
138 Sun in Frankish and Ramsey, *The Cambridge Handbook of Artificial Intelligence* (2014)114.
139 Berkeley, 'The Curious Case of Connectionism' (2019) 2(1) Open Philosophy p. 190, 200.
140 Morik in Bauer and others, *Applications in Statistical Computing* (2019) 130.
141 Alpaydin, *Machine Learning* (Revised Edition 2021) 32-35.

rium state in which a majority of (potentially contradictory) constraints are simultaneously satisfied'.[142]

Four overarching classes of ML are often outlined according to the purpose of the ML and the nature of its training. These are: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning.[143] Supervised learning refers to the situation where the aim is to draw an inference from an input (e.g. an X-ray) to an output decision (e.g. the identification of a pathology). Here there is the assumption that human experts possess the requisite knowledge to connect the two variables, with them labelling the output that should be reached.[144] Supervised learning is a means for the computer to teach itself this human knowledge by developing its own way of connecting the input to the output that the human supervisor says is correct. The focus of this type of learning is classification: 'the aim is to predict the class of each observation' – i.e. to predict the appropriate label.[145]

Unsupervised learning by contrast does not rely on labelling by humans. It seeks to discover ways of grouping data according to its own criteria, discovering 'patterns, classes or distinctive features that cannot be readily interpreted by a human observer or necessarily judged against established gold standards or ground truth.[146] One common application for unsupervised learning is clustering.[147] For example, the algorithm may cluster groups of patients together according to traits which it determines to be similar and arriving at new categories (labels) for them.[148] Other important tasks include detecting novelties or outliers in the data or reducing its

142 Boden in Frankish and Ramsey, *The Cambridge Handbook of Artificial Intelligence* (2014) 95.

143 E.g. Chang in Riaño, Wilk and Teije, *Artificial Intelligence in Medicine: 17th Conference on Artificial Intelligence in Medicine, AIME 2019, Poznan, Poland, June 26–29, 2019, Proceedings* (2019).

144 Deo, 'Machine Learning in Medicine' (2015) 132(20) Circulation: Cardiovascular Quality and Outcomes p. 1920, 1920.

145 Scarpazza and others in Mechelli and Vieira, *Machine Learning: Methods and Applications to Brain Disorders* (2019) 46.

146 Kellmeyer in Mechelli and Vieira, *Machine Learning: Methods and Applications to Brain Disorders* (2019) 337.

147 Igual and Seguí in Igual and Seguí, *Introduction to Data Science* (2017).

148 Li and others, 'Unsupervised Analysis of Transcriptomic Profiles Reveals Six Glioma Subtypes' (2009) 69(5) Cancer research p. 2091.

complexity ('dimensionality').[149] This may be done by grouping correlated features together – leaving only a smaller group of uncorrelated principal components for analysis.[150] Very often these forms of unsupervised learning are not applied in their own right, but are important elements in supervised learning approaches.[151] So-called autoencoders are one modality that may be used to generate labels for use in supervised learning in clinical research.[152] As such, unsupervised learning offers possibilities to automate dimensions of ML development, including aspects of feature selection and data processing.

This illustrates that, while it is important to understand the different purposes underlying the classes – and their contribution to AI capabilities and a more independent functioning – in evaluating practical applications it is often difficult to draw sharp distinctions. They must be considered as likely elements of a complex whole. The same thing is also intimated by semi-supervised learning. Under this head, ML techniques are utilised to learn from both labelled and unlabelled data, enabling the model to account for gaps in the human expertise that is provided. These algorithms are capable of recognising and incorporating patterns from the unlabelled data.[153] A concurrent benefit is that this mixture can limit the impact of biases that could be introduced into the programme by human labelling.[154]

The final class is reinforcement learning. Here the machine is target-oriented.[155] It reacts to signals of rewards/penalties, (usually in the form of a numerical value) rather exact labels for desired outcomes.[156] Obtaining such labels that sufficiently correspond to the relevant situation would be

149 Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective* (Second Edition 2020) 12.

150 Jolliffe, *Principal Component Analysis* (Second Edition 2002) 1.

151 Schmidhuber, 'Deep Learning in Neural Networks: An Overview' (2015) 61 Neural Networks p. 85, 89.

152 Stevens and others, 'Recommendations for Reporting Machine Learning Analyses in Clinical Research' (2020) 13(10) Circulation: Cardiovascular Quality and Outcomes 782-793, 783.

153 Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective* (Second Edition 2020) 12.

154 Chang in Wulfovich and Meyers, *Digital Health Entrepreneurship* (2020) 75.

155 Naeem, Rizvi and Coronato, 'A Gentle Introduction to Reinforcement Learning and its Application in Different Fields' (2020) 8 IEEE Access p. 209320, 209322.

156 Sun in Frankish and Ramsey, *The Cambridge Handbook of Artificial Intelligence* (2014) .111-112.

impractical in relation to many interactive, uncertain problems.[157] Thus positive/negative signals are automatically generated *via* a trial and error approach in the course of the relevant interactions with the environment.[158] The ML agent must both exploit ways in which it has effectively solved the relevant problem in the past (signalled by the relevant reward) and explore new ways in which the problem may be solved in order to progressively improve – these issues do not arise in either supervised or unsupervised learning.[159] Reinforcement learning too can be paired with other types of ML techniques and may fulfil functions that are closely analogous to those usually fulfilled by supervised learning in clinical research and practice.[160]

Each of the 'classes' of AI offer different capabilities. Their selection and combination can serve reduce the need for human input and oversight. Supervised learning is often associated with a greater degree of human involvement, as developers establish criteria for the machine's functioning, and with a more consistent operation, as functioning is optimised on a stable dataset.[161] This is to be contrasted with unsupervised learning, which was seen to do away with different kinds of human engineering, and with reinforcement learning. The latter's ability to learn from changing, uncertain environments means that it can pursue dynamic, interactive responses.[162]

Therefore, selecting unsupervised or reinforcement learning approaches – or pairing them with supervised learning – is indicative of the higher degree of automation (independence from human input and action) that makes AI technology less predictable and controllable.[163] A potentially un-

157 Sutton and Barto, *Reinforcement Learning: An Introduction* (Second Edition 2018) 2.

158 Vieira, Pinaya and Mechelli in Mechelli and Vieira, *Machine Learning* (2019) 13.

159 Sutton and Barto, *Reinforcement Learning: An Introduction* (Second Edition 2018) 3.

160 Stember and Shalu, 'Deep Reinforcement Learning With Automated Label Extraction From Clinical Reports Accurately Classifies 3D MRI Brain Volumes' (17.6.2021) <https://arxiv.org/pdf/2106.09812> accessed 6.3.2022: here reinforcement learning is used to classify 2D and 3D brain images.

161 Strauß, 'Deep Automation Bias: How to Tackle a Wicked Problem of AI?' (2021) 5(2) Big Data and Cognitive Computing p. 1, 5.

162 Shortreed and others, 'Informing Sequential Clinical Decision-Making Through Reinforcement Learning: An Empirical Study' (2011) 84(1-2) Machine Learning p. 109, 111.

163 Strauß, 'Deep Automation Bias: How to Tackle a Wicked Problem of AI?' (2021) 5(2) Big Data and Cognitive Computing p. 1, 5-9.

precedented degree of automation *via* ML techniques allows the technology to operate in ways that are less closely and less obviously aligned with the goals of human users in practice.[164]

## C. Specific features of ML models: the example of deep neural networks

To understand further, specific characteristics of ML it is worth illustrating a model through which these types of learning and their respective objectives are implemented. Hereby it is notable that one kind of algorithm is not necessarily restricted to one type of training outlined above. For instance, deep learning is one prominent approach that can be deployed to pursue all classes of AI learning: supervised, unsupervised, semi-supervised and reinforcement.[165] One deep learning algorithm may even be trained in sequence – first in an unsupervised and then in a supervised manner.[166] Again, one can see that a close connection between the different training methods often exists in practice. To provide some more concrete explanations of how ML functions, this section focuses on a type of deep learning, a deep neural network (DNN), which is trained in a supervised fashion.[167]

### 1. Sub-symbolic functioning

DNNs are species of artificial neural networks. Neural networks are made up of connected nodes that are often called neurons after their biological inspirations.[168] Neurons are essentially separate processing units where

164 ibid 5-6.

165 López-Rubio, 'Computational Functionalism for the Deep Learning Era' (2018) 28(4) Minds & Machines p. 667, 670.

166 ibid 671.

167 This particular modality features prominently in many viable clinical AI and illustrates many of ML's distinct attributes: 'Almost every type of clinician, ranging from specialty doctor to paramedic, will be using AI technology, and in particular deep learning, in the future. This largely involved pattern recognition using deep neural networks (DNNs) (…) that can help interpret medical scans, pathology slides, skin lesions, retinal images, electrocardiograms, endoscopy, faces, and vital signs': Topol, 'High-Performance Medicine: The Convergence of Human and Artificial Intelligence' (2019) 25(1) Nature Medicine p. 44, 44.

168 Buckner, 'Deep learning: A Philosophical Introduction' (2019) 14(10) Philosophy Compass p. 1, 2.

computations occur, taking values from the signals sent along the connections between them.[169] In a neural network these neurons are arranged in layers and function in parallel (i.e. many units can carry out their computations at the same time).[170] In the simplest form the network receives input signals in an input layer, processes this data in a hidden layer and provides a human-interpretable 'decision' *via* an output layer.[171] More exactly, the first layer receives inputs in the form of values from an input signal (e.g. corresponding to pixels in an image) and the inputs in the next layer come from the output signals generated by the activation of this initial layer of neurons.[172]

The activation of neurons is determined by three types of computation occurring in the input-neuron interaction, termed: weight, bias and activation function. Weight is the multiplication applied to each single input before it is computed in the neuron where it is received.[173] Within the relevant neuron two computations are then performed simultaneously.[174] Each neuron adds a constant term, *a bias*, to the sum of received weighted inputs. This pushes the sum in a direction that keeps the output in a desired range.[175] The resulting value is passed through a non-linear function that ultimately determines the degree of the neuron's activation and thus the output it provides to serve as an input for the next layer.[176] Hence this is the aforementioned activation function.

Overall one should note the sub-symbolic nature of this process. Each neuron is carrying out the simple task of utilising various values without clear semantic meaning and computing these *via* the outlined processes, which may be identifiable but are likewise without well-defined content. Depending on the size of the network, thousands or millions of these processes can occur simultaneously and in sequence.

169 Rumelhart, Hinton and McClelland in Rumelhart, James L. McClelland and PDP Research Group, *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations* (1999) 47.

170 ibid 47.

171 Buckner, 'Deep learning' (2019) 14(10) Philosophy Compass p. 1, 2.

172 Michelucci, *Applied Deep Learning: A Case-Based Approach to Understanding Deep Neural Networks* (2018) 84.

173 ibid 32-34.

174 ibid 34.

175 Erb, 'Introduction to Backpropagation Neural Network Computation' (1993) 10(2) Pharmaceutical Research p. 165, 167.

176 Vieira and others in Mechelli and Vieira, *Machine Learning: Methods and Applications to Brain Disorders* (2019) 159.

Particularly DNNs are liable to contain complex combinations of such computations. Instead of having just one hidden layer, they contain many hidden layers between the input and output. These can possess different activation functions, fulfilling different roles and they can combine the sub-symbolic operations of neurons to achieve increasingly abstract and sophisticated representations of the data. The ability to automatically learn hierarchical representations distinguishes deep learning from shallow networks and renders it so successful as an ML tool.[177] Yet these representations are still distributed – that is, spread across neurons that each represent small, feature-like entities rather than concepts.[178] It is difficult to isolate a concept by focussing on a pattern of neurons and to understand or anticipate how changing the functioning of some elements, will impact the representation as a whole.[179]

Consequently, it will not be possible to straightforwardly associate these representations in the hidden layers with human-understood concepts and looking into the hidden layers does little to indicate the interpretable criteria that are being used to reach the outcome.[180] This mode of functioning exemplifies many of the challenges that will be discussed later in terms of interpretable AI and the black box problem in Section IV.

## 2. The training process

With this outline of DNN structures we should turn to the training process. Exploring this further promotes our understanding of AI-related automation and interpretability. All ML algorithms will have some features that

177 ibid 158.

178 Rumelhart, Hinton and McClelland in Rumelhart, James L. McClelland and PDP Research Group, *Parallel Distributed Processing, Volume 1* (1999) 47.

179 Zednik, 'Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence' (2021) 34(2) Philosophy & Technology p. 265, 280: 'there is often no way of knowing in advance whether an intervention on a single parameter will change the relevant system's behavior entirely or else affect it in a way that is mostly or entirely imperceptible'. See also Rudin and others, 'Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges' (2022) 16 Statistics Surveys p. 1, 32: referring to DNN's they state: 'concepts that are completely unrelated could be activated on the same axis' and that 'vectors in the latent space are "impure" in that they do not naturally represent single concepts'.

180 There are attempts to extract certain kinds of criteria, I will come onto this below when I discuss explainable AI.

will be altered with experience; these are known as learnable parameters. In a DNN these are, for example, the weights of connections between neurons, which will be changed automatically in response to a learning rule.[181]

To illustrate this one can turn to a common type of supervised learning for DNN's known as backpropagation.[182] Here an input is provided to a model with randomly initialised learnable parameters. Signals are propagated forward through the network in the manner described above, activating a pattern of neurons and providing a given output.[183] This output is automatically compared with a labelled set of correct decisions (hence this is supervised learning) *via* an error function.[184] This 'represents how far off the network is from making accurate predictions based on the input'.[185] Initially it is to be expected that the DNN will be quite mistaken, given the random values it is created with. To optimise performance, the errors are propagated back up the layers to the front of the network, adjusting the weights differently according to a pre-determined rule. Hereby a single forward- and backpropagation can cause the machine to update its parameters across a complex, multilayer network.[186]

By incrementally adjusting these weights in response to training samples, which may be presented one at a time or in small batches, a DNN 'can converge on the solutions to a wide range of classification and decision problems'.[187] Such problems span the areas of natural language processing, computer vision, speech recognition and robotics.[188] One can therefore see how these capabilities are acquired without close human involvement and direction. This allows for a greater degree of automation, even in a super-

181 Rumelhart, Hinton and McClelland in Rumelhart, James L. McClelland and PDP Research Group, *Parallel Distributed Processing, Volume 1* (1999) 46.

182 Hosseini and others in Pedrycz and Chen, *Deep Learning: Concepts and Architectures* (2020) 2.

183 Erb, 'Introduction to Backpropagation Neural Network Computation' (1993) 10(2) Pharmaceutical Research p. 165, 167.

184 Buckner, 'Deep learning' (2019) 14(10) Philosophy Compass p. 1, 2. This can also be called loss function: Vieira and others in Mechelli and Vieira, *Machine Learning* (2019) 161; or cost function: Hosseini and others in Pedrycz and Chen, *Deep Learning* (2020) 9.

185 Hosseini and others in Pedrycz and Chen, *Deep Learning* (2020) 9.

186 ibid 10.

187 Buckner, 'Deep learning' (2019) 14(10) Philosophy Compass p. 1, 2.

188 Dube, *An Intuitive Exploration of Artificial Intelligence: Theory and Applications of Deep Learning* (2021) Part II.

vised learning context, and affects interpretability, as the system's design is not made dependent on human understanding.[189]

### 3. Data and performance evaluation

DNNs are only one example of ML models that learn to achieve a distinctive kind of functioning through a relatively automated training process. At the same time, the way in which human engineers select and structure the use of data in ML development remains a crucial prerequisite for success and for an assessment of the algorithm's performance. Stepping back from our analysis of DNNs, we can frame this aspect more widely, as an issue for ML models in general.

A basic distinction to be made here is between offline and online ML algorithms. This marks the difference between those that do not draw on data from the application environment to improve performance and those that do. The values of offline models are determined by their interactions with the dataset used for training and are then locked in. After the ML system is understood to be sufficiently capable of performing the human defined task on the training data, the learning algorithm ceases to be applied so that the model ceases to update its learnable parameters with the presentation of new data. When ML models are online they will initially still be trained on a training dataset, but they will continue to learn from incoming information during application. This entails a greater independence and unpredictability of performance, limiting the value of *ex ante* assessments.

These features of online AI do not recommend themselves to the sensitive healthcare field and for the moment it appears that the ML applications that have received regulatory approval are offline.[190] For this reason, the following discussion will be couched in terms of offline ML models. The characteristics of online learning will be mentioned where relevant.

189 Burrell, 'How the Machine 'Thinks'' (2016) 3(1) Big Data & Society p. 1, 6-7.

190 Minssen and others, 'Regulatory Responses to Medical Machine Learning' (2020) 7(1) Journal of Law and the Biosciences p. 1, 5. At the same time the FDA has, for instance, already made an action plan for the regulation of online algorithms: U.S. Food & Drug Administration, 'Artificial Intelligence and Machine Learning in Software as a Medical Device' (2021) <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device> accessed 6.3.2022.

Even in the case of offline learning, the relative independence of the learning process, the nature and quality of data and the way in which it is used to train and validate the machine's functioning are central to the performance it can learn to achieve. Additionally, given the limited knowledge of ML models' decision-making criteria, as exemplified by DNNs, close attention to these data-related processes will also provide critical information on the AI's performance.[191] It is during training where it is easy to follow which inputs are matched correctly to their outputs. For example, whether a patient's condition is correctly identified from their electronic health record.

Developers will then be able to derive evaluative metrics, assessing how accurately the algorithm is performing its function across the entire training data set. For example, it may arrive at the correct diagnosis in X% of cases. This has practical implications. If an algorithm performs poorly in categorising outputs in training data sets it may be subject to 'underfitting'.[192] That is, the model may be too simple to capture complex relations in the data, leading to a low accuracy.[193] In consequence, non-learnable parameters (also known as hyperparameters)[194] may have to be tuned, or a different model may have to be selected to accomplish the relevant task. By contrast, if the algorithm has a high accuracy, one may say it performs well in the training environment, although it will still be expected to fall short in some cases.

This indicator is still inadequate for a realistic judgment of AI utility. To truly understand whether there is a basis for applying an AI in a practical setting one further needs to know something about its generalisability. Even if a model is assessed as sufficiently accurate in a training environment, i.e. it has not underfit, this may be the result of a phenomenon known as 'overfitting'. Here machine learning is so flexible that it simply memorises the quirks of the training data.[195] A good example from the medical field

191 Krishnan, 'Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning' (2020) 33(3) Philosophy & Technology p. 487, 495-496.

192 Vieira, Pinaya and Mechelli in Mechelli and Vieira, *Machine Learning: Methods and Applications to Brain Disorders* (2019) 34-35.

193 Tayo, 'Simplicity vs Complexity in Machine Learning — Finding the Right Balance' (11.11.2019) <https://towardsdatascience.com/simplicity-vs-complexity-in-machine-learning-finding-the-right-balance-c9000d1726fb> accessed 6.3.2022.

194 Vieira, Pinaya and Mechelli in Mechelli and Vieira, *Machine Learning* (2019) 32.

195 Sejnowski, *The Deep Learning Revolution* (2018) 43.

is provided by Narla and others who designed an algorithm for diagnosing the malignancy of a skin lesion. They found that:

> the algorithm appeared more likely to interpret images with rulers as malignant. Why? In our dataset, images with rulers were more likely to be malignant; thus the algorithm inadvertently 'learned' that rulers are malignant[196]

This association between a ruler and malignancy is dependent on the precise nature of the training data. It incidentally holds for a particular set, but for obvious reasons it is unlikely to transfer to new examples. Thus, even if accuracy is high for one data set (the algorithm is not biased) it may fall dramatically when used for unseen examples (the algorithm has overfit). Given the complexity of DNNs they are particularly prone to exhibit this defect – even if various mitigation techniques are being developed.[197]

To have a basic indicator of generalisability, i.e. an absence of overfitting, AI development almost ubiquitously includes a testing stage. Here data that is not used during the training phase is presented to the trained algorithm in a separate testing phase and performance is assessed. If accuracy remains high, one can have some confidence that the model has not overfit during training and that, in consequence, it may be applicable to new data.

Going even further for medical AI, it is important to demonstrate that the ML can generalise to the specific clinical environment in which it is to be used. This is necessary for several reasons. Even if an algorithm is neither inappropriately underfit nor overfit, there is no guarantee that the data used for testing or training is representative of the groups on which it is to be deployed. A factor that must be assessed against the background of a well-documented propensity of ML, including medical ML, to reach decisions that disadvantage already vulnerable groups. This state of affairs can result from the selection of training and test data, from biased engineering of features and from a failure to consider the biases that may arise in human-machine interactions.[198]

---

196 Narla and others, 'Automated Classification of Skin Lesions: From Pixels to Practice' (2018) 138(10) The Journal of Investigative Dermatology p. 2108, 2108. See similarly: Afnan and others, 'Interpretable, Not Black-Box, Artificial Intelligence Should Be Used for Embryo Selection' [2021](4) Human Reproduction Open p. 1, 3.

