



WHAT IS RIGHT FOR ‘THIS’ PATIENT? AUTONOMY, TRUST AND TRANSPARENCY IN THE ERA OF MEDICAL AI *

O que é correto para ‘este’ paciente? Autonomia, confiança e transparência na era da inteligência artificial médica

Anna Smajdor **

Arnoldo Frigessi ***

Manuela Zucknick ****

Abstract: In this paper we compare the capabilities and limitations of AI algorithms and human doctors in the context of treatment recommendations. Ethical concerns have been raised in the literature about transparency, accountability, and interpretability in relation to AI algorithms in the medical context. Patients’ autonomy is not likely to be respected if treatment recommendations are opaque or incomprehensible. Medical AI algorithms are programmed to prioritise predefined goals that may not align with patients’ preferences, potentially undermining their autonomy without their awareness. Yet human doctors often face similar challenges, such as reliance on clinical trial data with endpoints that may not be fully understood by patients. We conclude that there are opportunities now to integrate patient autonomy considerations into AI algorithms at the design

* Arquivo recebido em 06.10.2025 e aprovado para publicação em 14.11.2025.

** Anna Smajdor (corresponding author) PhD, Professor of Practical Philosophy, University of Oslo. ORCID: <https://orcid.org/0000-0002-9752-6302> E-mail: anna.smajdor@ifikk.uio.no.

*** Arnoldo Frigessi, PhD, Professor of Statistics at the Oslo Centre for Biostatistics and Epidemiology, University of Oslo. ORCID: <https://orcid.org/0000-0001-7103-7589>. E-mail: arnoldo.frigessi@medisin.uio.no

**** Manuela Zucknick, PhD, Professor of Statistics at the Oslo Centre for Biostatistics and Epidemiology, University of Oslo. ORCID: <https://orcid.org/0000-0003-1317-7422>. E-mail: manuela.zucknick@medisin.uio.no.

stage, and that these goals and any relevant weightings are made explicit to end users, including both doctors and patients.

Keywords: Artificial intelligence. Autonomy Alignment. Medical ethics.

Resumo: Neste artigo, comparamos as capacidades e limitações dos algoritmos de inteligência artificial e dos médicos humanos no contexto de recomendações terapêuticas. A literatura tem levantado preocupações éticas acerca da transparência, da responsabilização e da interpretabilidade dos algoritmos de IA no âmbito médico. A autonomia dos pacientes dificilmente será respeitada se as recomendações de tratamento forem opacas ou incompreensíveis. Os algoritmos de IA médica são programados para priorizar objetivos predefinidos que podem não estar alinhados com as preferências dos pacientes, podendo comprometer sua autonomia sem que eles tenham consciência disso. No entanto, médicos humanos frequentemente enfrentam desafios semelhantes, como a dependência de dados de ensaios clínicos cujos desfechos podem não ser plenamente compreendidos pelos pacientes. Concluímos que existem oportunidades, já no estágio de projeto, para incorporar considerações sobre a autonomia do paciente nos algoritmos de IA e para tornar explícitos aos usuários finais — tanto médicos quanto pacientes — esses objetivos e os eventuais pesos atribuídos a eles.

Palavras-chave: Inteligência artificial. Autonomia. Alinhamento. Ética médica.

Introduction

In recent years, the degree to which AI algorithms can play a part in ethically-sound medicine has been an active topic of debate among ethicists.^{1,2,3} In particular, AI-aided medical decisions raise questions about trust and transparency.⁴ AI algorithms are trained by optimising a predefined utility, which represents the ultimate ‘intent’ of the algorithm, as incorporated by the programmers. However, depending on the type of AI model, its ‘intent’ in terms of its utility goal, might not be made public. Furthermore, it is not always possible to trace back the process of reasoning that leads to a specific AI decision or recommendation. This is the so-called ‘black box’ phenomenon.⁵

¹ GROTE, T. Trustworthy medical AI systems need to know when they don’t know, in *Journal of Medical Ethics*, v. 47, n. 5, May 2021, p. 337-338. DOI: 10.1136/medethics-2021-107463.

² KERASIDOU, C.; KERASIDOU, A.; BUSCHER, M.; WILKINSON, S. Before and beyond trust: reliance in medical AI, in *Journal of Medical Ethics*, v. 48, n. 11, Nov. 2022, p. 852-856. DOI: 10.1136/medethics-2020-107095.