197 Vieira and others in Mechelli and Vieira, *Machine Learning* (2019) 164.

198 Thomasian, Eickhoff and Adashi, 'Advancing Health Equity with Artificial Intelligence' (2021) 42(4) Journal of Public Health Policy p. 602. See also Alon-Barkat and Busuioc, 'Human-AI Interactions in Public Sector Decision-Making: 'Automation

Fundamentally one must remember that, unlike humans, AI are not able to 'basically recognize and adapt to changed situations or contexts'[199] and that performance may drop if there are relevant distinguishing characteristics between training or validation and deployment:

> Proper assessment of real-world clinical performance and generalisation requires appropriately designed external validation involving testing of an AI system using adequately sized datasets collected from institutions other than those that provided the data for model training. This will ensure that all relevant variations in patient demographics and disease states of target patients in real-world clinical settings are adequately represented in the system where it will be applied[200]

Training and validating ML performance on the right type of data is particularly important in healthcare where decisions have significant consequences and where there are numerous variables that can lead to different divisions between population groups. For instance, in a study investigating the IDx-DR device (which will be examined further below) it was found that its accuracy was robust for sex, race and ethnicity, but that there was a greater specificity for those aged over 65.[201]

Such variability will also often be co-determined by the highly variable clinical contexts. Price provides an excellent example of this when he distinguishes between AI application in high-resource settings (where they are often trained and tested) and low-resource settings (where they may be most needed and deployed):

> The most significant problem with applying algorithms developed in High-Resource Hospitals in lower-resource settings is that those algorithms are likely to make diagnoses and treatment recommendations that are systematically suboptimal in those lower-resource settings. These can arise in at least two different ways: differences in diagnoses and treatment recommendations based on systematically different patient

Bias' and 'Selective Adherence' to Algorithmic Advice' (2023) 33(1) Journal of Public Administration Research and Theory p. 153.

199 Strauß, 'Deep Automation Bias: How to Tackle a Wicked Problem of AI?' (2021) 5(2) Big Data and Cognitive Computing p. 1, 3.

200 Kelly and others, 'Key Challenges for Delivering Clinical Impact with Artificial Intelligence' (2019) 17(1) BMC Medicine p. 1, 4.

201 Abràmoff and others, 'Pivotal Trial of an Autonomous AI-Based Diagnostic System for Detection of Diabetic Retinopathy in Primary Care Offices' (2018) 1 NPJ Digital Medicine p. 1.

> populations, and differences in recommended treatments based on treatment rankings whose order shifts with available medical resources[202]

Similarly Cohen and others note:

> a model to identify patients with sepsis that was derived from data at ten community hospitals may need to be changed for use in a tertiary care center that serves a large transplant population or in hospitals that do not have an ICU[203]

This illustrates the many nuances at play in the healthcare context that point towards the need for clarity about the type of data that is used in AI training and testing and about the nature of its development.

Ideally it demands some form of external validation, in the sense of validation in the application-environment, definable in terms of specific regions or even hospitals.[204] There is also a need for such external analyses to be continually updated to account for any relevant shifting in patient populations that may occur over time.[205]

More generally it is necessary to not only assess the technical features of the algorithm's functioning, but also its interaction with the relevant medical environment. Knowing about generalisability in the abstract may not translate to improved outcomes in practice if the AI is not used properly, if it is not integrated into the clinical workflow or its recommendations are not accepted by users.[206]

Indeed, if the AI is making a recommendation without having accounted for the wider operation of the healthcare environment and the causal factors at play there, then it may be outright dangerous. This is exemplified by an algorithm that indicated that patients with a history of asthma have a

---

202 Price II, 'Medical AI and Contextual Bias' (2019) 33(1) Harvard Journal of Law and Technology p. 65, 91.

203 Cohen and others, 'The Legal and Ethical Concerns That Arise From Using Complex Predictive Analytics in Health Care' (2014) 33(7) Health Affairs (Project Hope) p. 1139, 1143.

204 NHSX, 'NCCID case study: Setting standards for testing Artificial Intelligence' (21.2.2022) <https://www.nhsx.nhs.uk/ai-lab/explore-all-resources/develop-ai/nccid-case-study-setting-standards-for-testing-artificial-intelligence/> accessed 6.3.2022.

205 Kelly and others, 'Key Challenges for Delivering Clinical Impact with Artificial Intelligence' (2019) 17(1) BMC Medicine p. 1, 3.

206 Garg and others, 'Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: A Systematic Review' (2005) 293(10) The Journal of the American Medical Association p. 1223, 1235-1236.

lower risk of dying from pneumonia.[207] This association only held because patients with such a history were prioritised for intensive treatment, which was the case in the training data. If the AI decision does not account for the specific proactive dimension, then it could disincentive intensive treatment, putting asthmatics at a greater risk of death. Ultimately, having some evidence from prospective, external validation of clinical ML models is a prerequisite for adequately gauging whether it alters clinical practice or improves clinical outcomes.[208]

It is therefore of some concern that a lack of evidence from such evaluations has been a noted defect in medical AI development.[209] Nor is such validation mandated to achieve market access. There is evidence that both the U.S. Food and Drug Administration (FDA) and Notified Bodies in the European Union (EU) have approved devices without such evidence.[210] Further it appears that some AI will, even if used in a clinical or hospital setting, not have passed through these regulatory procedures given that

207 Caruana and others, 'Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission' (Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney NSW Australia, 10.8.2015-13.8.2015).

208 Brajer and others, 'Prospective and External Evaluation of a Machine Learning Model to Predict In-Hospital Mortality of Adults at Time of Admission' (2020) 3(2) JAMA Network Open 1-14, 2.

209 Kelly and others, 'Key Challenges for Delivering Clinical Impact with Artificial Intelligence' (2019) 17(1) BMC Medicine p. 1, 4; Topol, 'High-Performance Medicine' (2019) 25(1) Nature Medicine p. 44, 45; Freeman and others, 'Use of Artificial Intelligence for Image Analysis in Breast Cancer Screening Programmes: Systematic Review of Test Accuracy' (2021) 374 BMJ (Clinical Research Edition) 1-15. Four common defects of AI studies in one field of medicine have also been outlined as: (1) No generalisability assessment (2) unbalanced data (3) small sample size and (4) a limited reporting of performance metrics: Curchoe and others, 'Evaluating Predictive Models in Reproductive Medicine' (2020) 114(5) Fertility and Sterility p. 921, 923.

210 Angus, 'Randomized Clinical Trials of Artificial Intelligence' [2020](11) The Journal of the American Medical Association p. 1043: 'the US Food and Drug Administration recently approved AI-enabled decision support tools (also called software as medical devices or SaMDs) for diagnosis of diabetic retinopathy on digital fundoscopy and early warning of stroke on computed tomography scans. In neither instance was approval based on any RCT evidence that the information provided by the SaMD improved care'. The former (IDx-DR) device has also received a CE-Mark: IDx LLC, 'Fully Automated Diagnostic Device Receives CE Certification; IDx LLC Planning For Rollout Across Europe' (6.5.2013) <https://www.prnewswire.com/news-releases/fully-automated-diagnostic-device-receives-ce-certification-idx-llc-planning-for-rollout-across-europe-206263101.html> accessed 7.3.2022.

they may not be classified as medical devices.[211] One can also see that in the UK the design of guidance for the testing of medical AI for implementation is very much still at the proof of concept stage.[212] Currently there is a situation where medical AI implementation is progressing rapidly, with a risk that the outlined performance metrics are not collected or not made available.

Lastly, it has also been thrown into doubt in how far traditional clinical trials or indicators, as well as enforced centralised oversight, can be achieved for AI in healthcare.[213] For instance, area under the curve (AUC) measures are 'a core professional method of evaluating diagnostic tools', allowing professional radiologists to gauge whether any such tool improves accuracy.[214] While developers of diagnostic AI also regularly cite this aggregate value as a primary indicator of performance – including in regulatory applications and research publications – its utility for this purpose has proven limited.[215] It has been hypothesised that this is due to the implicit, subjective and variable nature of the professional know-how that ML models are seeking to capture and the contestable assumptions that must therefore be made during training and testing.[216]

---

211 E.g. Currently it appears that Watson for Oncology (a device utilising ML techniques) has been implemented in hospitals in the U.S. without being subject to regulatory oversight. See: Ross and Swetlitz, 'IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close' (5.9.2017) <https://www.statnews.com/2017/09/05/watson-ibm-cancer/> accessed 28.3.2023. For a more general overview of the U.S. situation see: Price II, Sachs and Eisenberg, 'New Innovation Models in Medical AI' (2022) 99(4) Washington University Law Review p. 1121, 1125-1126, 1150-1151; Price II, 'Distributed Governance of Medical AI' (2022) 25(1) SMU Science & Technology Law Review p. 3: highlighting both the limits of central regulatory involvement and the limits on effective central oversight in the case of involvement.

212 NHSX, 'NCCID case study: Setting standards for testing Artificial Intelligence' (21.2.2022) <https://www.nhsx.nhs.uk/ai-lab/explore-all-resources/develop-ai/nccid-case-study-setting-standards-for-testing-artificial-intelligence/> accessed 6.3.2022.

213 Price II, 'Artificial Intelligence in Health Care: Applications and Legal Implications.' (2017) 14(1) The SciTech Lawyer p. 10, 11 and see: Price II, 'Distributed Governance of Medical AI' (2022) 25(1) SMU Science & Technology Law Review p. 3.

214 Lebovitz, Levina and Lifshitz-Assaf, 'Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What' (2021) 45(3) MIS Quarterly p. 1501, 1507.

215 ibid 1510.

216 ibid 1510: for instance it was found that 'only a narrow subset of the relevant diagnosis inputs was captured in the datasets underlying the ML model'.

## 4. Summary

On the basis of the described attributes of ML structure and development we should bear in mind the following aspects as we go forward. We are focussing on ML algorithms, which take multiple forms and can pursue different types of cognitive clinical problem-solving tasks. It was shown how the innate structure of one ML technique, a DNN, makes it difficult to comprehend the criteria by which it arrives at a decision. We also know that an ML model's performance is importantly shaped by the data on which it is trained as well as tested, giving rise to risks for certain groups and in certain implementation environments, and that a meaningful evaluation metric for this performance presupposes further validation procedures – especially in the protean medical context. That such an evaluation has taken place is not currently guaranteed by the regulatory environment and it is a shortcoming of many development procedures. As such, a degree of uncertainty around these metrics ought to be taken as a general, albeit not universal feature, of current medical AI.

With this description of AI one can now anticipate the sources of autonomy-related problems that will constitute the focus of this book's discussion. The first are the capabilities that were formerly the preserve of human professionals, but which are now, to some degree, exercised autonomously by the AI. Specifically, it must be elaborated how AI are solving the outlined cognitive problems of clinical practice and how they are interacting with human decision-makers. Secondly, one needs to have a deeper understanding of the ways in which ML may be understood to have or to lack the quality of interpretability in the medical context. This includes an elaboration of how existing AI are interpretable or not and the ways in which interpretability is or is not being enhanced by scientific or technological means.

The remainder of this chapter will focus on a description of these aspects. Section II. uses case studies to demonstrate the capabilities of medical AI and their relationship with established medical practice and human practitioners. Section III. elaborates on this relationship by exploring general limitations on human oversight over machine assistance in decision-making. This indicates why machine-generated problems persist in expert-mediated contexts. Section IV. details the state of the art on ML interpretability.

## II. Capabilities of clinical AI: case studies

At the time of writing there are already dozens of medical ML applications that possess the features outlined above and which have reached the implementation stage in the United States and the United Kingdom.[217] This is well-documented in survey studies cataloguing the regulatory approvals that grant market access to AI/ML-based medical devices.[218] Such approved devices may not include all instances of mature AI, but it can be supposed that they capture a significant proportion.[219]

Indeed, the nature of many of these applications has been verified through searches of medical device databases and through a review of the public resources offered by developers. In this respect it is important to note that medical devices from the United States may be overrepresented in this book's analysis, owing to the easily accessible federal database and the detailed information that is offered on devices' nature, benefits, risks and intended types of use. The EU and UK systems do not yet offer a functional equivalent, although the European EUDAMED database, which is currently under development, may remedy some of these omissions. In any case, by focussing on devices with market access, as well as those whose practical feasibility is supported by robust external evidence,[220] we are adhering to the outlined definition of AI. It is their capabilities as useful problem-solving tools that drive their approval and implementation.

---

217 One must recall in this regard also the caution in Chapter 1, that this is not to be equated with reimbursement and widespread adoption in the healthcare system. There are additional hurdles to this, beyond the exhibition of the discussed capabilities.

218 Muehlematter, Daniore and Vokinger, 'Approval of Artificial Intelligence and Machine Learning-Based Medical Devices in the USA and Europe (2015–20): A Comparative Analysis' (2021) 3(3) The Lancet Digital Health p. 195; Benjamens, Dhunnoo and Meskó, 'The State of Artificial Intelligence-Based FDA-Approved Medical Devices and Algorithms: An Online Database' (2020) 3 NPJ Digital Medicine. It is to be noted that existing U.S. and EU approvals are considered for these purposes, with the latter still conferring UK market access until 30 June 2023: GOV.UK, 'Regulating medical devices in the UK: What you need to do to place a medical device on the Great Britain, Northern Ireland and European Union (EU) markets' (1.1.2022) <https://www.gov.uk/guidance/regulating-medical-devices-in-the-uk> accessed 7.3.2022.

219 See for example the reference to Watson for Oncology *supra*.

220 This may take the form of evidence that the relevant ML is in being trialled for use or that it has demonstrated utility in clinical practice.

This section presents several case studies from this subset of devices. The purpose of these case studies is manifold. They illustrate concrete cognitive problems that ML is capable of solving. At the same time, they will also touch upon some of the problems associated with the opacity of these devices. At bottom however, they are intended to emphasise the non-uniform manner in which AI are interacting with human clinical expertise and are assuming new roles for machines in medical decision-making.[221] They offer a context-specific differentiation of the kinds of automation that are taking place and highlight that 'automation is not a straightforward perspective, but a choice'.[222] Understanding that choices are being made, and the kinds of trade-offs involved, provides concrete reference points for the subsequent legal analysis.

The following section is accordingly structured by reference to the different types of interaction that are envisaged between human and ML agents: (A.) the complementation of human expertise, the form of which essentially remains unchanged (B.) the partial replacement of human cognitive capabilities (C.) direct control over dimensions of clinical decision-making.

## A. Devices complementing human expertise

There are medical ML devices that do not lessen the human cognitive capabilities that are brought to bear on a particular patient's treatment. Rather, they are intended to provide an additional resource that guides and supports such human decision-making.

A good example is provided by the AI-Pathway Companion Prostate Cancer that is on the EU market.[223] It is an AI that provides treatment advice. Specifically:

> Natural Language Processing is used to extract and compile data relevant to the decision-making process from the radiology, pathology,

221 See Strauß, 'Deep Automation Bias: How to Tackle a Wicked Problem of AI?' (2021) 5(2) Big Data and Cognitive Computing p. 1 for a general comment on the different types of automation underlying AI implementation.

222 Tsoukias in Papathanasiou, Zaraté and Freire de Sousa, *EURO Working Group on DSS: A Tour of the DSS Developments Over the Last 30 Years* (2021) 156.

223 Siemens Healthineers, 'AI-Pathway Companion Prostate Cancer from Siemens Healthineers approved for use in Europe as medical device' (3.3.2020) <https://www.siemens-healthineers.com/fr-be/press-room/press-releases/pr-aipathwaycomp-ce.html> accessed 7.3.2022.

> genetics, and lab results (…) Algorithms search through the prostate cancer guidelines for recommendations that suit the patient's individual disease status based on his or her current available data. (…) Based on this data, AI-Pathway Companion Prostate Cancer displays the patient's current clinical situation and offers guideline-based recommendations for further steps to provide treatment in accordance with the medical evidence[224]

A common problem is encountered here, in that it is not clarified what types of techniques are leveraged to achieve the relevant functioning.[225] However, the explicit role of natural language processing in extracting data and offering guideline-conform recommendations is almost certain to utilise ML and, more than likely, deep learning.[226] This helps to accomplish the demanding cognitive task of providing bespoke judgments in individual cases. The device offers therapeutic advice that is tailored to the specific patient, to the system within which they are being treated and to the medical evidence. Its capabilities are brought to bear to visualise the current situation and the possible treatment options. The device thereby frames the exercise of human expertise and seeks to optimise treatment decisions. There is no intention to replace elements of human decision-making; the AI is a sophisticated support tool that assists multidisciplinary teams of human experts.[227]

---

224 ibid.

225 Muehlematter and others note this shortcoming even for the more transparent database: Muehlematter, Daniore and Vokinger, 'Approval of Artificial Intelligence and Machine Learning-Based Medical Devices in the USA and Europe (2015–20)' (2021) 3(3) The Lancet Digital Health p. 195, 201.

226 'Various NLP architectures—including rule-based, machine learning-based, and hybrid models—have been developed and studied to enhance the accuracy of clinical concept extraction. With the emergence of deep learning models, research on clinical concept extraction has shifted from traditional machine learning models that rely heavily on semantic and lexical features manually crafted by domain experts to deep learning models that can automatically learn feature representations (eg, word embeddings) from large volumes of unlabeled clinical text': Yang and others, 'Clinical Concept Extraction Using Transformers' (2020) 27(12) Journal of the American Medical Informatics Association p. 1935, 1935-1936.

227 Siemens Healthineers, 'AI-Pathway Companion Prostate Cancer from Siemens Healthineers approved for use in Europe as medical device' (3.3.2020) <https://www.siemens-healthineers.com/fr-be/press-room/press-releases/pr-aipathwaycomp-ce.html> accessed 7.3.2022.

Another example is the Acumen Hypotension Prediction Index Software, which utilises machine learning to carry out prognoses. Specifically it indicates the 'patient's likelihood of future hypotensive events' during surgery – by offering a risk score for a specific range of time.[228] It is a CE-marked and FDA approved device.[229] In addition it is to be noted that the implementation of this device was supported by a randomised control trial that demonstrated an 'ability to influence physician actions and change proximate patient outcomes'.[230] As with the AI-Pathway Companion Prostate Cancer, this software does not seek to replace physician expertise. It serves to provide information to a clinician who is responsible for the patient's condition and 'no therapeutic decisions should be made based solely on the Hypotension Prediction Index (HPI) parameter'.[231] Nevertheless, one can see that this device is introducing a unique form of judgment and influencing physician behaviour. Rather than relying on the established, overt clinical signs of hypotension, which occur relatively late, the data-driven model extracts and analyses subtle information from arterial waveforms to arrive at a prediction score in a way that was previously impossible.[232] This machine-determined prediction score and the early action it allows are the primary benefits that are offered to clinical decision-makers.[233]

228 U.S. Food & Drug Administration, 'De Novo Classification Request for Acumen Hypotension Prediction Index Feature Software' (16.3.2018) <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/denovo.cfm?id=DEN160044> accessed 7.3.2022; Angus, 'Randomized Clinical Trials of Artificial Intelligence' [2020](11) The Journal of the American Medical Association p. 1043.

229 Edwards Lifesciences, 'Edwards' Acumen Hypotension Prediction Index Launches In The U.S.' (18.3.2022) <https://www.edwards.com/ns20180319> accessed 7.3.2022; the latter's use of the de novo classification pathway is in itself something remarkable as it (1) illustrates the novelty of this type of device (2) imposes higher regulatory burdens than the more commonly used 501(k)procedure.

230 Angus, 'Randomized Clinical Trials of Artificial Intelligence' [2020](11) The Journal of the American Medical Association p. 1043.

231 U.S. Food & Drug Administration, 'De Novo Classification Request for Acumen Hypotension Prediction Index Feature Software' (16.3.2018) <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/denovo.cfm?id=DEN160044> accessed 7.3.2022.

232 Hatib and others, 'Machine-learning Algorithm to Predict Hypotension Based on High-fidelity Arterial Pressure Waveform Analysis' (2018) 129(4) Anesthesiology p. 663, 664.

233 U.S. Food & Drug Administration, 'De Novo Classification Request for Acumen Hypotension Prediction Index Feature Software' (16.3.2018) <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/denovo.cfm?id=DEN160044> accessed 7.3.2022.