³ LUXTON, D. D. AI decision-support: a dystopian future of machine paternalism?, in *Journal of Medical Ethics*, v. 48, n. 4, Apr. 2022, p. 232-233. DOI: 10.1136/medethics-2022-108243.

⁴ MAINZ, J. T. Medical AI: is trust really the issue?, in *Journal of Medical Ethics*, 2023 (Online first, 26 jul. 2023). DOI: 10.1136/jme-2023-109414. Available at: <https://jme.bmj.com/content/early/2023/07/26/jme-2023-109414>. Accessed on 26 Nov. 2025.

⁵ DURÁN, J. M.; JONGSMA, K. R. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI, in *Journal of Medical Ethics*, v. 47, n. 5, May 2021, p. 329-335. DOI: 10.1136/medethics-2020-107096.

If an AI algorithm cannot produce an explanation for its treatment recommendations, nor even state what its programmed utility goal is, this poses problems for patient autonomy, and perhaps for medicine in general. Modern medicine seeks for clarity, replicability and transparency. Before evidence-based medicine became the gold standard, patients might have expected at the most to be told what to take and when. Today's doctor must explain *why* they believe a particular medication is right for a specific patient. This explanation, in ideal circumstances, should include a comparison of the advantages and disadvantages of the preferred treatment, relative to other treatment options. In the absence of this explanatory capacity, it seems that AI based recommendations cannot fully satisfy the ethical demands of respect for patient autonomy. If the AI cannot give an account of its reasoning, the doctor cannot relay this information to the patient.

One response to this might be to focus more narrowly on the question of whether the recommendation is the right one; if the AI-recommended treatment *works* then perhaps this is all that should count. Effectively, on this view, we should believe that it is acceptable for beneficence to trump autonomy in the context of medical AI. This is an example of the tension between reliabilism and transparency.^{6,7} However, the problem here is that without reference to autonomy it becomes difficult, if not impossible, to define what beneficence *is* in regard to a specific patient. Even if there were such a thing as the clinically-best solution in a given case, it is not invariably true that this is right for *this* patient. For example, a cancer patient approaching the end of life stage might prefer a treatment that allows him to spend his remaining days at home with family rather than in hospital, even if this treatment would be sub-optimal with regards to life expectancy.

If medical AI fails both in terms of respect for autonomy *and* in terms of beneficence, seemingly its place in medicine is questionable. However, this is too hasty a conclusion. It may well be that today's AI is inadequate to fulfil the most stringent demands of beneficence and autonomy, but can we really be sure that the everyday human doctor is so much better? And if we agree that AI treatment recommendations do not currently meet ethical requirements, how might future AIs be designed in such a way as to better satisfy these requirements?

In this paper, we analyse the question of how a treatment recommendation that comes from a human doctor might fare versus an AI recommendation, in terms of respect for patient autonomy.

⁶ FERRARIO, A. Design publicity of black box algorithms: a support to the epistemic and ethical justifications of medical AI systems, in *Journal of Medical Ethics*, v. 48, n. 7, Jul. 2022, p. 492-494. DOI: 10.1136/medethics-2021-107482.

⁷ MISHRA, A. Transparent AI: reliabilist and proud, in *Journal of Medical Ethics*, v. 47, n. 5, May 2021, p. 341-342. DOI: 10.1136/medethics-2021-107352.

1. *The goals of treatment*

When a doctor considers which treatment to recommend to a particular patient – let us call him John – there are many things she takes into account. There is usually a limited range of drugs licenced for use in the treatment of any specific disease. For each possible choice, there are different features. Some may have better results in prolonging the life of the patient; others may be more effective in improving quality of life, for example, by reducing the unpleasant symptoms of the disease. Some may be simple, and easy to administer for the patient himself, others require repeated blood tests and trips to the clinic. Each treatment option will have side effects, some more serious than others; some of these side-effects may be common, others extremely rare. Some treatments may be very expensive, and others relatively cheap.

When John thinks about what treatment is right for him, many factors play into his perception of how to trade off different risks and benefits. Commonly, one might assume that ‘saving the patient’s life’ or ‘curing the patient’ is the single most important aim. If so, there is no need to agonise over what goal the treatment is designed to achieve. It is given from the outset. However, what is meant by ‘cure’ may be open to question. Because of this, when treatments are developed and tested in randomised controlled clinical trials (RCTs), the goal is not the ‘cure’ of the patient per se. It is a single primary endpoint that defines whether the novel treatment in the intervention arm is more effective than the control treatment.⁸

For some cancer trials the endpoint is 6 months’ survival probability. Other trials might have a different endpoint; some might take into account other factors than overall survival time, such as, for example, the length of time the patient is cancer-free following treatment. These endpoints are highly precise and inherently inflexible. Since these goals are not patient-sensitive, they may not be well aligned with an individual’s own interests. If a patient’s remaining days or years are likely to be dominated by unpleasant side effects, reduced quality of life or time-consuming treatments, they may prefer a different option, even if the first recommendation is the one that is most likely to guarantee a longer period of survival.