This category represents the least independent mode of AI functioning, being closely aligned with established uses of computers as decision support systems. The AI essentially collects, curates and presents knowledge to a fully-qualified human specialist to instrumentalise as they see fit.[234] Nevertheless, there is empirical data that documents the, significant influence of automated tools in clinical decision making. For instance, non-AI evidence-retrieval tools have been shown to have the capacity to improve the accuracy of decisions, boost confidence in conclusions but also, negatively, introduce new kinds of errors.[235] As will be discussed in Section III. below, there is an increased potential for ML-based devices, purporting to contribute a degree of specialist judgment, to have such impacts.

### B. Devices (partially) replacing pre-existing cognitive capabilities

The models in the previous section introduced new types of judgment into the medical decision-making process, but they left intact existing human cognitive resources which are to be exercised in a patient's treatment. In this section we will see that the use of ML devices may form part of a choice to diminish such resources, even if they do not eliminate them entirely. Two examples serve to make this point.

The first is the IDx-DR device, which has received FDA approval and the CE-mark.[236] This uses multiple algorithms to look for different types of lesions that indicate diabetic retinopathy in recorded images of patients' eyes.[237] This is a condition that can lead to blindness in people with

234 See Kazzazi, 'The Automation of Doctors and Machines: A Classification for AI in Medicine (ADAM framework)' (2021) 8(2) Future Healthcare Journal p. 257, 260.

235 Westbrook, Coiera and Gosling, 'Do Online Information Retrieval Systems Help Experienced Clinicians Answer Clinical Questions?' (2005) 12(3) Journal of the American Medical Informatics Association p. 315.

236 U.S. Food & Drug Administration, 'De Novo Classification Request for IDx-DR' (11.4.2018) <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/denovo.cfm?ID=DEN180001> accessed 7.3.2022; IDx LLC, 'Fully Automated Diagnostic Device Receives CE Certification; IDx LLC Planning For Rollout Across Europe' (6.5.2013) <https://www.prnewswire.com/news-releases/fully-automated-diagnostic-device-receives-ce-certification-idx-llc-planning-for-rollout-across-europe-206263101.html> accessed 7.3.2022.

237 Grzybowski and Brona, 'Analysis and Comparison of Two Artificial Intelligence Diabetic Retinopathy Screening Algorithms in a Pilot Study: IDx-DR and Retinalyze' (2021) 10(11) Journal of Clinical Medicine p. 1, 4-5.

diabetes if not identified and treated. Without the device human-based screening would have to be carried out. This requires specially trained graders or ophthalmologists.[238] With the device, an eye care institution can rely on non-physician operators to detect this condition, without a need for specialists to over-read the results.[239] On the basis of the device's output a decision can be made on whether to refer the patient for further examination. The AI is said to function autonomously in this respect, although it remains embedded in a clinical institution.[240] It is envisaged that this institution will provide relevant information to patients - for instance that 'IDx-DR does not treat retinopathy'[241] – and will provide for their wider eye-examination needs. Essentially IDx-DR therefore replaces the patient's access to one relatively narrow type of proficiency: that necessary for screening for diabetic retinopathy. The level of human expertise that is brought to bear on this decision is reduced, supplanted by the machine. The recourse to equivalent human capabilities may become *de facto* barred but the AI does not abrogate human involvement altogether. Again, it bears emphasising that this situation is different to the previous two examples where the patient's access to human resources in their care remained undiminished.

The second, comparable example is Mia (Mammography Intelligent Assessment). This is a deep learning software that, inter alia, has been trialled in several NHS Trusts for use as a second reader in mammograms, diagnosing breast cancer and determining whether women should be recalled for further examination.[242] The expertise that is being supplied here is that of a second human expert. One human professional and the AI will initially be involved. If the human and AI agree, then the relevant decision to recall or not to recall is made. If they disagree then a further human arbitrator

238 ibid 1.

239 'No need for specialist overread or telemedicine call backs': Digital Diagnostics, 'IDx-DR' <https://www.digitaldiagnostics.com/products/eye-disease/idx-dr/> accessed 7.3.2022.

240 'IDx-DR is for medical professionals who want to provide patients with rigorously validated and ethically designed diagnostic results at the point-of-care': ibid.

241 U.S. Food & Drug Administration, 'De Novo Classification Request for IDx-DR' (11.4.2018) <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/denovo.cfm?id=DEN160044> accessed 7.3.2022.

242 United Lincolnshire Hospitals NHS Trust, 'ULHT trialling artificial intelligence software to support breast cancer screening' (16.8.2019) <https://www.ulh.nhs.uk/news/ulht-trialling-artificial-intelligence-software-to-support-breast-cancer-screening/> accessed 7.3.2022.

is engaged to make the final decision.[243] Evidently therefore the AI is altering the established pattern of human expert involvement. Whereas previously two human specialists would have been routinely involved in recall decisions, this is now reduced to one. But unlike with the IDx-DR example, the human knowledge brought to bear on every decision should remain – qualitatively if not quantitatively – at the same level: that of a human specialist.

Of course one could go further than these examples; the strongest sense in which AI are often portrayed as replacing the expertise of professionals is by granting the patient direct access to the device's abilities, without the involvement and/or oversight of other humans.

This appears to be occurring for certain very narrow diagnostic tasks, such as the detection of atrial fibrillation. For example, the FibriCheck device is 'indicated for self-testing by patients who have been diagnosed with, or are susceptible to developing, atrial fibrillation and who would like to monitor and record their heart rhythms on an intermittent basis'.[244] However, it is ambiguous quite how far these decisions are really removed from human mediation. It is noted on the FibriCheck website that 'state-of-the-art machine learning algorithms (…) automatically interpret these results and create an output towards the healthcare professional, who can visually confirm these findings'.[245] A degree of physician involvement is therefore very much envisaged.[246]

It is also to be emphasised that the capabilities of such devices appear to be at the lower end of the scale – fulfilling a highly targeted diagnostic task, which will usually have to be located in a broader care context. Topol makes

243 See Sharma and others, 'Large-Scale Evaluation of an AI System as an Independent Reader for Double Reading in Breast Cancer Screening' (2021) Pre-Print p. 1.

244 U.S. Food & Drug Administration, 'FibriCheck 510(k) Summary' (28.9.2018) <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K173872> accessed 7.3.2022.

245 Fibricheck, 'What is FibriCheck and how does it work?' <https://www.fibricheck.com/what-is-fibricheck-and-how-does-it-work/> accessed 7.3.2022.

246 This can also be found in comparable devices. For instance one study evaluating a similar device emphasised the importance of transmitting relevant information to the physician: 'The [Kardia Band] is the first smartwatch accessory cleared by the FDA and available to the general public without a prescription that claims to instantaneously detect AF and transmit this information to a patient's treating physician': Bumgarner and others, 'Smartwatch Algorithm for Automated Detection of Atrial Fibrillation' (2018) 71(21) Journal of the American College of Cardiology p. 2381.

this point by arguing that medical AI will not lead to a lack of clinician oversight 'across all conditions, across all time' and that forms of partial and conditional autonomy (where there is more general human oversight in the background) are likely the highest degrees of automation that ML will introduce into healthcare.[247] Overall human decision-makers are seen as 'critical checks, in their roles as decisional mediators'[248] and they are likely to appropriately fulfil this purpose in relation to targeted, verifiable tasks. Much more of this will be made in Section IV. below.

Consequently the focus of this book's analysis is not on fully automated clinical ML models. There may be important ways in which machines replace human cognitive capabilities, but human professional involvement is not altogether precluded. Again, it will be seen in Section III. how this constellation nevertheless enables AI to influence medical processes. Moreover, it will be seen in Part III. that this has significant implications for the types of legal analyses that are conducted.

### C. Devices determining dimensions of clinical decision-making

Among devices that augment or partially replace human expertise a separate category is due to those that effectively determine a significant aspect of the clinical decision-making process, even in the case of physician involvement. Such a distinction must necessarily be a matter of degree, as the framing of various medical decisions, or the weakening of professional oversight, can prove influential in various ways (see Section III. below and Chapter 3). In this sense, this category of AI poses merely the most direct challenge to the notion of human mediation by altogether precluding the possibility for effective oversight of certain decisions.

Accipio Ix provides an illustration of a medical device making such interferences. It has received a CE-mark and FDA approval.[249] It utilises a form

247 Topol, *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again* (2019) 87. Kazzazi also notes the majority of current AI applications 'demonstrate a clear and narrow function that is both accurate and safe (…) but require human instrumentation in order to prove effective' Kazzazi, 'The Automation of Doctors and Machines' (2021) 8(2) Future Healthcare Journal p. 257, 260.

248 Alon-Barkat and Busuioc, 'Human-AI Interactions in Public Sector Decision-Making' (2023) 33(1) Journal of Public Administration Research and Theory p. 153, 156.

249 U.S. Food & Drug Administration, 'Accipiolx 510(k) Summary' (26.20.2018) <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K182177>

of DNN known as a convolutional neural network to identify instances of acute intercranial haemorrhages in non-contrast computerised tomography (NCCT) head scans.[250] It thereby provides a diagnostic indication that is intended to assist, not replace, the assessment of a clinical professional. The professional can now draw on the device's assessment, but: '[i]ts results are not intended to be used on a stand-alone basis for clinical decision-making (…) or otherwise preclude clinical assessment of CT cases'.[251] To this extent, the device aims to complement human expertise, which remains fully applicable, much like the AI-Pathway Companion Prostate Cancer or the Acumen Hypotension Prediction Index Software.

What sets Accipio Ix apart is the fact that it triages the analysed NCCT scans for the clinician, who will often be engaged in the time-intensive task of analysing a large number of such scans. The machine is able to leverage its capabilities to read the scan almost instantaneously and this allows it to prioritise the review of patients without the possibility for human oversight. This benefits some and disadvantages others. Even if the human makes the ultimate decision on diagnosis and treatment, the machine has essentially made a triaging choice that represents a judgment on how to balance the needs of different patients.[252] Without effective human mediation this choice is susceptible to the shortcomings associated with AI decision-making. As Voter and others note in this context: 'a poorly performing [decision support system] can hinder a clinician by highlighting false-positive studies and promoting premature closure in falsely negative studies'.[253]

---

accessed 7.3.2022; MaxQ AI, Ltd, 'MaxQ-AI Receives CE Mark Approval for Accipio™Ix Intracranial Hemorrhage Artificial Intelligence Software Platform' (22.5.2018) <https://www.prnewswire.com/news-releases/maxq-ai-receives-ce-mark-approval-for-accipioix-intracranial-hemorrhage-artificial-intelligence-software-platform-300652488.html> accessed 7.3.2022.

250 MaxQ Artificial Intelligence, 'ACCIPIO®—Solution Architecture and Design: A White Paper' <https://www.maxq.ai/resources> accessed 7.3.2022.

251 U.S. Food & Drug Administration, 'Accipiolx 510(k) Summary' (26.20.2018) <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K182177> accessed 7.3.2022.

252 One should note attempts to generalise these kinds of triage capabilities: For instance in the UK's NHS there has been a trial of an AI symptom checker seeking to '"triage out" avoidable attendances that present at emergency departments, to combat growing demand': Mahase, 'Birmingham Trust and Babylon Health Discuss Pre-A&E Triage App' (2019) 365(l2354) BMJ (Clinical Research Edition).

253 Voter and others, 'Diagnostic Accuracy and Failure Mode Analysis of a Deep Learning Algorithm for the Detection of Intracranial Hemorrhage' (2021) 18(8) Journal of the American College of Radiology p. 1143, 1144.

## III. Interpretability of AI

A task that is necessarily prior to understanding how decision-makers interact with these forms of ML functioning, is to determine what they will be in a position to know about them. As has been referred to, the technology underlying many cutting-edge medical AI is subject to a pervasive problem of 'opacity', sometimes referred to as the 'black box problem'.[254] Hereby human users of ML algorithms are said to systematically lack epistemic access to a relevant element;[255] an element that is often referred to as a specific kind of knowledge or understanding.[256] This section seeks to expand upon the nature of this element for a clinical specialist and a patient that are using AI. This involves stipulating what factors are (1) relevant and (2) systematically obscured by the outlined uses of ML.

The relevance of certain facts on AI functioning is often discussed in the literature under the head of interpretability, understandability, transparency or accountability.[257] The precise meaning of these concepts, representing opacity's counterpart, is hotly contested and it is left intentionally ambiguous here. The aim is to provide an overview of the kinds of factors that fall under this head, the reason for their being obscured and whether there are trends that are likely to resolve these issues as AI-deployment progresses in the near-future. As interpretability and opacity are relative concepts[258] it is worth noting again that the perspectives of the physician and patient are deemed central to this discussion and that the latter's understanding is often mediated through the former.

254 Zednik, 'Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence' (2021) 34(2) Philosophy & Technology p. 265, developing the concept of Humphreys, 'The Philosophical Novelty of Computer Simulation Methods' (2009) 169(3) Synthese p. 615, 618.

255 Zednik, 'Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence' (2021) 34(2) Philosophy & Technology p. 265, 268-269.

256 See e.g. Rudin and others, 'Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges' (2022) 16 Statistics Surveys p. 1, 2-11.

257 Krishnan, 'Against Interpretability' (2020) 33(3) Philosophy & Technology p. 487, 488; Binns, 'Algorithmic Accountability and Public Reason' (2018) 31(4) Philosophy & Technology p. 543, 544.

258 Humphreys, 'The Philosophical Novelty of Computer Simulation Methods' (2009) 169(3) Synthese p. 615, 618.

## A. Prevalent types of opacity

An objection that is often levelled at ML algorithms is that their 'inner workings' are not accessible to users.[259] One way in which this criticism has been framed is as an inability to examine the code of the model, 'the variables – learnable parameters and/or abstract representational structures – that mediate the transformation of inputs to outputs'[260] and how these behave or have behaved when interacting with data. The missing piece of information is that users are not aware of the internals of the black box model.[261] A possible source for this inability is technical; it stems from the characteristic complexity of ML models, paired with time- and capability-limitations on users.[262] An adequate inspection of the internal operations becomes unfeasible in practice.

More commonly, such a lack of knowledge of the makeup of an ML tool, will also stem from a non-technical source, from corporate secrecy.[263] This was seen to generate uncertainty even around the basic fact of whether devices employ ML techniques. While such secrecy is neither an inevitable fact of AI deployment nor exclusive to it,[264] the fostering of opacity as an 'intentional form of self-protection by corporations intent on maintaining their trade secrets and competitive advantage'[265] with respect to devices using AI is well-documented in medicine and beyond.[266] Specifically with regard to ML, the maintenance of corporate secrecy around model func-

259 Lakkaraju and Bastani, '"How Do I Fool You?": Manipulating User Trust via Misleading Black Box Explanations' (AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, USA, 07.02.2020-09.02.2020); Krishnan, 'Against Interpretability' (2020) 33(3) Philosophy & Technology p. 487, 495.

260 Zednik, 'Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence' (2021) 34(2) Philosophy & Technology p. 265, 271.

261 Lakkaraju and Bastani, '"How Do I Fool You?": Manipulating User Trust via Misleading Black Box Explanations' (AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, USA, 07.02.2020-09.02.2020).

262 Guidotti and others, 'A Survey of Methods for Explaining Black Box Models' (2019) 51(5) ACM Computing Surveys p. 1, 6.

263 Burrell, 'How the Machine 'Thinks'' (2016) 3(1) Big Data & Society p. 1, 3-4.

264 Rudin cites the proprietary COMPAS risk prediction tool as an example of a non-ML model that would, but for corporate secrecy, be interpretable: Rudin, 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead' (2019) 1(5) Nature Machine Intelligence p. 206, 209.

265 Burrell, 'How the Machine 'Thinks'' (2016) 3(1) Big Data & Society p. 1, 3.

266 Rudin and Ustun, 'Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice' (2018) 48(5) Interfaces p. 449.

tioning can be seen as an intentional strategy that companies adopt to avoid responsibility for the quality of individual predictions, while at the same time being able to capitalise from the purported capabilities of providing accurate predictions at the granular level.[267]

Ultimately, whether it is caused by corporate secrecy or technical complexity, an inability of users to understand the form, content and functioning of the underlying computational mechanisms is one prevalent form of opacity in ML.

The second form of opacity concerns the way in which the functioning of ML algorithms is related to external factors and forms of justification. Hereby there are difficulties in understanding the 'environmental patterns and regularities' that are being tracked, which 'features of the environment' are represented by variables in the relevant model and how the generation of outputs *via* causal mechanisms in the model relates to justificatory reasons for it.[268] The technical features of common ML techniques, like DNNs, contribute substantially to a disjuncture between (1) a technical knowledge of the computations being performed by the model (in so far as the user or another actor is able to access these) and (2) a set of general criteria that can serve to indicate to (any) human what information is flowing through it and how and why it is being manipulated and classified in a relevant way.[269] It is neither possible to understand all of the model at once, nor to intuitively explain the role of specific elements.[270] In consequence, even when access to the internal workings of an algorithm is granted, obtaining knowledge about its 'reasoning process',[271] 'the underlying rationale of …

267 Rudin, 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead' (2019) 1(5) Nature Machine Intelligence p. 206, 210.

268 Zednik, 'Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence' (2021) 34(2) Philosophy & Technology p. 265, 279; Krishnan, 'Against Interpretability' (2020) 33(3) Philosophy & Technology p. 487, 493-494.

269 Funer, 'The Deception of Certainty: How Non-Interpretable Machine Learning Outcomes Challenge the Epistemic Authority of Physicians' (2022) 25(2) Medicine, Health Care and Philosophy p. 167, 172.

270 Krishnan, 'Against Interpretability' (2020) 33(3) Philosophy & Technology p. 487, 490.

271 Afnan and others, 'Interpretable, Not Black-Box, Artificial Intelligence Should Be Used for Embryo Selection' [2021](4) Human Reproduction Open p. 1, 2.

machine-learning components?'[272] and the link of this to some form of ground truth is not forthcoming in many ML devices.

For example, if an AI is being used to select embryos for in vitro fertilisation (IVF) treatment, then one may be left asking what categories have been learned and applied. One may wonder whether cleavage rate, symmetry, etc. are relevant and, if so, how are they combined with other factors.[273] It may not be deducible whether a particular aspect of the patient's personal life has already been taken into account,[274] or what benefits, risks and limitations can be associated with a utilisation of the device.[275] It may further not be possible to check for errors, including the identification of forms of reasoning that are obviously wrong, in real-time.[276]

Beyond this, there are two important relations between mathematical processes and general criteria, which are particularly relevant to the physician-patient interaction, but which are obscured by this disjuncture. On the one hand there is the 'difficult task of translating information about causal processes into considerations relevant to the justification of a categorization'.[277] For users of AI and for decision-subjects this makes it unclear what kind of justification-strategy is being pursued for the output, since there are many different kinds for any given decision.[278] As a result, a physician or patient will struggle to incorporate the recommendation into their wider decision-making processes.[279] It is hidden from examination whether the selection and implementation of a justification is of a kind that would be deemed adequate by the patient.

---

272 Guidotti and others, 'A Survey of Methods for Explaining Black Box Models' (2019) 51(5) ACM Computing Surveys p. 1, 2.

273 Afnan and others, 'Interpretable, Not Black-Box, Artificial Intelligence Should Be Used for Embryo Selection' [2021](4) Human Reproduction Open p. 1, 4.

274 Funer, 'The Deception of Certainty' (2022) 25(2) Medicine, Health Care and Philosophy p. 167, 175.

275 Afnan and others, 'Interpretable, Not Black-Box, Artificial Intelligence Should Be Used for Embryo Selection' [2021](4) Human Reproduction Open p. 1, 4.

276 Afnan and others, 'Interpretable, Not Black-Box, Artificial Intelligence Should Be Used for Embryo Selection' [2021](4) Human Reproduction Open p. 1, 3; Parikh and others, 'Why Interpretable Causal Inference is Important for High-Stakes Decision Making for Critically Ill Patients and How To Do It' (2022) Preprint.

277 Krishnan, 'Against Interpretability' (2020) 33(3) Philosophy & Technology p. 487, 494.

278 Binns, 'Algorithmic Accountability and Public Reason' (2018) 31(4) Philosophy & Technology p. 543, 544.

279 Funer, 'The Deception of Certainty' (2022) 25(2) Medicine, Health Care and Philosophy p. 167, 173.

On the other hand, there are inevitable normative and epistemological assumptions that are embedded in the finished device and in the way that it has been developed and tested. Specifically in medicine, there will have to be a 'selection of parameters deemed relevant, the weighting of each possible treatment goal, the choice of means to achieve the selected goals with their respective consequences for the patient's life'.[280] In addition, there may be constraints embedded in an ML algorithm that represent an attempt to prevent algorithmic discrimination.[281] Whether intentional or inadvertent, these are choices with normative implications that are contestable[282] and they may be obscured by the second type of opacity, especially if they are made against a backdrop of corporate opacity.

The same may be said of epistemic assumptions that can be found in the code itself, as well as in the ways in which it is optimised and tested. For instance, an AI may be trained and assessed only with one kind of uncertainty in mind, ignoring another.[283] Or, ML results may be categorised into 'good' and 'poor' quality, giving the impression of high capabilities, when a distinction should in fact be made between results of a similar 'good' quality.[284]

In light of the omnipresence of such assumptions – and without knowing how the mathematical processes of the algorithm are incorporating them into their reasoning – the provision of external epistemic justifications for the AI's use by the developer gains added significance. External evaluation metrics – such as accuracy and reliability – will have to be drawn upon to serve as arguments for the deployment of clinical ML devices.[285] Users will have to assess the rigour and the underlying assumptions incorporated into

280 ibid 173.

281 Binns, 'Algorithmic Accountability and Public Reason' (2018) 31(4) Philosophy & Technology p. 543, 547.

282 One useful example of such choices in the domain of fairness can be found in: Grgic-Hlaca and others, 'Human Perceptions of Fairness in Algorithmic Decision Making' (WWW '18: Proceedings of the 2018 World Wide Web Conference, Lyon, France, 23.04.2018-27.4.2018).

283 Bhatt and others, 'Explainable Machine Learning in Deployment' (Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27.01.2020-30.01.2020).

284 Afnan and others, 'Interpretable, Not Black-Box, Artificial Intelligence Should Be Used for Embryo Selection' [2021](4) Human Reproduction Open p. 1, 2.

285 Krishnan, 'Against Interpretability' (2020) 33(3) Philosophy & Technology p. 487, 495-496.

the training data (so-called 'pre-model techniques'),[286] as well as those that are implicit in the performance metrics gathered afterwards.

A good example is provided by Lebovitz and others.[287] In their study, U.S. hospital managers were considering whether to adopt five ML models and they took into account, amongst other things, the expertise of those who created 'ground truth' labels for the supervised training of ML models, and the compatibility of such labels with professional standards and the professional 'know-how' of local experts.[288] They also performed their own local studies of the devices.[289]

For the adoption and clinical use of otherwise opaque ML devices, external evaluative metrics gain an added significance. Yet it is contested whether they can truly stand in for an analysis of the machine's reasoning process.[290] Corporate opacity continues to pose issues, even for this alternative rationale,[291] and given the problems highlighted in Section I.C.3. there are still pitfalls that remain common, perhaps intractable,[292] in their collection.[293]

286 Okay, Yildirim and Ozdemir, 'Interpretable Machine Learning: A Case Study of Healthcare' (2021 International Symposium on Networks, Computers and Communications (ISNCC), Dubai, United Arab Emirates, 10.31.2021-11.2.2021).

287 Lebovitz, Levina and Lifshitz-Assaf, 'Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What' (2021) 45(3) MIS Quarterly p. 1501.

288 ibid 1512-1517.

289 ibid 1510.

290 Funer, 'The Deception of Certainty' (2022) 25(2) Medicine, Health Care and Philosophy p. 167, 172-173.

291 Lebovitz and others highlight this when they document the frustration of managers who were not able to assess the source of the ground truth labels for one ML tool: Lebovitz, Levina and Lifshitz-Assaf, 'Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What' (2021) 45(3) MIS Quarterly p. 1501, 1509.

292 Funer, 'The Deception of Certainty' (2022) 25(2) Medicine, Health Care and Philosophy p. 167, 176; Bjerring and Busch, 'Artificial Intelligence and Patient-Centered Decision-Making' (2021) 34(2) Philosophy & Technology p. 349, 368.

293 Curchoe and others, 'Evaluating Predictive Models in Reproductive Medicine' (2020) 114(5) Fertility and Sterility p. 921, 923.

## B. Solutions to interpretability problems

Given that this is a forward-looking analysis, it is important to ask whether problems around interpretability will persist as AI-based devices are deployed more widely in medicine, or whether solutions can be anticipated. The focus in this regard is not so much on how the veil of corporate secrecy may be pierced, but on how the underlying technical difficulties may be alleviated. The former is dependent on contingent regulatory interventions that do not appear to be existent or forthcoming at this stage. Some extent of secrecy is therefore taken as a pervasive, albeit not inevitable, fact of clinical AI. By contrast, there is widespread agreement that, before regulatory interventions can be effective in dealing with the black box problem, there would have to be adjustments to the currently unavoidable difficulties associated with the underlying technology.[294] Two avenues can be pursued in this regard that are often distinguished by the labels explainability and interpretability.[295]

### 1. Explainability

Explainability is used where ML devices are deployed without alterations to their opaque functioning but with an additional model bootstrapped to them.[296] This latter model will be designed to be more understandable and to approximate the device's functioning, reconstructing an *explanation* for the way in which the model is working, without actually tracking the causal processes within it.[297] Such techniques may seek to indicate the logic behind a model's overall performance,[298] to measure one specific property,

294 Guidotti and others, 'A Survey of Methods for Explaining Black Box Models' (2019) 51(5) ACM Computing Surveys p. 1, 2; Rudin and Ustun, 'Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice' (2018) 48(5) Interfaces p. 449, 450.

295 Rudin, 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead' (2019) 1(5) Nature Machine Intelligence p. 206.

296 Laugel and others, 'The Dangers of Post-Hoc Interpretability: Unjustified Counterfactual Explanations' (Twenty-Eighth International Joint Conference on Artificial Intelligence, Macao, China, 8.10.2019-8.16.2019) 2802.

297 Guidotti and others, 'A Survey of Methods for Explaining Black Box Models' (2019) 51(5) ACM Computing Surveys p. 1, 10-11.

298 ibid 11. This is sometimes referred to as global explainability.

such as 'sensitivity to attribute changes',[299] or to establish reasons for specific results.[300]

For instance, saliency maps are employed to provide grounds for the output of an image analysis.[301] Such maps are intuitive because they visualise the parts of an image that an ML device is concentrating on. In medicine they may be used to indicate whether the algorithm is picking up on an area or pathology that a human expert would deem relevant or irrelevant.[302]

Counterfactual reasoning techniques are another prominent type of approach that explains ML decisions. They aim to identify a factor in the input data that, if changed, would alter the prediction.[303] Such models are easily interpretable by humans as they 'provide a minimal amount of information capable of altering a decision, and they do not require the data subject to understand any of the internal logic of a model in order to make use of it'.[304]

## 2. Interpretability

Interpretability is used to refer to devices that are developed with the aim of being more transparent and understandable to humans. By contrast with explainable approaches, the model itself and the actual causal reason for a given output are made more accessible to users.[305]

In effect this means that different kinds of constraints are imposed on the device to align its computational functioning with human capacities

299 ibid 14.

300 ibid 13-14. This is often termed local explanation to contrast it with the aforementioned global type.

301 Saporta and others, 'Benchmarking Saliency Methods for Chest X-Ray Interpretation' [2022](4) Nature Machine Intelligence p. 867.

302 Saporta and others, 'Benchmarking Saliency Methods for Chest X-Ray Interpretation' [2022](4) Nature Machine Intelligence p. 867, 867-868.

303 Laugel and others, 'The Dangers of Post-Hoc Interpretability: Unjustified Counterfactual Explanations' (Twenty-Eighth International Joint Conference on Artificial Intelligence, Macao, China, 8.10.2019-8.16.2019) 2802.

304 Wachter and others, 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR' (2017) 31(1) Harvard Journal of Law & Technology p. 841, 851.

305 Rudin, 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead' (2019) 1(5) Nature Machine Intelligence p. 206, 206.

for reasoning.[306] For instance, a constraint of sparsity (a low number of features)[307] on some forms of data may align the model with humans' limited capacity for handling 'at most 7 ± 2 cognitive entities at once'.[308]

Such constraints cannot, however, be determined in the abstract. Sparsity would make little sense in different contexts: the analysis of an image would not be easier to comprehend if it uses only a limited number of pixels.[309] Here a useful interpretability constraint would be for the model to pursue a process of case-based reasoning. This emulates a well-established human problem-solving technique: by leveraging known solutions to past situations, it solves a new problem.[310] For example, by highlighting what feature of an image has been extracted, what features in past images it compares this to and how it has been combined with other information to reach the new result, the computer would engage in a human-interpretable form of reasoning.[311] In a healthcare context a physician could then understand that a given output has been provided for a current patient because of the way(s) they compare to past patients.[312]

For the outlined DNN's more interpretable models may be designed by disentangling the flow of information through the neurons. Whereas it has been described above how neurons are typically associated with several concepts. so that it is difficult to determine how altering individual variables will impact overall model performance, there are attempts to associate neurons in a certain layer with individual concepts.[313] For instance, when classifying a bedroom, different neurons in one layer may be associated with '"lamp," "bed," "nightstand," "curtain."' and '[a]ll information about the concept up to that point in the network travels through that concept's

306 ibid 206.
307 Carvalho, Pereira and Cardoso, 'Machine Learning Interpretability: A Survey on Methods and Metrics' (2019) 8(8) Electronics p. 832.
308 Rudin, 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead' (2019) 1(5) Nature Machine Intelligence p. 206, 206.
309 Rudin and others, 'Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges' (2022) 16 Statistics Surveys p. 1, 5.
310 ibid 25.
311 ibid 26-28.
312 Parikh and others, 'Why Interpretable Causal Inference is Important for High-Stakes Decision Making for Critically Ill Patients and How To Do It' (2022) Preprint.
313 Rudin and others, 'Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges' (2022) 16 Statistics Surveys p. 1, 31-33.

corresponding neuron'.[314] By seeing what concepts are being identified, the flow of information becomes more transparent to human users.

### 3. Evaluation

Both types of approaches to solving the opacity problem of AI have benefits and disadvantages. Yet the reason why this section refers to the interpretability of AI, rather than their explainability, is that only the former offers satisfactory solutions in the sensitive healthcare context. While the explainable AI movement purports to offer a solution, and thereby to increase public trust, protect individual rights and mediate the risks posed by black box tools,[315] it has the potential to worsen many of the issues associated with them.

The most glaring issue has been pointed out by Rudin: by definition the models approximating to the black box are only that, approximations, and they may generate 'explanations that are not faithful to what the original model computes'.[316] In medicine there is a resulting risk that for a percentage of cases, or a specific subset of them, the provided reasons do not match with the predictions that the medical device is making. This is evidenced by the experiments of Saporta and others, where the utilised heatmaps were more likely to be erroneous for certain kinds of pathologies.[317] This would limit trust in the explanation,[318] re-invigorating or compounding doubts about the original ML device.

Moreover, even where explanations do match predictions, the two may be appealing to 'completely different features'.[319] Laugel and others frame this problem in the context of the aforementioned counterfactual explana-

314 ibid 28.

315 Such arguments are deployed in Wachter and others, 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR' (2017) 31(1) Harvard Journal of Law & Technology p. 841.

316 Rudin, 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead' (2019) 1(5) Nature Machine Intelligence p. 206, 207.

317 Saporta and others, 'Benchmarking Saliency Methods for Chest X-Ray Interpretation' [2022](4) Nature Machine Intelligence p. 867, 873-874.

318 Rudin, 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead' (2019) 1(5) Nature Machine Intelligence p. 206, 207.

319 ibid 207.

tions. They point out that explanations may be generated by unjustified counterfactuals which are disconnected from the ground truth.[320] Similarly, Lakkaraju and Bastani have highlighted how explainable models may offer predictions that are highly faithful to the original black box models and yet hide the fact that they rely on morally problematic factors (such as race and gender) by offering alternative reasons for the decisions.[321]

Lastly, where predictions and common features are utilised by the two models, the explanations may not provide enough details to make sense of what the black box is doing with them.[322] A heatmap indicating that a black box is concentrating on parts of the image does not indicate how that image is used.[323]

Explanations will be liable to mislead users in some cases, either regarding the predictions *per se* or the factors used in generating them, and the knowledge they provide about the operation of the original model is inherently limited. Rather than opening the black box, they merely deliver 'summary statistics' and 'trends in how predictions are related to the features'.[324] For these reasons this work takes the stance that explainable AI techniques do not offer a satisfactory solution to the types of opacity outlined above.

Things are different with interpretable AI techniques. By definition they provide insights into what the relevant device is actually doing. Therefore, they do provide a solution to the opacity problem, although, even here it remains contested how the techniques used for these purposes relate to an epistemological concept of interpretability.[325] Moreover, there is no universal or perfect technique for interpretability. Within a given context there are outstanding questions, such as: how should one select between

320 Laugel and others, 'The Dangers of Post-Hoc Interpretability: Unjustified Counterfactual Explanations' (Twenty-Eighth International Joint Conference on Artificial Intelligence, Macao, China, 8.10.2019-8.16.2019) 2.

321 Lakkaraju and Bastani, '"How Do I Fool You?": Manipulating User Trust via Misleading Black Box Explanations' (AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, USA, 07.02.2020-09.02.2020).

322 Rudin, 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead' (2019) 1(5) Nature Machine Intelligence p. 206, 208.

323 ibid 208.

324 ibid 208.

325 Krishnan, 'Against Interpretability' (2020) 33(3) Philosophy & Technology p. 487, 492-493.

them in a given context? When is one aiming for sparsity, monotonicity, decomposability or other factors?[326]

What's more, there are considerable challenges in the adequate technological realisation of interpretable models. Even their proponents acknowledge that it is often easier to construct accurate black box models, than to develop comparable interpretable ones.[327] This means that, although perhaps not inevitable, there are still technological limits on the knowledge that can be generated about high-performing ML models. For instance, the aforementioned association of neurons in DNNs with individual concepts is not comprehensive: residual neurons are needed within the relevant layer to deal with uncategorised information, insights into other layers will be limited and opacity around the way in which the concepts are combined persists.[328] More generally, there are serious challenges to the satisfaction of user-defined interpretability constraints.[329]

Perhaps most relevant for the present work is the fact that such interpretable techniques are still not required, or sufficiently ubiquitous, in medical devices.[330] Regulators should bear in mind that the right pressure on developers may lead to accurate, interpretable solutions. However, for the foreseeable future, users of many ML devices will be dealing with technologies that, although desirable on other grounds, and although purportedly offering explanations, remain opaque in the outlined manner.

### *IV. Human-AI collaboration in the healthcare environment*

Equipped with an understanding of ML technology, having seen the capabilities ML devices possess, and also anticipating the difficulties surrounding AI interpretability, we are now in a position to describe how humans and

326 Rudin and Ustun, 'Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice' (2018) 48(5) Interfaces p. 449, 450. These authors ask: 'What are the desired characteristics of an interpretable model, if one exists?'.

327 ibid 450.

328 Rudin and others, 'Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges' (2022) 16 Statistics Surveys p. 1, 33-35.

329 Rudin and Ustun, 'Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice' (2018) 48(5) Interfaces p. 449, 451.

330 ibid 449. The authors hypothesise that this is partly due to a misaligned incentive, whereby the black box nature of models shields developers from accountability.

AI are likely to work together when the technology is implemented in the healthcare environment.

The point of departure for this section, and the assumption underlying much of the previous analysis, is the need for AI to collaborate with human experts. Although ML techniques are demonstrating that they can execute demanding cognitive tasks at scale, quickly and effectively, they lack the requisite contextual understanding and empathy with the individual circumstances of the patient that are so central to medical decision making.[331] Moreover, as will be discussed in the next chapter, there is an important difference between the trust that is placed in a human professional and the reliance placed upon a technical tool. Amongst other things, patients can and do assume that a physician is representing and acting in their interests. No such assurances can be made regarding AI.

For these reasons it is said that human actors and ML devices have different strengths that may complement each other to achieve optimal outcomes.[332] Three features that frame this collaborative effort are picked out here: the nature of the choices made by healthcare actors when they use AI, the knowledge that such actors must have to sensibly utilise ML technology and the potential for the device to influence the human element of the interaction.

## A. Choices in the use of ML devices

Clinicians and hospitals may choose whether to use AI and which AI to use for a given kind of task,[333] although both choices will be influenced by the institutional setting, such as a system's reimbursement framework. Once this choice has been made, a further contextual decision will often be necessary in relation to a particular patient. Namely, whether to use

331 Holley and Becker, *AI-First Healthcare: AI Applications in the Business and Clinical Management of Health* (2021), 58-61.

332 The idea between man-computer symbiosis is well-established and its implications are now being worked out in the healthcare sphere: see ibid 49-71.

333 Providing case studies from the U.S. context see: Lebovitz, Levina and Lifshitz-Assaf, 'Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What' (2021) 45(3) MIS Quarterly p. 1501. It is also the assumption running through: Nix, Onisiforou and Painter, 'Understanding Healthcare Workers' Confidence in AI' (2022) <https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning/understanding-healthcare-workers-confidence-in-ai> accessed 11.11.2022.

the AI in that specific clinical encounter and how to use it: for what purposes, in preference over which other techniques, in combination with what other sources of information and with what degree of confidence its results should be accepted and acted upon.[334]

From the perspective of professional-AI cooperation it is particular this latter dimension that is significant. Where an AI complements human expertise there must still be a contextual decision as to how to treat the AI's results. For example, even where an AI does not necessarily offer an accurate diagnosis for a given decision, it may provide an efficient prompt for the kind of analysis that is likely to lead to a better diagnosis (if properly integrated into the clinical workflow and considered as one piece of evidence alongside others).[335]

Things may be slightly different where devices partially replace cognitive expertise or determine dimensions of the clinical decision-making process. The very choice to rely on the AI for a task – e.g. as a tool for triaging or as a second reader in radiology – may constitute a standardised solution that is applied without individual consideration of patients. Nonetheless, there will still be an operational decision to introduce the patient to this system and a localised assessment of how to act after the patient has been processed by the AI device. For instance, if an emergency department triages chest X-rays as a matter of course and the patient's chest X-ray is classified as non-urgent, and the AI is known to be relatively reliable in this regard, then this may shape the assessment of the patient's condition by the doctor that makes the post-triage decisions.

The key point is that it is a pervasive choice whether and how to divide one's specialist labour with an AI tool. In the final instance this will involve a human decision that is granular, having to be specified for the circumstances of a particular patient, conducted against the backdrop of operational decisions that were made at a more abstract level.

334 'During clinical decision making, clinicians should determine appropriate confidence in AI-derived information and balance this with other sources of clinical information': ibid.

335 Lebovitz and others provide the example of Bone Age and Brain Tumor Segmentation tools, which were explored as a way of improving diagnostic processes by professionals and managers in spite of known flaws in their outputs: Lebovitz, Levina and Lifshitz-Assaf, 'Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What' (2021) 45(3) MIS Quarterly p. 1501, 1512.

## B. User knowledge of ML devices

In Section III.A. above it has been argued that, relative to the physician and patient, current ML models will be opaque or uninterpretable. Specifically in the sense that the reasoning process of the AI, the range of factors it considers and the goals that it pursues will be relatively inaccessible. By contrast, given the collaborative context assumed here and the role that medical professionals play in imparting information to patients, the focus of this section is the knowledge that a healthcare provider will *possess*, since it is a prerequisite for a realistic use of the technology.

There can be little doubt that some such knowledge is necessary to make the choices outlined in the previous section. A recent policy document published by the UK's NHS AI Lab and Health Education England highlights that: 'clinicians will need to understand when AI-derived information should and should not be relied upon, and how to modify their decision making process to accommodate and best utilise this information' and they will need to be in a position 'to confidently evaluate, adopt and use AI'.[336] In the U.S. context Lebovitz and others present a nuanced evaluation of the different kinds of information that the managers and professionals of a hospital had to possess (and partially generate) in order to use AI effectively in their practice.[337] Moreover, forceful cases are being made for the integration of AI literacy courses into the curriculum of medical students.[338] These assessments support the position advocated earlier: understanding AI performance is a complex task, requiring knowledge and skills that are not readily transferable from those exercised in the evaluation of existing technologies by healthcare providers.

To decide how much confidence to place in an ML device, a professional user must be assumed to have an awareness of the broad tasks that an AI is

336 Nix, Onisiforou and Painter, 'Understanding Healthcare Workers' Confidence in AI' (2022) <https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning/understanding-healthcare-workers-confidence-in-ai> accessed 11.11.2022.

337 Lebovitz, Levina and Lifshitz-Assaf, 'Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What' (2021) 45(3) MIS Quarterly p. 1501.

338 See: Wood, Ange and Miller, 'Are We Ready to Integrate Artificial Intelligence Literacy into Medical School Curriculum: Students and Faculty Survey' (2021) 8 Journal of Medical Education and Curricular Development 1-5; Ngo, Nguyen and van Sonnenberg, 'The Cases for and against Artificial Intelligence in the Medical School Curriculum' (2022) 4(5) Radiology: Artificial intelligence p. 1.

designed to accomplish and have access to some type of performance evaluation that measures its suitability for this task, as highlighted in Section 1.C.3.

Precisely what further information is needed will depend on a range of contextual factors. A professional may be able to place partial reliance on research studies and official guidance as to the quality and applicability of the device to a given context.[339] As was seen however, AI may also capture aspects of medical know-how that are contested and subjective and they may be useful only for a relatively specific clinical environment. This places an additional burden on health professionals using the AI: they must be aware of these limitations and they ought to critically evaluate how they can employ these tools in the context of their own experience, the applicable professional standards and other clinical evidence.

Towards this end, a basic understanding of AI model design, data collection and processing, as well as validation will be essential. For example, understanding that a supervised learning model functions on the basis of human designed labels will direct attention towards the individuals who have labelled the data and the assumptions that they have made.[340] This is the foundation for a realistic assessment of the purposes, anticipated benefits and risks that are posed by the tool,[341] as well as allowing the doctor to relativize some of these influences through their own input.

Engaging with the limitations of performance evaluations should further point practitioners towards more general features of AI, such as their relative independence and the ambiguity of the goals that they pursue. The selection of 'good' embryos for implantation, or the ranking of 'best' treatment options requires a great deal of careful analysis. The nature of these labels and how they can be combined with human preconceptions must (almost inevitably) be considered in AI deployment.

339 'Several standards and tools have or are being developed for medical devices and clinical research to guide approaches to the evaluation of AI products, including the National Institute for Health and Care Excellence (NICE) evidence standards framework.': Nix, Onisiforou and Painter, 'Understanding Healthcare Workers' Confidence in AI' (2022) <https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning/understanding-healthcare-workers-confidence-in-ai> accessed 11.11.2022.

340 Lebovitz, Levina and Lifshitz-Assaf, 'Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What' (2021) 45(3) MIS Quarterly p. 1501, 1509-1510.

341 ibid 1515.

Moreover, the fact that a relatively unique engagement with external studies is required is itself an outgrowth of the AI's novelty and opacity. The physician must have an awareness of the direct inaccessibility of conceptual features – as well as the reasoning processes – that the machine has relied upon and they must see that AI use gives rise to pervasive forms of uncertainty. Once again this can be expected to frame a purposive collaboration with the tool and the assessment of its risks and benefits – requiring reference to independent human expertise (whether one's own or that of a colleague) or to objectively ascertainable clinical indicators.

### C. ML influence

Leading on from the previous discussion, the professional ought to be aware of the capability of ML technology to shape the human element of the collaborative decision through cognitive biases. Such biases are a pervasive element in human action and it is well-documented that healthcare professionals are susceptible to heuristic replacements in response to automated decision aids.[342]

Several different kinds of biases have been deemed relevant to the deployment of medical AI,[343] but a consideration of automation bias (AB) serves to highlight the evidence that ML will shape the judgement of clinical professionals even under a collaborative framework.

AB can be defined as the tendency of humans to overly rely on automation.[344] It can lead to 'errors resulting from the use of automated cues as

342 In this sense it contrasts with other areas where AI may be deployed: Alon-Barkat and Busuioc, 'Human-AI Interactions in Public Sector Decision-Making' (2023) 33(1) Journal of Public Administration Research and Theory p. 153, 165.

343 A non-exhaustive list includes: automation bias, aversion bias, alert fatigue, confirmation bias and rejection bias: Nix, Onisiforou and Painter, 'Understanding Healthcare Workers' Confidence in AI' (2022) <https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning/understanding-healthcare-workers-confidence-in-ai> accessed 11.11.2022.

344 Although the term is often used alongside related concepts like 'automation-induced complacency', it has emerged as a central focus for discussions of automation misuse in healthcare. See: Parasuraman and Manzey, 'Complacency and Bias in Human Use of Automation: An Attentional Integration' (2010) 52(3) Human Factors p. 381, 394-395; Goddard, Roudsari and Wyatt, 'Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators' (2012) 19(1) Journal of the American Medical Informatics Association p. 121; Lyell and Coiera, 'Automation Bias and

a heuristic replacement for vigilant information seeking and processing'.[345] Although the phenomenon remains underexplored,[346] Lyell and Coiera state the general problem in this way:

> When it performs well, automation can reduce errors and improve decision performance. It also, however, has the potential to introduce *new types of errors*. One particularly significant risk is that users may become overreliant on automation, especially when a [clinical decision support system] tool is less than perfectly accurate or reliable, leading to decision errors[347]

Taking such mental shortcuts is especially ill-suited to the task of arriving at accurate decisions with the novel and singular ML technology.[348] It was seen to be exceptionally difficult to calibrate one's confidence in the outputs of ML models, even in light of an in-depth critical analysis. How, then, would such calibration be achieved in an abbreviated form? Moreover, AI decision making introduces controversial value judgments and relies on an independent human element to provide context and to tailor a decision to the specific needs of a situation. A biased engagement with an ML device would not only lead to an inaccurate decision, but to one that is missing the broader human component.

---

Verification Complexity: A Systematic Review' (2017) 24(2) Journal of the American Medical Informatics Association p. 423.

345 See: Mosier and others, 'Automation Bias: Decision Making and Performance in High-Tech Cockpits' (1998) 8(1) The International Journal of Aviation Psychology p. 47; Goddard, Roudsari and Wyatt, 'Automation Bias' (2012) 19(1) Journal of the American Medical Informatics Association p. 121, 121; Parasuraman and Manzey, 'Complacency and Bias in Human Use of Automation' (2010) 52(3) Human Factors p. 381, 391: Lyell and Coiera, 'Automation Bias and Verification Complexity' (2017) 24(2) Journal of the American Medical Informatics Association p. 423, 423.

346 Goddard, Roudsari and Wyatt, 'Automation Bias' (2012) 19(1) Journal of the American Medical Informatics Association p. 121; Lyell and Coiera, 'Automation Bias and Verification Complexity' (2017) 24(2) Journal of the American Medical Informatics Association p. 423; Schemmer and others, 'On the Influence of Explainable AI on Automation Bias: Research in Progress' (19.4.2022) <https://arxiv.org/pdf/2204.08859> accessed 6.6.2022.

347 Lyell and Coiera, 'Automation Bias and Verification Complexity' (2017) 24(2) Journal of the American Medical Informatics Association p. 423, 423 (emphasis added).

348 It is notable how seriously this issue was considered in Nix, Onisiforou and Painter, 'Understanding Healthcare Workers' Confidence in AI' (2022) <https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning/understanding-healthcare-workers-confidence-in-ai> accessed 11.11.2022.

The significance of the problem should therefore be clear. However, on the basis of existing studies it is difficult to assess its precise scope. On the one hand, errors associated with AB appear more prone to occur in the medical sphere. This has been associated with the complexity of decision-making and the large amounts information, ambiguity and detail that characterise it.[349] As a result of these circumstances, biases arise even in single-task environments (i.e. where the human user only performs one task concurrently).[350] For example, in the assisted detection of abnormalities in mammography scans it has been found that in the absence of a prompt, physicians were less likely to classify abnormal cases correctly.[351]

On the other hand, there are many case-specific factors that will co-determine the influence of automated tools, including: the intensity of the workload,[352] the skill of the user, and the difficulty of the cases.[353] The generally higher skills of medical professionals, and their ability to synthesise many different forms of knowledge,[354] may suggest a more limited role for AB.

More specifically for AI, the technology's nature may also be argued to increase the likelihood that there is an undue dependence on automation. One reason for this is the task complexity for which the devices are designed. A distinguishing feature of ML algorithms is the capacity to accomplish cognitive tasks that require sophisticated capabilities. It is likely therefore that they will function in environments that will induce AB, even if they focus on a single task – as many of the examples outlined in Section II. do.

The second, related, reason why AB is hypothesised to occur more frequently is that some uses of ML were seen to lower the levels of human

349 Lyell and Coiera, 'Automation Bias and Verification Complexity' (2017) 24(2) Journal of the American Medical Informatics Association p. 423, 429.

350 ibid 426.

351 Alberdi and others, 'Effects of Incorrect Computer-Aided Detection (Cad) Output on Human Decision-Making in Mammography' (2004) 11(8) Academic Radiology p. 909.

352 Goddard, Roudsari and Wyatt, 'Automation Bias' (2012) 19(1) Journal of the American Medical Informatics Association p. 121, 124-125.

353 Povyakalo and others, 'How to Discriminate Between Computer-Aided and Computer-Hindered Decisions: A Case Study in Mammography' (2013) 33(1) Medical Decision Making p. 98, 106.

354 Funer, 'The Deception of Certainty' (2022) 25(2) Medicine, Health Care and Philosophy p. 167, 169.

expertise brought to bear on a case. This has been found to increase the vulnerability to AB.[355]

Thirdly, a greater reliance is likely to be placed on ML systems. The very fact that humans may attribute more sophisticated capabilities to them than previous automatic systems, and that the performance of many will be advertised as having ultra-high accuracy and/or reliability, may be a cause for AB.[356] This is because it has been found that a greater accuracy may engender reliance or trust and trusting users are less likely to detect failures of machines.[357]

Explainable AI techniques may also feed into this misplaced reliance. It has been argued that 'the sole existence of an explanation could increase the reliance of the human on the AI which increases AB'.[358] Furthermore, it has been seen that explanations may be targeted toward inducing certain behaviours in human users, such as placing confidence in the AI, rather than truly offering an interpretation of considered factors and applicable forms of reasoning.

Similar issues will no doubt arise from the design of user interfaces for ML devices. Framing treatment options to indicate the levels of confidence that an AI has in different predictions, or offering an AI output as a default option, will aim to bring about human responses to the technology that are deemed desirable. Presenting such information, and the form of that presentation, will itself contribute to ML devices' potential for inducing cognitive biases and influencing judgments, going even beyond AB.

With the significance and probability of ML's potential to induce cognitive biases thus stated, it can be understood as an important piece of information about AI functioning that medical professionals must consider in reaching collaborative judgments.[359] A healthcare provider should be

355 Povyakalo and others, 'How to Discriminate Between Computer-Aided and Computer-Hindered Decisions' (2013) 33(1) Medical Decision Making p. 98, 106.

356 Sujan and others, 'Human Factors Challenges for the Safe Use of Artificial Intelligence in Patient Care' (2019) 26(1) BMJ Health & Care Informatics p. 1, 3.

357 Lyell and Coiera, 'Automation Bias and Verification Complexity' (2017) 24(2) Journal of the American Medical Informatics Association p. 423, 424.

358 Schemmer and others, 'On the Influence of Explainable AI on Automation Bias', (19.4.2022) <https://arxiv.org/pdf/2204.08859> accessed 6.6.2022.

359 'The propensity towards these biases may be affected by choices made about the point of integration of AI information into the decision making workflow, or the way such information is presented. Interviewees for this research highlighted that enabling clinicians to recognise their inherent biases, and understand how these affect their use of AI-derived information should be a key focus of related training and

aware of AI-induced biases and they must attempt to assess and mitigate their impact on decision making.[360]

## V. Conclusion

The aim of this chapter was to provide the empirical, technical characteristics that are necessary for an assessment of AI's autonomy-related challenges in the following work. Stated succinctly the dimensions that should be emphasised going forward for these purposes are: (1) the choices represented by the introduction of new human-machine interactions that provide alternatives to the *status quo*: human specialists, (2) the variable quality of these interactions, including different degrees of automation, (3) the deficiencies in the performance evaluation of clinical AI, especially where deployed for variable groups and/or environments, (4) a lack of understanding about AI functioning, which will remain in spite of emerging technological solutions, (5) the fact that AI use will involve choices for healthcare providers in individual clinical encounters and that these must normally be able to regulate their cooperative decision making with AI on the basis of their broader knowledge of the technology and how it functions, and (6) the ML devices' influence on this cooperative interaction.

education. Failure to do so may lead to unnecessary clinical risk or the diminished patient benefit from AI technologies in healthcare.': Nix, Onisiforou and Painter, 'Understanding Healthcare Workers' Confidence in AI' (2022) <https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning/understanding-healthcare-workers-confidence-in-ai> accessed 11.11.2022.

360 How far this and other strategies can serve to combat AB is admittedly contested: Lyell and Coiera, 'Automation Bias and Verification Complexity' (2017) 24(2) Journal of the American Medical Informatics Association p. 423, 429-430.

# Chapter 3: Bioethical autonomy and artificial intelligence

This chapter investigates how the application of artificial intelligence (AI) may undermine the adherence to the value of autonomy in medicine. By combining an understanding of the factual background with a bioethical analysis, the aim is to identify those challenges that represent comparable problems across the UK and the U.S. jurisdictions. As outlined in Chapter 1, for the purposes of the legal comparison underlying this work, autonomy challenges stemming from AI constitute the *tertium comparationis*: a comparable pre-legal problem between legal systems to which each may offer a different response.

An appropriate definition of autonomy is central to the establishment of the comparison. If the definition is drawn too narrowly, then there is a risk that diverging conceptions of that value, which can be found reflected in the law, could be excluded *ex ante*. If the definition is not detailed enough however, then one risks missing or underconceptualising the core challenges posed by medical machine learning (ML).

The way to strike this balance is to recognise an important facet of the common law's private law reasoning that is the subject of both of our jurisdictional analyses: within limitations this reasoning is policy-based and aspirational.[361] Although national specificities must be respected, there is often an (implicit) assumption that judges are involved in a collective process of reasoning in line with recognised principles.[362] The nature of these principles will be unspecified to a certain degree and contested – allowing for comparable argumentation regarding their nature.[363] Much more of this will be made in Part II.

For present purposes this means that a conception of autonomy will be developed that is initially external to the law, lends itself to an identification of AI problems and yet can be inserted into its argumentative structure. Section I. develops this understanding of autonomy, it's broad nature. Section II. draws on elements of this theory to frame the autonomy challenges posed by AI as going towards: certain key beliefs about the goals being

361 Robertson in Robertson and Tang, *The Goals of Private Law* (2009).
362 Duxbury, 'The Law of the Land' (2015) 78(1) The Modern Law Review p. 26, 47-48.
363 Sunstein, *Legal Reasoning and Political Conflict* (1998) 35-61.

pursued in the patient's care, the changing expertise of human users, the risks related to AI and the technology's potential for manipulation.

### *I. The procedural conception of autonomy*

It is a trite observation that autonomy derives from the terms of self (*autos*) and law (*nomos*). It is also widely recognised that what this entails is controversial. Broad distinctions have been drawn between variously defined theories of autonomy, such as 'libertarian', 'liberal' and 'communitarian'.[364]

In the practice-oriented field of bioethics there have been attempts to reconcile these different approaches to arrive at a workable middle path. Most notably this is purportedly done by the widely operationalised theory of principlism developed by Beauchamp and Childress.[365] These authors outline a principle according to which autonomy is 'self-rule that is free from both controlling interference by others and from limitations, such as inadequate understanding, that prevent meaningful choice'.[366] The principle is thereby elaborated at a level of abstraction that allows diverging specifications of a core normative content and it is often described as a middle-level guide to action that can yield more specific rules in specific scenarios.[367]

While this theory therefore points towards two potential sources of autonomy violations, control of the patient and their limited understanding, it does not offer very much by way of specification. One is not provided with the kind of general grounds that determine which controlling interferences are permissible and which are not, or which deficiencies of understanding render it inadequate.[368] Even if one can agree that violations are related to interferences and limitations this does not take one that much further, since it is generally recognised that humans are always

364 Maclean, *Autonomy, Informed Consent and Medical Law: A Relational Challenge* (2009) 11-17.

365 Beauchamp and Childress, *Principles of Biomedical Ethics* (Fifth Edition 2001).

366 ibid 58-59.

367 Wolf, 'Shifting Paradigms in Bioethics and Health Law: The Rise of a New Pragmatism' (1994) 20(4) American Journal of Law & Medicine p. 395, 400; that Principlism's distinction between rules and principles does not follow a Dworkinian model, but is merely a matter of abstractness has also been noted by Paulo, *The Confluence of Philosophy and Law in Applied Ethics* (2016) 119.

368 Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 11.

subjected to some forms of entirely legitimate limitation and interference.[369] This account does not enable us to deal with the central issue at hand: to identify *which* AI-influences or AI-related limitations are challenging autonomy and *why*.

To offer general grounds that help with this identification, this chapter will utilise a procedural and, specifically, rationalist account of bioethical autonomy that has been developed by Johnathan Pugh.[370] Hereby the autonomy of an individual depends on the way in which they came to make their decision.[371] It presupposes that beliefs and actions flow from the values that individuals hold *and* from a certain responsiveness to reasons.[372]

The strength of this account stems especially from the fruitful way in which it allows one to conceptualise the challenges raised by AI in medicine and from the fact that it can be integrated into the legal doctrines of the selected jurisdictions, as described in Part II. Below it will be seen that the use of AI in medicine concerns the insertion of a device into the deliberative process, changing the way in which the doctor and ultimately the patient, arrive at their decision. What is problematic about such an insertion, is the sense that it upsets the deliberative-facilitative process of doctor-patient decision making and that it undermines the ability to act on the basis of reasons.

Pugh begins his work with the declared aim of elucidating 'the nature and forms of influence that can subvert autonomy'.[373] By distinguishing

369 This is recognised by relational accounts of autonomy, which take the social-embeddedness of individuals and their autonomy as a central premise: Stoljar, 'Informed Consent and Relational Conceptions of Autonomy' (2011) 36(4) The Journal of Medicine and Philosophy p. 375, 376.

370 Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020). Pugh is building on an existing strand of thought that has developed in bioethics, see: Savulescu in Rhodes, Francis and Silvers, *The Blackwell Guide to Medical Ethics* (2008); Ploug and Holm, 'Doctors, Patients, and Nudging in the Clinical Context--Four Views on Nudging and Informed Consent' (2015) 15(10) The American Journal of Bioethics p. 28.

371 Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 5. Fischer and Ravizza develop a similar approach for moral responsibility and highlight that their account focuses on 'the characteristics of the *actual sequence that leads to the action*': Fischer and Ravizza, *Responsibility and Control: A Theory of Moral Responsibility* (2000) 37.

372 Again there are established theories that call for a similar responsiveness outside of the bioethical context: Wolf, *Freedom Within Reason* (1993); Fischer and Ravizza, *Responsibility and Control: A Theory of Moral Responsibility* (2000).

373 Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 3.

between a decisional dimension of autonomy and a practical dimension of autonomy, he is able to provide interlocking, general reasons for distinguishing interferences that are legitimate from those that constitute autonomy violations.

## A. Decisional autonomy

Decisional autonomy relates to the ability to make one's own decisions. It incorporates both a cognitive and a reflective element. Under the cognitive element the agent ought to follow the norms of theoretical rationality, which means arriving at beliefs in ways that are based on appropriate evidence, inductive reasoning, are not inconsistent, etc.[374] These beliefs must then also be placed in the broader context of the agent's other convictions 'about both descriptive and evaluative features of the world'.[375]

The reflective element of decisional autonomy responds to the intuition that autonomy requires one's reasons for action to be one's own.[376] That is, regardless of what is generally accepted or treasured, an individual can shape their own unique system of beliefs and desires. Determining exactly how to identify the relevant, internalised states is not straightforward. We each have many beliefs and desires that occur to us without being endorsed, we may consider them fleetingly or non-seriously. These may have some importance too, but they are not taken to define who we are or what we want. To make decisions truly one's own, there is a philosophical tradition that holds it necessary for an agent to form evaluative judgments about what they have reasons to do.[377] To act autonomously, individuals must

374 Pugh draws on well-established conceptions of practical reasoning. Similar accounts can be found in: Baron, *Rationality and Intelligence* (2005) 90; Fischer and Ravizza, *Responsibility and Control: A Theory of Moral Responsibility* (2000) 71. The latter maintain that individuals must be receptive to what reasons there are, which involves an understandable pattern of (actual and hypothetical) reasons-receptivity.

375 Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 38.

376 Hyun, 'Authentic Values and Individual Autonomy' (2001) 35(2) The Journal of Value Inquiry p. 195, 196. See also: Christman, 'Autonomy and Personal History' (1991) 21(1) Canadian Journal of Philosophy p. 1. A version of this intuition is also expressed by: Frankfurt, 'Freedom of the Will and the Concept of a Person' (1971) 68(1) The Journal of Philosophy p. 5; Dworkin, *The Theory and Practice of Autonomy* (2012).

377 Christman, 'Autonomy and Personal History' (1991) 21(1) Canadian Journal of Philosophy p. 1, 4-6.

reflect upon their 'desires and beliefs, forming attitudes towards them'.[378] This evaluative *reflection* should occur at some point, although it need not be at the point of action[379] and although it may be unconscious.[380]

In essence this means that self-governance requires a hierarchical structuring of beliefs and desires, with some being higher, in the sense that they have lower beliefs or desires as their object.[381] It also means that there is an internal condition of evaluation. Individuals exercise a degree of control in making beliefs and desires their own: they apply personal standards in deciding whether these are true and whether they are good to act upon.[382] Given his rationalist emphasis, Pugh holds that evaluative attitudes, whether about beliefs or about one's understanding of the good, must be theoretically rational and not just aimed at one's own subjective understanding of the truth or the good.[383]

These more targeted beliefs and desires are respectively termed acceptances and preferences.[384] Pugh follows Ekstrom in going further and distinguishing a subset of these artefacts that constitute one's self. Ekstrom develops a coherentist model whereby one authorises certain preferences and acceptances 'when they cohere with one's other preferences and acceptances', they 'hold together firmly, displaying consistency and mutual support'.[385] There must be some judgement as to the value of the belief and action in the context of one's wider character system in order for them to relate to the agent's true self.[386]

---

378 Ekstrom, 'A Coherence Theory of Autonomy' (1993) 53(3) Philosophy and Phenomenological Research p. 599, 599.

379 Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 49, citing Savulescu, 'Rational Desires and the Limitation of Life-Sustaining Treatment' (1994) 8(3) Bioethics p. 191, 199-200.

380 Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 49; Ekstrom, 'A Coherence Theory of Autonomy' (1993) 53(3) Philosophy and Phenomenological Research p. 599, 603.

381 This is inspired by Frankfurt's seminal theory: Frankfurt, 'Freedom of the Will and the Concept of a Person' (1971) 68(1) The Journal of Philosophy p. 5.

382 Ekstrom, 'A Coherence Theory of Autonomy' (1993) 53(3) Philosophy and Phenomenological Research p. 599, 606-607.

383 Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 51-52.

384 Ekstrom, 'A Coherence Theory of Autonomy' (1993) 53(3) Philosophy and Phenomenological Research p. 599, 600.

385 ibid 608.

386 ibid 610-612. Pugh elaborates on the details of this and amends some part of the theory: Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 50-54. His amendment ensures that autonomous action is possible in light of conflicting,

Cohering acceptances and preferences can be identified by reference to their being: long-lasting, defensible (being well-supported with reasons and thus resilient to challenge) and comfortably owned by the individual (they are not conflicted in acting on them).[387] In this manner, one has a rational justification of a given element that is in line with one's wider character. The outcomes are authorised preferences or acceptances.[388] This will be used as a shorthand for desires and beliefs that reflect the agent's self, since they are an outcome of the reflective process.

In addition, Pugh argues that certain beliefs pertaining to a decision need not only be arrived at rationally and upon reflection, but must also be true.[389] These are decisionally necessary true beliefs. If one is mistaken about them, then one cannot arrive at an autonomous decision.[390] The key thought behind this approach is that the content of these beliefs is so central to the relevant choice – so significant for the connection of one's values to one's actions – that, without them, one does not really control one's decision at all.[391] Given this connection to action, the nature of these beliefs will be defined below, with a view to the practical dimension of autonomy.

Overall, these conditions may make it seem like autonomy is difficult to exercise for a normal patient and thus an interest that seldom requires protection. Beauchamp has claimed that 'the conditions of (…) reflective control are so demanding in this theory that either many human actors will be excluded as persons or their actions will be judged nonautonom-

difficult choices where one may act in a less than fully rational manner, but not irrationally. Sub-optimal choices can still cohere with one's character sufficiently.

387 Ekstrom, 'A Coherence Theory of Autonomy' (1993) 53(3) Philosophy and Phenomenological Research p. 599, 608-609.

388 'One's preferences, I suggest, are authorised-or sanctioned as one's own-when they cohere with one's other preferences and acceptances': ibid 608.

389 Pugh makes clear that he draws on a long-established intellectual tradition by relating this to the 'Aristotelian claim that actions performed from reasons of ignorance are non-voluntary': Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 131.

390 ibid 35.

391 This is encapsulated in Wolf's claim that: 'an agent cannot have the kind of freedom and control necessary for responsibility unless, when making choices about values and actions, she can understand the significant features of her situation and of the alternatives among which her choice is to be made. That is, an agent cannot be free and responsible unless she can sufficiently see and appreciate the world for what it is': Wolf, *Freedom Within Reason* (1993) 117.

ous'.[392] This may be termed an anti-paternalist concern and, although it is impossible to mount a full defence of Pugh's theory here, it is important to counter this argument because, as will be seen in subsequent chapters, this is a kind of objection that has been raised within the law itself and within academic analyses of the law.[393]

As Beauchamp rightly points out, the outlined theory requires individuals to be able to engage in certain processes of rational evaluation. It would be misguided to deny that such a bar is set. Yet its demandingness is arguably grossly overstated. In particular, the requisite capabilities are not of a kind that is unachievable for those that we would want to describe as autonomous agents.[394] Nor is particular emphasis placed on the intellectual calibre of these agents – the requisite processes may even be engaged entirely unconsciously.[395] If one takes seriously the criterion that autonomy requires self-government, as per the definition with which we started this chapter, then setting conditions that enable this control will be unavoidable. Pugh's theory seems well-placed to strike the balance.

### B. Practical autonomy

Practical autonomy is used to refer to conceptions of positive and negative freedom that constitute distinct prerequisites for effective action.[396] Freedom in this sense is used to refer to some restraint – this can be positive or negative – whereas autonomy is broader and encapsulates the above, deliberative elements.[397] Negative freedom requires that there be no debilitating factor or force preventing an agent from achieving an end that they are motivated to achieve and positive freedom holds that no factor enabling that end is absent (this is taken to refer to capacities, but also to the possession of information, understanding etc.)[398]

---

392 Beauchamp, 'The Failure of Theories of Personhood' (1999) 9(4) Kennedy Institute of Ethics journal p. 309, 313.

393 Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 199.

394 ibid 201.

395 ibid 200-201.

396 Pugh is drawing on an established tradition of discussions of freedom/liberty: Berlin and Harris, *Liberty* (Second Edition 2017); MacCallum, 'Negative and Positive Freedom' (1967) 76(3) The Philosophical Review p. 312.

397 Cf. Christman, 'Autonomy and Personal History' (1991) 21(1) Canadian Journal of Philosophy p. 1, 2-4.

398 Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 123-125.

Negative freedom appears relatively self-explanatory and need not be specified further for our purposes. Positive freedom, by contrast, ought to be considered because it points one towards a broader dimension of autonomy and lies beneath the problems to be elaborated in the clinical sphere. Namely, respecting another's autonomy sometimes entails assisting another, enabling them to achieve their ends. This may involve education, dialogue and decision aids, as well as an expansion of options and/or information about those options.[399]

One may also note the interconnectedness with theoretical rationality here; adhering to or enhancing this dimension also makes the realisation of practical autonomy more likely, being conducive of acting in ways that are likely to further the pursuit of one's ends, even if it does not guarantee success.[400] Significantly, one can see that the facilitation of understanding depends on putting the patient in a position where they are able to grasp the relation of inputs to the realisation of their own values.[401]

Furthermore, returning to the notion of decisionally necessary true beliefs, positive freedom requires that, for certain discrete categories of beliefs, the agent's understanding must map onto the true nature of the world.[402] In Parfit's terminology, utilised by Pugh, certain apparent reasons must be real reasons.[403] These categories of reasons are vital for the realisation of one's ends, so that they must not only be arrived at rationally, but they must actually be true.

To flesh out what these categories are Pugh draws on a modal test. Hereby he asks whether the agent could hold a relevant false belief and still achieve their desire's objective in a relevantly similar situation (a 'nearby possible world').[404] If they can, then the belief is not decisionally necessary, if they cannot, then it is decisionally necessary. An example that is given is 'the fact that an intervention will be painful or invasive'.[405] It is implied that without being aware that their treatment involves such factors the patient

399 ibid 144-145.
400 ibid 22.
401 ibid 134-135, 157.
402 For a similar, but by no means identical account (given that Pugh merely refers to true belief rather than knowledge) see: Mueller, 'The Knowledge Norm of Apt Practical Reasoning' (2021) 199(1-2) Synthese p. 5395. This provides an overview on the debate surrounding a knowledge criterion.
403 Parfit, *On What Matters: Volume One* (2011) 111; Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 25-26.
404 Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 128-136.
405 ibid 173.

will not be able to achieve their goals in their care. Anticipating a discussion that will be elaborated upon in Section II., one could also cite the example of information that concerns the fundamental purpose of a procedure, such as the conditions that are to be ascertained *via* a genetic test. Without such information the patient cannot align the goals that are pursued by the procedure with their own desires and beliefs.

The precise nature of the modal test, its validity and implications, need not be elaborated upon for our purposes. Rather, it is taken to illustrate the differing levels of significance that can be attached to information and that a particularly significant interference occurs where a lack of understanding altogether precludes a patient from aligning their care with their objectives. Contrast this with a situation where a patient decides to pursue a certain intervention, but does not know the future state of the world (whether a risk will eventuate or not).[406] This does not preclude effective action. Rather, making such judgment calls may be seen as a condition of practical agency.[407]

All in all, the notion that the absence of important beliefs may sever the connection between the agent's situation and the pursuit of their desired ends, is a fruitful one for identifying particularly severe autonomy violations, which may demand a unique institutional response. It does not, however, require a patient to conform their decision making to an overarching objective framework. It must always be borne in mind that the kinds of necessary beliefs are limited both in number and scope. Under the reflective dimension the pride of place in the balancing of reasons and interests must be left to the individual patient.

### C. Summation

In sum, AI's challenges will be conceptualised as affecting one or more of these dimensions. In the course of this I will draw on the kinds of considerations adduced under Pugh's approach. These include a cognitive dimension of autonomy – which posits a decisional process where a minimal degree of theoretical rationality and an agent's reflective capabilities are brought to bear – as well as a practical dimension, which demands that

406 ibid 132.

407 Rid and Wendler, 'Risk-Benefit Assessment in Medical Research – Critical Review and Open Questions' (2010) 9(3-4) Law, Probability and Risk p. 151, 154.

the patient's decision-making process is not interfered with and, above all, that this process is also adequately facilitated by a suitable informational environment.

### *II. The challenges posed by clinical AI to procedural autonomy*

On the above account patient autonomy demands that the patient be able to adequately reason about their clinical choices, so that these can be aligned with their own desires and goals. There are several ways in which the embedding of AI into clinical decision making can cause this endeavour to go wrong.

AI will sometimes constitute an external (non-agential)[408] influence that undermines the cognitive aspect of the patient's decisional autonomy. Most clearly this is the case where its use causes a patient to fail to hold a decisionally necessary true belief, but it may also happen where that individual is led to sustain theoretically irrational beliefs. Both of these may be referred to as forms of informational manipulation.[409]

ML technology may also affect the practical dimension of autonomy. One must distinguish here between an interference with negative and positive freedom. The former is limited where one is restricted from pursuing the end that one has decided to pursue. Ordinarily the functioning of AI cannot be expected to undermine this dimension: it is a supporting element in a decision-making process that does not have the ability to restrain autonomous action.[410] Where a patient's negative freedom may become relevant, is where AI use is made non-optional by a medical professional. Yet this purported autonomy interference must be located in the wider context

408 Pugh is clear that his account differs from Principlism in that non-agential forces can violate autonomy: Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 61. This is a further reason why the rationalist conception of autonomy is well-suited to identify the challenges posed by narrow AI.

409 ibid 60. I do not address another form of manipulation brought up by Pugh, psychological manipulation, as this occurs where the patient's desires are changed without appealing to their cognition. AI are less likely to bring about such changes, as they engage only indirectly with patients' ends, especially in the context of physician mediation. For example, automation bias does not change the desires of the patient.

410 I hold the same to be true in relation to coercion: AI technology is not designed to issue coercive threats and the possibility of AI-mediated use having the features of coercion does not warrant discussion in the abstract.

of medical practice, which is premised upon offering individuals a highly restricted choice-set that is shaped by the importance of other values.[411] Without more, offering medical treatment that is conditional on AI use is unlikely to constitute a significant autonomy violation and will not be considered further here.

Things are different with the dimension of positive freedom. Recall that such freedom is limited where one is not provided with the enabling factors to act in pursuit of one's goals. It focuses our attention on creating those conditions of decisional autonomy in AI use that allow for effective individual action. For example, while one may restrict a patient's negative freedom by denying them access to a non-AI alternative, one may still have an obligation to promote their awareness of the existence of non-AI alternatives. Indeed, an individual's autonomy may be substantially undermined where there are relevant differences between these alternatives and an individual is not put in the position to make a meaningful choice amongst them.

With these broad categories of interreference in mind, three challenges will be focussed on: causing a patient to fail to hold decisionally necessary true beliefs about goal-directed AI action, failing to assist the patient's in forming beliefs about the nature of AI-human cooperation and failing to facilitate the patient's understanding of AI devices, including their general risk characteristics and their influence on decision making.

### A. The need to form true beliefs about AI's goal-directed action

Drawing on the insights of the last chapter we know that, when AI is inserted into medical decision making, it influences that process. Sometimes this influence has implications for the patient's pursuit of their plans and policies. Specifically, one way in which AI threatens to undermine autonomy stems from its ability to pursue goals relatively independently; if not independently from the designer, then at any rate independently from the user and subject.

This has been discussed in the bioethical literature in relation to Watson for Oncology, an AI similar to our case study of the AI-Pathway Companion Prostate Cancer. Both devices draw on ML techniques to complement

411 Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 140.

human expertise, providing an additional resource for decision-making. Moreover, Watson analyses the patient's condition and proposes tailored treatment options, ranking them in different colours to indicate their desirability, which appear to be related to the chances of achieving 'disease-free survival'.[412]

One commentator, McDougall, has argued that because this end is not tailored to the individual patient, since it drives the treatment decision and since there is no encouragement to reflect on the value-laden nature of decisions, AI like Watson for Oncology violate patient autonomy.[413] This violation is said to be incurable by physician mediation because '[r]espect for patient autonomy means that patients' values should drive the ranking process. The patient's own values should be overtly shaping treatment decision making as a primary parameter, not a secondary consideration'.[414]

This account arguably overstates the significance of AI autonomy violations, missing the key situations in which the technology's goal-directed action becomes challenging. The rationalist approach to autonomy provides a more nuanced explanation of relevant violations, one that fits better with our understanding of the technology and with the rationale behind the legal protections of patient autonomy.[415]

Before critiquing her position, we can note that McDougall is surely correct in the claim that many medical devices based on ML do not tailor their recommendations to patient values. As such, they potentially pursue ends that diverge from those that some patients may wish to realise in their care.[416]

One cannot deny that this technology will, almost by necessity, have to have recourse to financial considerations. At a minimum, to be useful, a recommendation must accord with the treatments that the particular

412 Di Nucci, Jensen and Tupasela, 'Ethics of Medical AI' (5.12.2019) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3432317> accessed 5.4.2020.

413 McDougall, 'Computer Knows Best?: The Need for Value-Flexibility in Medical AI' [2019](45) Journal of Medical Ethics p. 156, 157-158.

414 McDougall, 'Computer Knows Best?' [2019](45) Journal of Medical Ethics p. 156, 158.

415 One relatively brief attempt has already been made to utilise Pugh's theory in this respect: Debrabander and Mertes, 'Watson, Autonomy and Value Flexibility' [2021] Journal of Medical Ethics p. 1043.

416 Bjerring and Busch highlight the conceptual issues of sensibly encoding preferences and values into artificial systems: Bjerring and Busch, 'Artificial Intelligence and Patient-Centered Decision-Making' (2021) 34(2) Philosophy & Technology p. 349, 360.

health system reimburses.[417] For many ML devices an additional goal may be to increase resource conservation and efficiency. At its most extreme, the designers of a particular AI may covertly seek to shape its reasoning to further their own financial interests.[418] Moreover, ML models may also weigh considerations that, while not financial, are nevertheless extrinsic to the patient's medical need. For instance, where an AI is online (i.e. continuously learning) a need to improve performance may influence how, and indeed what, decisions are made. The aim of furthering the 'research' that advances the AI would then become an operative goal in the treatment of the patient.[419]

Even here, however, the claim must be qualified. For some tasks assigned to AI the goals will be so narrowly defined that a greater capacity for independent action does not become relevant.[420] The IDx-DR case study illustrates how AI can be used to pursue a narrowly defined, singular goal: to diagnose diabetic retinopathy on the basis of an image-analysis of the patient's retina. If the patient knows the broad purpose of the procedure (say to check for one detrimental eye condition associated with diabetes), then it is unlikely that there will be a divergence of goals between human and machine. This is so even where there is no further attempt to incorporate patient values.

Admittedly, things are different for AI like Watson or the AI-Pathway Companion Prostate Cancer. They perform a broad array of complex tasks,

417 '[IBM] said that the system can be customized to reflect variations in treatment practices, differences in drug availability and financial considerations': Ross and Swetlitz, 'IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close' (5.9.2017) <https://www.statnews.com/2017/09/05/watson-ibm-cancer/> accessed 28.3.2023.

418 Although not involving AI, a situation like this has already arisen with drug-prescription software: Mann, 'Health Care Software Firm Fined $145M In Opioid Scheme With Drug Companies' (1.2.2020) <https://www.npr.org/2020/02/01/801832788/healthcare-software-firm-fined-145m-in-opioid-scheme-with-drug-companies?t=1615393792393> accessed 10.3.2021.

419 Cf. Grubb and Pearl's distinction between therapeutic and non-therapeutic touching, noting that '[t]he law does not assume that a patient who consents to treatment also consents to research': Grubb and Pearl, *Blood Testing, AIDS, and DNA Profiling: Law and Policy* (1990) 12-13.

420 Funer makes a similar distinction by reference to AI that do and do not depend 'on statistically based, normatively uncontroversial information': Funer, 'Accuracy and Interpretability: Struggling with the Epistemic Foundations of Machine Learning-Generated Medical Information and Their Practical Implications for the Doctor-Patient Relationship' (2022) 35(1) Philosophy & Technology, Article no 5.

involving the analysis of different data sources to visualise the patient's situation and possible treatment options. This visualisation and the presentation of treatment options will inevitably contain ideas about the ends that are to be pursued. As already discussed, this could involve a preference for the pursuit of longevity over quality of life. Here there is real potential for misalignment between the agent's desired ends and the one's that the AI is programmed, or has programmed itself, to pursue.

Whether this potentiality is then actualised depends on further factors that restrict the subset of problematic AI. Most importantly, as de Nucci has pointed out[421] and as we already saw in Chapter 2, it is still unclear whether an assisting AI/ML device can be said to drive a medical decision. No doubt it exerts some influence and capitalises upon some biases,[422] but the intention behind human involvement, especially human expert involvement, is to serve as a check on these influences and it is at least unclear when, or how far, these checks are ineffective.[423]

McDougall's assertion that AI drives medical decisions is a strong generalisation, especially if one uses Watson for Oncology, an AI merely complementing human expertise, as an example. If this generalisation is questioned, then it is plausible that any divergent patient goals can be meaningfully introduced into the subsequent deliberations between healthcare professional and patient. McDougall simply stipulates that autonomy demands that patients' values should drive the ranking process and that they should overtly shape treatment decision making as a primary parameter.

Pugh's autonomy account is intended to move beyond stipulations. Hereunder, AI could only drive clinical decision making in such a strong manner, irreversibly subverting patient values, if it hid its objectives and

421 Di Nucci, 'Should We Be Afraid of Medical AI?' (2019) 45(8) Journal of Medical Ethics p. 556.

422 For instance, Debrabander and Mertes note that Watson taps into the 'order effect': Debrabander and Mertes, 'Watson, Autonomy and Value Flexibility' [2021] Journal of Medical Ethics p. 1043.

423 McDougall appears to acknowledge that the influence of AI is an open question, calling for further empirical assessments: McDougall, 'No We Shouldn't Be Afraid of Medical AI; It Involves Risks and Opportunities' (2019) 45(8) Journal of Medical Ethics p. 559. Similarly Kudina and de Boer acknowledge that 'the line between supporting medical decisions and determining them may be thin if not carefully reflected upon': Kudina and Boer, 'Co-Designing Diagnosis: Towards a Responsible Integration of Machine Learning Decision-Support Systems in Medical Diagnostics' (2021) 27(3) Journal of Evaluation in Clinical Practice p. 529, 533.

thereby induced a failure to hold a decisionally necessary true belief.[424] This would sever the connection between the patient's actions and the goals that they wish to pursue. Under Pugh's modal test, this requires there to be no relevantly similar scenario in which the patient could fail to hold the relevant belief about the AI's objectives and yet achieve their end.

Such a state of affairs seems highly unlikely where a human expert meaningfully mediates AI use for the patient. Where an AI merely complements a human professional, this professional can use their specialised body of knowledge to assess the goals that AI outputs can achieve, as well as the implications of this and what other values and evidence are in play. On this basis they can discursively engage with the patient about their plans, desires and goals. This is arguably suited to generate the kind of *post facto* critical reflection that Pugh alludes to elsewhere.[425] Subsequently it is hard to identify elements of deliberative or practical autonomy that are inescapably undermined by such an AI output. Contrary to McDougall's claim, it is not necessary for a patient to know the AI's goals in order to align their action with their preferences in most circumstances.

Watson's case itself exemplifies the outlined distinction. Even if users do not know that the machine is optimising for life expectancy, the professional should still be able to incorporate the machine's outputs into their wider body of knowledge and identify the options that they judge to promote life expectancy, promote quality of life, etc. and to discuss this with the patient. The patient must then only be able to reflect upon those commitments and to incorporate them into their practical reasoning, choosing to endorse or reject a course of action accordingly.

This is not to say that an agent's autonomy remains entirely unimpaired however. Deliberating reflectively about the utilisation of AI outputs would admittedly be more straightforward if one knew the goals that an AI is striving to accomplish – or if one could exert a degree of control over these goals.[426] The problems associated with weaker forms of AI influence will be explored separately below.

For now it is found that, if one can deliberate in a roundabout way and reach a decision independently from the AI, then neither dimensions of autonomy is fatally undermined. Such general reflection, accompanied

424 In Section II.C.2. below we will discuss the weaker influence that AI can have on decision making.

425 Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 138.

426 This is McDougall's proposed technological solution: McDougall, 'Computer Knows Best?' [2019](45) Journal of Medical Ethics p. 156, 158-159.

by an assisted decision process for the patient, is arguably maintained in many instances where AI complement or partially replace expert cognitive capabilities.

Where an AI's pursuit of goals irredeemably violates autonomy, is where it is not only used to accomplish a task that has no single, well-defined aim, but where this also partially determines an aspect of the medical procedure (i.e. where a device falls under the third class of case studies discussed in Chapter 2). One may refer back to the triaging example of Accipio Ix in this respect. The influence of this ML device's output irreversibly manifests itself as soon as it brings certain cases to the physician's attention above others. A patient or physician cannot counteract this influence by incorporating it into their reasoning because there is no opportunity to reflect on the ends underlying it.[427] Let us exemplify the resulting autonomy violations through two hypothetical examples of ML that pursue a general, or not very well-defined, aim and partially determine a decision.

The first example leans on Accipio Ix. While this AI arguably has a relatively narrow purpose, the identification and prioritisation of acute intracranial haemorrhages, there is significant potential for AI to be used in the triaging of patients more widely, with one NHS Foundation Trust for example already seeking to implement AI to support a general triaging service.[428] The very broad purpose of such devices will be to prioritise different clinical options in light of patients' conditions. Many different goals could be pursued under this head – focusing on the most acute conditions, ensuring an efficient allocation of resources, counteracting potential

427 Of course, in emergency settings autonomy may not even come into play in relation to triaging, as the patient may lack capacity, but it is not clear that this will always be so. Moreover, the point applies to the use of AI in the prioritisation of AI more generally, as the next example shows.

428 E.g. Mahase, 'Birmingham Trust and Babylon Health Discuss Pre-A&E Triage App' (2019) 365(l2354) BMJ (Clinical Research Edition).The Babylon Symptom checker app discussed here appears to rely on some forms of AI for some aspects of its service provisions, but they do not appear as sophisticated as the AI techniques discussed by Marchiori and others, 'Artificial Intelligence Decision Support for Medical Triage' [2020] AMIA Annual Symposium Proceedings p. 793. In the U.S. context see, for example: Johns Hopkins Technology Ventures, 'Digital Health Start-up That Assists Emergency Department Decision Making Acquired' (2022) accessed 17.3.2023.

discrimination, ensuring some targeted access to human contact and care, etc.[429]

Moreover, the rationale for giving AI this role is that they alleviate the intense pressure on a limited workforce.[430] As such, it stands to reason that AI will play a preliminary, largely unsupervised role: engaging the patients in a dialogue, undertaking a personalised assessment of their symptoms, giving them healthcare recommendations and in ensuring a targeted use of telehealth and physical services.[431] This could not only considerably speed up the availability of triaging services in-house, but even make it generally available to patients *via* web- and mobile-applications. Even if there is a backstop of human decision making therefore, the AI initially determines an aspect of the decision for which it will be hard to incorporate patient preferences.

The second example draws on the recognition that making and disclosing certain diagnoses has irreversible impacts on patients' lives (their plans, policies, etc). This is well-illustrated by discussions surrounding autonomy and predictive genetic testing, including the 'right not to know' certain classes of information.[432] Concerns arise particularly from the uncertain scope of many such tests and from the possibility of incidental diagnoses being made.[433] In some circumstances the concern is not only that the patient is unaware of certain risks or dangers (a false positive or negative),

429 '[W]hen we leverage AI to detect people with high potential for chronic disease, we must understand the AI goal. Is the purpose of the AI to make people with that disease process healthier, or is it to reduce costs? In the former case, the AI would prioritize the highest-risk patients to receive healthcare that would intervene in or even prevent development of the disease. In the latter case, the AI would appear discriminatory if it prioritized on the low end those who were at highest risk for poor outcomes, with the result that they received the least amount of healthcare': Holley and Becker, *AI-First Healthcare: AI Applications in the Business and Clinical Management of Health* (2021) 59-60.

430 Kim and others, 'A Data-Driven Artificial Intelligence Model for Remote Triage in the Prehospital Environment' (2018) 13(10) PloS One.

431 Marchiori and others, 'Artificial Intelligence Decision Support for Medical Triage' [2020] AMIA Annual Symposium Proceedings p. 793.

432 E.g. Andorno, 'The Right Not to Know: An Autonomy Based Approach' (2004) 30(5) Journal of Medical Ethics 435-9.

433 Bunnik and others, 'The New Genetics and Informed Consent: Differentiating Choice to Preserve Autonomy' (2013) 27(6) Bioethics p. 348, 350-351.

but that the patient has not consented to categories of tests with certain aims at all.[434]

AI that function by reference to widely defined goals will involve similar concerns. They may be initiated in a procedure for one broad, unspecified purpose and then generate knowledge in surprising ways, without the possibility for direct oversight. For example, an analysis of an ECG may, when supplemented by ML techniques, provide a non-invasive method for determining potassium levels – a feat that was previously impossible.[435] Once this knowledge is generated, a doctor may not be able to discuss its disclosure without revealing its existence.[436]

Drawing on the AI-Pathway Companion Prostate Cancer and case studies like it, one can imagine an AI tool that is used to conduct a multi-purpose analysis of existing patient data and which finds a rare form of cancer in an unanticipated way.[437] This finding cannot be brought up abstractly to the patient after the fact in order to engage their reasoning and judgment. It is important that the patient is made aware *ex ante* that a diagnosis with this purpose and the intendent aims was a possibility.

Lastly, it is also important to realise that these issues do not arise in a binary fashion, but on a sliding scale. Some AI that partially replace human expert decision-making will be suited to drive medical decision-making, although they do not determine an aspect of the clinical process. The purposes of other AI may appear relatively narrow, as with Accipio Ix, but this may obscure contentious judgments. To provide the clearest illustration of this category of autonomy challenges posed by AI, the two examples given here are admittedly futuristic. Yet they draw on existing clinical studies and

434 ibid 352; It is telling that Hostuic also frames the relevant issues in terms of the autonomy to initiate actions: Hostiuc in Hostiuc, *Clinical Ethics at the Crossroads of Genetic and Reproductive Technologies* (2018) 233. See also: Herring and Foster, '"Please Don't Tell Me": The Right Not to Know' (2012) 21(1) Cambridge Quarterly of Healthcare Ethics p. 20.

435 Dillon and others, 'Noninvasive Potassium Determination Using a Mathematically Processed ECG: Proof of Concept for a Novel "Blood-Less, Blood Test"' (2015) 48(1) Journal of Electrocardiology p. 12; Topol, *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again* (2019) 61-68.

436 Andorno, 'The Right Not to Know' (2004) 30(5) Journal of Medical Ethics 435-9, 436. This offers another analogy to genetic testing.

437 Although Watson for Oncology is used in a relatively restricted context, it already managed to surprise doctors by identifying a rare form of the disease: Rohaidi, 'IBM's Watson Detected Rare Leukemia In Just 10 Minutes' (16.8.2016) <https://www.asianscientist.com/2016/08/topnews/ibm-watson-rare-leukemia-university-tokyo-artificial-intelligence/> accessed 4.9.2022.

the features of approved devices and, more importantly, they point towards problems that exist more widely and less visibly.

It is my contention that such scenarios – those where ML devices are used to pursue broadly defined goals and in a way that is partially determinative of the clinical decision – result in a situation where a patient fails to hold a decisionally necessary true belief. If a patient wishes to realise certain authorised preferences through a clinical intervention and it involves this kind of AI use, then they will need to accurately gauge the goals that the AI is intended to pursue. Without such beliefs they simply have no opportunity to engage their practical reasoning and cannot align their decisions and actions with their desired ends. If a patient wishes to be triaged in order to receive the most effective care, while the operating AI in fact maximises for efficiency in resource use, then the patient must be aware of this discrepancy before their case is triaged. If an open-ended AI analysis is being conducted of the patient's data, then they must know the aims of this analysis before it is conducted and that this may yield significant categories of diagnoses.

Again, we may note that things are different were there to be an effective physician-mediator. Even if the AI pursues a covert goal and, in spite of expert involvement, this is allowed to determine the actual process of decision-making, there would at least be some similar scenarios where the expert intervenes to correct the misalignment. With meaningful physician mediation there is, at most, a risk of divergence. Without it, there will be a systematic discrepancy, the extent of which depends on how well the particular patient's ends align with the AI's assumed framework.

In summation, the argument in this section has been that there is a subset of AI where there is a systemic mismatch between the general, non-personalised goals that the technology pursues and the ones that the agent may wish to structure their clinical reasoning and decision making by. Moreover, this mismatch cannot be remedied by physician mediation *ex post*, since there will be no practical reasoning left to be done on the relevant point. In consequence, there is a very strong mandate that information be provided about the goal-directed nature of AI, including some general information about the aims of its use.

## B. Theoretical rationality and changes in human-AI expertise

Recall that another injunction of our autonomy concept was that the patient must not be placed in a position where they sustain theoretically irrational beliefs. Drawing on this notion, the starting point for this section is the hypothesis that the reasoning of patients depends on certain justifiable assumptions about the expertise of medical professionals, which includes a degree of deference to such professionals' authority. The latter's privileged epistemic status appears widely accepted. Accordingly, it will be assumed that medical expertise confers some sort of domain-relative epistemic authority on medical professionals (they are 'experts').[438] In consequence, a patient will justifiably believe in the truthfulness of a medical opinion because it is provided by a medical professional and they have reason to believe, or conversely have no reason to doubt, the expert status of this professional.[439] To form a justified, theoretically rational belief they do not need to comprehensively evaluate the validity of every claim for themselves.[440]

Against this background, it is concerning that the reliance on certain AI tools is likely to give rise to situations where the patient has no grounds for changing their assumptions regarding the human professional, even though the basis for their expert status has fallen away. In particular, in our second class of case studies one can assume that there will be circumstances where the patient believes their medical professional to be an expert, or to be exercising their expertise, when in fact their relevant specialist knowledge is being derived primarily, if not solely, from an AI device. For example, in cases like IDx-DR, given the nature and complexity of the diagnosis, a patient may expect that there will be a person evaluating them with expert knowledge, when in fact they are only an augmented assistant.[441]

438 Funer, 'The Deception of Certainty' (2022) 25(2) Medicine, Health Care and Philosophy p. 167, 169.

439 Wagemans states this succinctly: 'the claim that the opinion involved is an ''expert opinion'' functions as an argument for the truth or acceptability of that opinion': Wagemans, 'The Assessment of Argumentation from Expert Opinion' (2011) 25(3) Argumentation p. 329, 331.

440 Goodwin, 'Accounting for the Appeal to the Authority of Experts' (2011) 25(3) Argumentation p. 285.

441 This is of course subject to the caveat that the patient may be informed of the assistant's status. But this is itself a remedy for a potential violation and as such will be returned to below.

Such examples constitute grounds for asking whether the patient's justified reliance on expert knowledge cannot be translated into a justified reliance on a human non-expert whose technical capabilities are augmented by AI. Why would the reasoning of a patient be justified when the relevant information comes from an expert, but not justified when it comes from a human-AI combination?

Two grounds will be adduced here. First, the theoretical appropriateness of relying on expert knowledge is premised in part on the ability of the advisee to critically interrogate it, which will not be possible where the only or primary source of a piece of knowledge is an opaque AI. Second, and relatedly, an individual's practical reasoning with expert knowledge is often premised on trust. The kind of justification that AI relies on may provide grounds for believing in its reliability, but it is not capable of instilling trust in the same way.

Regarding, the first ground, Walton has highlighted 'the critical nature of dialogue between the expert and the user of expert advice in argumentation' as a basis for the rational acceptance of an argument from expert authority.[442] Such a dialogue may not always be easy, even with human experts,[443] but 'interpretation, questioning and clarification' is possible.[444]

Concretely this means that, where information is conveyed in a practical decision-making setting such as medicine,[445] by someone claiming to possess expert status, the presumptive force of this information can be tested by the advisee. Even if they lack expert knowledge they can assess, for example, the credibility of the expert status of an individual, the consistency of the claim made by them with the claims of other experts, and the evidence that is relied upon to back up their claims.[446] Likewise, if someone puts forward another expert's views, then a secondary dialogue can be presupposed, in relation to which the same kinds of critical questions can be asked.[447] If these questions cannot be adequately answered, the appeal to expert opinion loses much of its force.[448]

AI's challenge to theoretical rationality emerges partially from the fact that it precludes these kinds of questions from being answered. Funer

442 Walton, *Appeal to Expert Opinion: Arguments from Authority* (1997) 112.
443 ibid 113-114.
444 ibid 114.
445 ibid 118-119.
446 ibid 258.
447 ibid 121-122.
448 ibid 258-259.

is one commentator who has picked up on this dimension, arguing that relevant AI techniques evade 'such deliberative scrutiny'.[449] But one can also see that such a conclusion follows directly from our discussion of AI opacity in the previous chapter. Assessing AI credibility is much more complex and not easily demonstrated by established methods. The precise evidence that the AI relies upon will not be apparent – even in cases of explainable AI. For it will not be possible to provide an intuitive, faithful explanation of individual elements being processed through the ML model and the kinds of human-interpretable justifications being applied to them. Moreover, comparing an ML device's output with the opinions of other expert will be more difficult in scenarios where a non-interrogable device is providing a non-specialist professional with the requisite expertise. It is entirely plausible that they themselves will lack the context to aid the patient's practical decision making in this respect.

Stated simply, a situation involving a replacement of overall human expertise with AI expertise – i.e. where the ML device is the only agent possessing a requisite level of knowledge and skill with regard to an aspect of the patient's care – a patient will at least have to engage with this expertise differently to make a rational decision. As our previous discussion of explainable AI demonstrates, this shortcoming cannot be remedied by technological means alone; by attempting to make a secondary ML model that is responsive to critical inquiry. The responses given by an explainable model only approximate to the actual decision-making process and they will sometimes provide the patient with a misleading impression. Therefore, even if a patient knows an AI is used and they attempt to exercise their theoretical rationality *vis-à-vis* expertise normally, by critically interrogating ML functioning, they ought not to attribute the same kind of weight to the responses they receive.

Such a concern leads on to our second ground. It could be rational to adhere to the recommendations of actors, even where other avenues of critical questioning are limited, if they are sufficiently reliable or trusted.[450] Indeed, this will be the natural recourse of an individual who is uncertain whether a certain bias has influenced a human expert, or whether they are lying – possibilities that are arguably comparable to the shortcomings

449 Funer, 'The Deception of Certainty' (2022) 25(2) Medicine, Health Care and Philosophy p. 167, 174.

450 Krishnan, 'Against Interpretability' (2020) 33(3) Philosophy & Technology p. 487.

of explainable ML models.[451] Walton submits that a patient must make a subjective assessment in these circumstances of reliability or trustworthiness.[452]

For our purposes it is important to distinguish between these two concepts. There is a fundamental differentiation between trust in human agents and reliance on technical tools. Ryan has elaborated this distinction in depth, stating that trust exhibits a 'concern about the trustee's motivation for action',[453] whereas reliability lacks this concern and is preoccupied with the evaluation of past behaviour and its extrapolation in order to make predictions.[454] Reliability indicates that 'the network tends to track the truth'.[455] By contrast, the motivational aspect, which is highly pertinent to the clinical sphere, appeals to the anticipation of the trustor that the trustee will act with good will towards them, exercising their skill in their interest.[456] This posits a subjective, affective attitude towards the trustee that allows for the trustor's rational evaluation of their actions and recommendations.[457]

With regard to AI, it has been convincingly argued that such an anticipation is simply not possible.[458] Nothing in our technical analysis alters this viewpoint. Quite the reverse. The connection between the values an ML device tracks, and the values of patients was argued to be tenuous, especially where the AI could exercise a wide discretion. Furthermore, nothing in the technology's design suggested an affective attitude towards an individual or a specific case. It was found to be an important role of human professionals to exercise their practical judgment to introduce relevant qualifications in the context of individual decisions.

In conclusion, if an AI replaces an element of human expertise in the clinical decision-making process, then a patient must be given the oppor-

451 Walton, *Appeal to Expert Opinion: Arguments from Authority* (1997) 115.
452 Walton himself recognises this 'trustworthiness question': ibid 115, 258. Although he does not distinguish it from reliability, as will be done below.
453 Ryan, 'In AI We Trust: Ethics, Artificial Intelligence, and Reliability' (2020) 26(5) Science and Engineering Ethics p. 2749, 2752.
454 ibid 2759.
455 Smart and others, 'Why Reliabilism Is Not Enough: Epistemic and Moral Justification in Machine Learning' (AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, USA, 07.02.2020-09.02.2020), 374.
456 Ryan, 'In AI We Trust' (2020) 26(5) Science and Engineering Ethics p. 2749, 2752-2753.
457 ibid 2760-2761.
458 ibid 2761-2762.

tunity to adjust their assessment of these recommendations accordingly. In such situations it will not be possible, or at least more difficult, for them to adequately bring their theoretical rationality to bear on the relevant recommendation. Nor will they be in a position to trust the technology in a manner that they are accustomed to trust a human expert. In consequence, they must be put in a position where they can adjust their confidence in the relevant output or recommendation.

Given our analysis in Chapter 2, this challenge is most likely to arise where an AI device leads to a diminishment of the overall level of expertise brought to bear on a decision: where a less qualified individual performs a role that was previously reserved for a more qualified one. By comparison, it will not occur in the case of, for example, Mia (Mammography Intelligent Assessment) where the technology merely offers a second opinion without influencing the initial, separate human decision, which remains trusted and interrogable.

The distinction between these categories must necessarily remain fluid. Not least due to the dimension of informational manipulation that will be evaluated below. This may diminish the extent to which a human actor can independently bring their expertise to bear on a decision, even if they do possess a requisite degree of skill.

## C. Positive freedom and the task of ensuring an adequate understanding of AI

The autonomy of a patient may also be challenged on the basis that they do not possess an adequate understanding of the AI device itself. Two categories of information will be adduced here. First, understanding the way in which the ML technology underlying AI is generically related to the risk profile surrounding its use allows a patient to act effectively in pursuit of their ends. It enables them to surpass the threshold level of positive freedom that is required for them to deliberate and act effectively. Second, given that AI devices will increase the potentially manipulative influences that a patient is exposed to, understanding AI's relation to such influences is necessary to assess the risk that this poses to the patient's autonomy.

### 1. General risk characteristics of AI

Risks are a function of the probability that a harmful event occurs in the future and of the extent of that event's harmful impact.[459] Such risks are often related to, and balanced against, the future, expected beneficial impact that is associated with the same event.[460]

There can be no question that specific clinical AI are intended to provide such clinical benefits and that they simultaneously pose risks in terms of the physical well-being of the patient, which constitutes one commonly accepted way to understand harm in the bioethical context. A diagnostic AI such as the Acumen Hypotension Prediction Index Software, described in Chapter 2, exemplifies how the technology shifts the risk-profile of an intervention. By expanding 'the diagnostic and monitoring abilities currently available in operating rooms, which fail to predict hypotension at an early stage',[461] it ameliorates an existent risk (provides a relative benefit). Yet it also creates new dangers, such as the prospect of incorrect, unnecessary interventions in response to false positive readings. This would create the chance of a different kind of physical harm materialising, whose exact likelihood is the result of the design and use of the AI.[462] Like most clinical interventions, we may therefore suppose that AI devices give rise to the possibility of a harmful event occurring in the future and that they co-determine its likelihood.

Holding beliefs about such risks is highly relevant to patient autonomy, although their absence does not disconnect a patient from the pursuit of their ends. Rather, it is a piece of information that increases a patient's understanding so that they can exercise their practical rationality more effectively, for example by placing them in a better position to assess and

459 Rid and Wendler, 'Risk-Benefit Assessment in Medical Research – Critical Review and Open Questions' (2010) 9(3-4) Law, Probability and Risk p. 151, 152. See also: Perry in Lewens, *Risk: Philosophical Perspectives* (2007) 190.

460 Rid and Wendler, 'Risk-Benefit Assessment in Medical Research – Critical Review and Open Questions' (2010) 9(3-4) Law, Probability and Risk p. 151, 151-152.

461 Davies and others, 'Ability of an Arterial Waveform Analysis-Derived Hypotension Prediction Index to Predict Future Hypotensive Events in Surgical Patients' (2020) 130(2) Anesthesia and Analgesia p. 352, 352.

462 U.S. Food & Drug Administration, 'De Novo Classification Request for Acumen Hypotension Prediction Index Feature Software' (16.3.2018) <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/denovo.cfm?id=DEN160044> accessed 7.3.2022.

reason about their different options.[463] It is imaginable that such information will play a significant part in a patient's decision-making process, allowing them to gauge whether they should rely on a procedure or choose an alternative course of action.

Concretely, thinking back to Chapter 2 one could imagine a patient who is made aware that their X-Ray is being prioritised according to an automated procedure with a certain risk profile or even with a propensity for certain mistakes. They may then be in a position to consider whether their case is being subjected to an obvious error and respond accordingly – seeking out additional attention from their care team. Risk disclosure thereby becomes an enabling factor that is necessary to secure an appropriate degree of positive freedom for the patient.

This relation between AI and particular risks does not secure the patient very much information about the AI being used in their care, however. Indeed, as Schönberger has noted, 'even AI applications in riskier areas would not add anything novel' to the information that the patient requires.[464]

Arguably, this relatively mundane disclosure provides insufficient information about AI-related risk. In particular, a patient may assume that the risk assessment around AI is comparable to that carried out for other procedures involved in their medical care. Yet there are general characteristics of ML technology that complicate the picture and which a patient must be placed in a position to assess for themselves.

An analogy can be drawn to the characterisation of innovative treatments. The innovative nature of a procedure is not merely an aspect of the specific type or magnitude of risks that the patient is facing in their care. Rather, the classification as 'innovative' itself conveys generic information to the patient that it is important for them to accommodate in their reasoning process if they are deliberating about their options. Especially when this involves a choice between an innovative and non-innovative alternative.

The first piece of generic information is that there is normally a lack of scientific testing of novel procedures. These do not require one to proceed according to the ordinary standards of validated scientific knowledge: 'in offering innovative treatment, the physician is working on a hunch or

463 Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 166.

464 Schönberger, 'Artificial Intelligence in Healthcare' (2019) 27(2) International Journal of Law and Information Technology p. 171, 188.

scientific theory that has not been adequately investigated or researched'.[465] Second, and relatedly, there is a general assumption that there is a greater degree of uncertainty in such situations about the nature or degree of physical harms to which the patient is exposed.[466] Similar arguments in favour of a category-driven approach have been advanced in relation to the unlicensed or off-label use of medical products and devices.[467]

While it cannot be assumed that the classification of procedures as novel or off-label exactly tracks their risk profile – and that this profile differs categorically from 'normal' medical interventions[468] – nevertheless a relevant categorisation is seen as an important risk-related factor for the patient to be aware of, to consider and to weigh in their decision-making.[469]

One explanation, under Pugh's account, stems from the fact that the provision of some information about risks can aid reasoning, but that an overly complex description of relevant factors does not further, and will in fact hinder, it.[470] Given that a detailed explanation of the (lack of) scientific evidence or approval, or the mechanisms associated with this, and the different reasons for uncertainty could negatively impact a patient's capacities for reasoning, it is not supportive of autonomy to disclose them. Yet the positive dimension of practical autonomy can require that the patient has some understanding of the procedure's type and the general way in which this is likely to impact the risk calculus. This provides the patient with an indication that their treatment is subject to distinct considerations and provides them with an opportunity to deepen their understanding of the more detailed implications of this classifier if they so wish.

465 Chan, 'Legal and Regulatory Responses to Innovative Treatment' (2013) 21(1) Medical Law Review p. 92, 94.

466 ibid 94. Uncertainty was also emphasised as a characteristic in the English case of *Simms v Simms* [2003] Fam 83.

467 'Lack of approval does not necessarily mean that the drug is dangerous or ineffective, but it should raise a concern about safety that the patient should weigh in deciding whether to consent to the treatment since it has not been proven safe and effective for the prescribed purpose. Since most off-label uses are not supported by scientific evidence, they may be ineffective or even detrimental': Johns, 'Informed Consent: Requiring Doctors to Disclose Off-Label Prescriptions and Conflicts of Interest' (2007) 58(5) Hastings Law Journal p. 967, 1015.

468 Beck and Azari, 'FDA, Off-Label Use, and Informed Consent: Debunking Myths and Misconceptions' (1998) 53(1) Food and Drug Law Journal p. 71, 84.

469 Price, 'Remodelling the Regulation of Postmodern Innovation in Medicine' (2005) 1(2) International Journal of Law in Context p. 121, 137.

470 Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 169, 176.

On similar grounds, it is suggested that such a categorical disclosure obligation should extend to the use of AI. It has generic features that are closely analogous to those associated with innovative treatments. These features may not always be present, or bear a direct relation to the intervention's risk profile, but they are prone to have this effect and this may influence the balance of the patient's considerations. At the same time, as Schoenberger rightly remarks,[471] it would be too much to demand of the patient to understand many of the underlying technical details.

One comparable ground is provided by ML-interventions' lesser reliance on established scientific knowledge. The dominant ML techniques driving AI will not rely on scientific knowledge in the same way that a human doctor might. Rather than basing recommendations on the validated testing of proffered hypotheses, it was seen that the algorithm's judgment stems from the establishment of probabilistic correlations.[472] Such correlations may align well with the state of scientific knowledge, but they may also offer completely new insights. This is one of the promises of medical AI after all. Even where AI outputs are seemingly generated according to established medical wisdom, they may latently be subject to certain confounders, causing the AI not to recognise a 'true signal' or causal relationship, and relying instead on a correlated but causally insignificant indicator, such as the presence/absence of a ruler in an image.[473]

Moreover, one cannot ignore the added difficulties in scientifically validating the use of AI. The multidimensional factors that these machines consider, as well as the multiple dimensions of uncertainty that surround their deployment, make its assessment according to the 'gold standard' of randomised control trials very difficult. Assessing even offline AI has been said to be more akin to the 'evaluation of highly complex health care delivery interventions'.[474] Where randomised control trials have been carried out, there have still been calls for caution based on the size and diversity of the sample.[475] Once one moves to online AI such difficulties

471 Schönberger, 'Artificial Intelligence in Healthcare' (2019) 27(2) International Journal of Law and Information Technology p. 171, 188.

472 Bennett and Doub in Luxton, *Artificial Intelligence in Behavioral and Mental Health Care* (2016) 30.

473 Narla and others, 'Automated Classification of Skin Lesions' (2018) 138(10) The Journal of Investigative Dermatology p. 2108.

474 Angus, 'Randomized Clinical Trials of Artificial Intelligence' [2020](11) The Journal of the American Medical Association p. 1043.

475 ibid and see Chapter 2 for a more comprehensive analysis of these issues.

are compounded by an evolving model. This fits itself to the population it serves and it would be self-defeating to test it every time it changed the nature of its functioning. Ultimately, as with innovative treatments, one should note that procedures implementing AI can be expected to have a more tenuous relationship with scientific knowledge and optimal forms of validation.

On top of this, one must account for the uncertainty stemming from the black box nature of any individual decision. The patient is faced with a decision that cannot be easily comprehended and they cannot know what factors contributed to it, how they were weighed or according to what criteria. As a consequence, there is a risk that the procedure does not confer the benefit that the individual patient would want or expect it to confer or that they are exposed to harms, which are only associated with certain (unidentified) populations that the AI is ill-suited to. Such uncertainties can be expected to persist even if the AI is validated and found accurate in a general sense.[476]

Of course, these relations to scientific knowledge and the degrees of uncertainty that arise are partly dependent on the features of any given ML device. They will be impacted *inter alia* by: the ML techniques being used, how they have been validated and by the information that is made available to users by the developer. Uncertainties about the factors going into a decision could be influenced by the type of interpretability or explainability mechanisms that are implemented, with the latter compounding uncertainties and adding a potential for misleading or incomplete responses (regarding which, see the next section). As was seen in Chapter 2, evaluations can also offer more or less detailed breakdowns of differential impacts on population groups.

Nevertheless, the general trends that AI exhibit in this regard are grounds for singling out ML devices as a type of tool whose challenges are comparable to those of innovative treatments. As a class they have a peculiar relation to forms of scientific knowledge and they incorporate significant types of uncertainty. In this way, the use of ML in medical care is indirectly related to the risk assessment of patients regarding their physical well-being. Understanding AI-based treatment as an accessible categorisation of risk would provide patients with relevant information that furthers their

476 Abràmoff and others, 'Pivotal Trial of an Autonomous AI-Based Diagnostic System for Detection of Diabetic Retinopathy in Primary Care Offices' (2018) 1 NPJ Digital Medicine p. 1.

positive autonomy, without overloading and hindering their decision making.

## 2. Informational manipulation

Sometimes an AI will challenge autonomy because, even where the technology does not determine a choice directly (as in Section II.A.), it exerts an influence on the patient's decision and this influence is manipulative: it causes them to reason in a theoretically non-rational manner and/or to act in practically non-rational ways.[477] This influence can be conceived of as a nudge, a concept that was popularised by Thaler and Sunstein.[478] They define this as 'any aspect of the choice architecture that alters people's behaviour in a predictable way without forbidding any options or significantly changing their economic incentives'.[479]

Although this work does not purport to offer a novel analysis of nudges, it is worth amending and clarifying this definition for our purposes. First, we may say that it is not just any aspect of the choice architecture that has normative implications. Rather, as Hansen and Jespersen have argued, it applies to 'attempts at influencing choice' that are 'directed towards any well-defined consistent end'.[480] This distinguishes targeted interferences that are liable to align individual behaviour with external aims from mere accidental environmental noise.[481] Second, particularly in the healthcare sphere it should be self-evident that we are not only concerned with

477 Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 80.

478 Thaler and Sunstein, *Nudge* (The Final Edition 2021).

479 ibid 8.

480 Hansen and Jespersen, 'Nudge and the Manipulation of Choice: A Framework for the Responsible Use of the Nudge Approach to Behaviour Change in Public Policy' (2013) 4(1) European Journal of Risk Regulation p. 3, 9-10.

481 Hansen and Jespersen frame this distinction more broadly in terms of intentionality: ibid 10. If intentionality were necessary, then it would be problematic, as it is questionable whether the relevant kind of intentional agency applies to AI: Froese and Ziemke, 'Enactive Artificial Intelligence: Investigating the Systemic Organization of Life and Mind' (2009) 173(3-4) Artificial Intelligence p. 466. Yet I do not believe that this step is crucial for distinguishing random behavioural influences of the environment from external manipulations, especially as autonomy violations can be non-agential under our procedural theory of autonomy: Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 61.

economic incentives, but with much broader benefits and costs.[482] Third, there is a sense that the influence of the nudge is in some way separated from the provision of reasons that can be deliberated upon. While nudges may partly appeal to deliberation and reasons, they aim to direct cognitive processes beyond this, by exploiting unconscious, hidden processes that are not necessarily in line with one's reasoned preferences.[483] It is in this sense that they undermine both the cognitive and, especially, reflective element of deliberative autonomy and further irrational or non-rational tendencies.

### i. AI nudging

AI's ability to induce biases was already examined in Chapter 2. By leveraging its unique capabilities, it is predicted that AI will increase the influence that design-induced psychological mechanisms have on making individualised decisions about patient care.

Crucially, it must be noted at this stage that even in the case of physician mediation, various AI will be given the opportunity to influence patient behaviour directly. This is significant as, if the AI's presented information would only be used by the physician, then it would be only the physician's choice that would be affected. For example, as the Acumen Hypotension Prediction Index Software information is only intended to reach the physician, there is no opportunity to impact the patient's decision-making. So long as the patient has sound grounds for relying on the physician (see Section II.B.) patient autonomy is not engaged.

It is where the AI is incorporated as another party into the deliberative decision-making process, that its nudges take on the quality of challenges to autonomy. Such a role is clearly envisioned for certain clinical AI, especially where they are involved in relatively broad decision-making tasks. For example, Watson for Oncology is capable of generating 'shareable individual treatment plans and patient education materials to engage the patient'.[484]

482 Hansen and Jespersen, 'Nudge and the Manipulation of Choice' (2013) 4(1) European Journal of Risk Regulation p. 3, 7.

483 ibid 13-15.

484 IBM, '5725-W51 IBM Watson for Oncology: Sales Manual' (2020) <https://www.ibm.com/common/ssi/cgi-bin/ssialias?appname=skmwww&htmlfid=897%2FENUS5725-W51&infotype=DD&subtype=SM&mhsrc=ibmsearch_a&mhq=IBM%20WATSON%20ONcology> accessed 18.3.2023.

The use of AI as a patient-facing triage device is another example where nudging becomes a tangible possibility.

### ii. Impermissible manipulation

Granting their existence, it is important to indicate precisely how such influences may be challenging for autonomy. This links to the widespread claim that nudges are problematic for autonomy because they work by manipulating choice.[485] On such a view an AI that nudges will be problematic without more.

Yet this sits uncomfortably with Pugh's process-oriented account of autonomy. It is possible to be influenced without having one's deliberative or practical autonomy diminished.[486] Indeed, prompting people to determine their organ donor status is a classic example of a nudge.[487] Even if there is an appeal to non-rational mechanisms, overall this can promote one's autonomy by causing one to make a decision that reflects one's wider, authentic preferences.[488] Benign AI nudging would occasion an opportunity for the kind of evaluative judgments regarding one's preferences and rational acceptances that furthers the patient's reflective autonomy in reaching care decisions.

The task thus becomes identifying the features of *certain* nudges that make them problematic for autonomy ('manipulative') and determining whether the influences introduced into individual clinical decisions by AI exhibit these factors. The dominant features that can be found in the literature in this respect relates to goal-divergence. There is a sense that a nudge presents a greater danger for autonomy – or only presents a danger for autonomy – when it diverges from the goals of a patient. This is evident with respect to Thaler and Sunstein's claim that nudges are liberty preserving when they 'aim to influence choices in a way that will

485 Hansen and Jespersen, 'Nudge and the Manipulation of Choice' (2013) 4(1) European Journal of Risk Regulation p. 3, 5. More recently see: Ploug and Holm, 'Doctors, Patients, and Nudging in the Clinical Context--Four Views on Nudging and Informed Consent' (2015) 15(10) The American Journal of Bioethics p. 28.

486 Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 79-82.

487 Thaler and Sunstein, *Nudge* (The Final Edition 2021) 269-271.

488 Hansen and Jespersen, 'Nudge and the Manipulation of Choice' (2013) 4(1) European Journal of Risk Regulation p. 3, 21.

make choosers better off, *as judged by the choosers themselves*'.[489] Hereby nudges do not merely track an individual's anticipated choice, but they seek to realise the objects of their desires, which they may otherwise be too weak-willed or inattentive to pursue.[490] One may say that they trigger a process that enhances an individual's ability (their positive freedom) to engage in the deliberative, reflective dimension of autonomy.

The link between the goals of the chooser and the nudges that they are subjected to has also been appealed to directly in the bioethical literature. Specifically, Blumenthal-Barby and Naik maintain that nudges in the individual clinical context will tend to have a comparable autonomy-maintaining or autonomy-enhancing effect because they are responsiveness to patient values. This is framed more abstractly from the libertarian paternalistic notion of tracking the objects of an individual's desires, which is already distinguished from an anticipation of their specific choices.

For these authors it is crucial that clinicians can and do 'use their understanding of patient's values and informed preferences to guide their nudges in ways that help patients to more efficiently work through competing goals'.[491] This does not entail that the nudges exactly promote the goals that the patient wants to promote. It simply entails that the patient is able to deliberate about and clarify their own objectives in response to the nudged behaviour. From the perspective of procedural autonomy this would be acceptable: it can be conducive of rationality to promote diverging goals and choices in the clinical interaction.[492]

What all of these accounts therefore require is that nudges have some positive contribution to make to the maintenance of reflective autonomy. They either track general preferences or are in some way responsive to an individual's circumstances to promote reflection and thus deliberation.

As has been elaborated in Section II.A. above, AI design will not be closely tailored to individual preferences. It will be shaped by the goal-directed behaviour of the AI and its designers. For example, the content and form of information will plausibly be connected to its pursuit of different incentives provided by the designer and by the sub-goals that the AI

489 Thaler and Sunstein, *Nudge* (The Final Edition 2021) 7.

490 ibid 7-8. For a statement of how it is possible for a disjuncture to arise between one's effective motivations and the objects one desires most see: Mele in Mele and Rawling, *The Oxford Handbook of Rationality* (2004).

491 Blumenthal-Barby and Naik, 'In Defense of Nudge-Autonomy Compatibility' (2015) 15(10) The American Journal of Bioethics p. 45, 45.

492 Pugh, *Autonomy, Rationality, and Contemporary Bioethics* (2020) 61-63.

develops to approximate to the successful performance of these tasks. If an AI's goal is the promotion of disease-free survival, then the options that it provides, and their ordering, will naturally reflect this goal. In this manner a bias is introduced into decision-making that does not align with a patient's general or specific interests. One can see how guidance issued to professionals is beginning to respond to this danger,[493] although as discussed in Chapter 2, the precise nature of this phenomenon still needs to be adequately researched and assessed.

For a nudge to be justifiable, this leaves open the possibility of it generating a prompt for reflective thinking. To the extent that the AI pursues a therapeutic goal, and its use is accompanied by human mediation, it is arguable that such thinking is a probable consequence of the technology's use. A patient who is confronted with a default therapeutic choice, but is provided with contextual information surrounding this, is able to form their own judgment in response to that prompt – whatever interests it exactly promotes.

The most problematic situations will occur where a patient cannot contextualise a prompt by reference to wider clinical reasoning. In other words, where an ML device exerts influence in favour of a non-therapeutic objective. There are expected to be instances where this will occur. It was seen that AI must account for non-therapeutic considerations, such as the clinical resources available in a particular care setting or for the reimbursement conditions of a particular healthcare institution and system in which it operates. It also stands to reason that explainable AI or online AI would covertly pursue other goals that are not strictly speaking therapeutic, such as rendering a persuasive representation of ML reasoning to the user or improving ML performance. Where such factors are incorporated into the presented information, there is the starkest disconnect to the kind of nudges that could trigger individual reflection regarding a patient's therapeutic situation.

In sum, the outlined non-obvious pursuit of certain values can preclude an AI from serving as a tool that encourages reflective thought. Instead, it becomes a tool that surreptitiously introduces certain commitments that are never reflectively assessed or endorsed. Our analysis from the previous

493 Nix, Onisiforou and Painter, 'Understanding Healthcare Workers' Confidence in AI' (2022) <https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning/understanding-healthcare-workers-confidence-in-ai> accessed 11.11.2022.

chapter further suggests that it would require considerable effort to detect and counteract these.[494]

Therefore, to the extent that there are such mismatches in values between patient and AI, and the patient cannot be expected to identify and reflect upon these, there is a problem for patient autonomy. The magnitude of this interference must depend on the circumstances of the case. Yet, as a general rule, it may be thought that this challenge is on the lower end of the scale. The patient will be influenced by a number of factors after all and even if an AI nudge is real, it will be difficult to establish that it impaired a relevant decision process in a way that significantly impacted an individual's positive freedom.

## *III. Conclusion*

In conclusion, this chapter sets out the link between the technical knowledge that was collated in Chapter 2 and the legal analysis that is to follow. Pugh's rationalist theory of autonomy has provided us with a means of doing so: identifying four categories of AI challenge that are detrimental to a procedural account of autonomy. In moving on to our legal analysis of these challenges, it is worth remembering that they are not all of a comparable scale. Disconnecting a patient from the pursuit of their goals was understood as a particularly problematic prospect. The different nature of AI risks and of AI expertise may also be understood as something that is of considerable importance to the maintenance of the patient's theoretical and practical autonomy. By contrast, the prospect of AI nudging and manipulation was categorised as an existent, but lesser interference with the patient's positive freedom.

494 For an argument elaborating how some nudges are more difficult to resist – weakening 'attention-bringing and inhibitory capacities' – see: Saghai, 'Salvaging the Concept of Nudge' (2013) 39(8) Journal of Medical Ethics p. 487.