Because of this, the doctor’s job in arriving at a treatment recommendation is highly complex. In order to identify the best treatment to achieve a particular goal, *that goal itself* needs to be identified. Only then will the doctor be able to evaluate the options in relation to the overall aim. If the goal is never discussed or made explicit, the patient’s autonomy will be further undermined, in that they do not have access to this information.

⁸ FLEMING, T. R.; POWERS, J. H. Biomarkers and surrogate endpoints in clinical trials, in *Statistics in Medicine*, v. 31, n. 25, Sept. 2012, p. 2973-2984. DOI: 10.1002/sim.5403.

So how are these goals of treatment established? Is there any reason to suppose that human doctors are better at identifying autonomy-satisfying goals than AI algorithms are? As noted, doctors are constrained in their treatment recommendations: they can choose from a limited number of drugs. These drugs will have been developed in trials that have specific, fixed goals built in as outcome measures. In addition, the options available to doctors are likely to be determined directly or indirectly by tools such as quality-adjusted life years (QALYs). QALYs are in effect a mechanism whereby the goals of treatment are pre-factored into an assessment of a particular intervention, along with its cost. These goals include extended life (the 'year' part of the measure) and a variety of quality-of-life measures.⁹

While it is relatively easy to measure the number of years a patient survives, it is notoriously difficult to identify and measure all the factors that can affect a person's quality of life. Therefore, the quality component is always to some degree a proxy or artificial representation. Pain is a fairly obvious undesirable side effect, but there are many, many other possible issues that can impact a patient's wellbeing (mobility issue, nausea, skin rashes, drowsiness, impotence, incontinence, anxiety, tremors, hair loss, insomnia, etc.) sometimes to the extent that a drug which is highly effective at extending life, may be intolerable to a patient because of the side effects associated with it.¹⁰

QALYs are a useful tool for health services and funders, because they take into account both length and quality of life as an outcome of a treatment, at population level. This in turn can enable a fairly simple comparison between the different available treatments, in terms of cost per QALY. But they are nevertheless a crude means by which to determine what treatment might be best for a specific individual. They are inflexible, in that the factors that feed into them are pre-fixed and the doctor cannot adjust them in the light of new information about a specific patient's needs or wishes. Despite this, they play a key role in defining what treatments are funded by health services or recommended by organisations such as the National Institute for Health and Care Excellence (NICE) in the United Kingdom.

In the case of our doctor, it is likely that the array of potential treatments from which she can choose is itself determined on the basis of QALY assessments, or something similar, over which she has no control, and perhaps limited understanding. The doctor in our case, however, does not simply consult the QALY recommendation. Where the doctor's human skill and insight come into play are in her ability to take into account information

⁹ PETTITT, D. A. et al. The limitations of QALY: a literature review, in *Journal of Stem Cell Research and Therapy*, v. 6, n. 4, 2016, p. 334.

¹⁰ KIRKDALE, R. et al. The cost of a QALY, in *QJM: An International Journal of Medicine*, v. 103, n. 9, Sept. 2010, p. 715-720. DOI: 10.1093/qjmed/hcq081

that goes beyond the QALYs, information that helps to arrive at a treatment recommendation that is ‘right for this patient’. That is, the patient himself contributes to the process of identifying the goal that treatment is intended to meet.

In order to identify this goal together with her patient, a doctor weighs many considerations often in the space of a very short time, drawing on existing knowledge, prior experience and perhaps to some extent intuition. The doctor’s analysis will lead her to a preliminary clinical judgement. Now, suppose that the patient — John — tells the doctor that he will become a father in a few weeks. He wants local treatment options that allow him to stay at home rather than make expensive and time-consuming trips to a distant hospital (even if it is the specialist hospital for his condition). John is also anxious to avoid treatment that will leave him debilitated in the months after his baby’s birth, so that he can participate as fully as possible in the care of his child. These elements of personal information about John also inform the doctor’s judgement, and eventually she recommends that John should take drug X, in the light of his treatment goals.

In short, a human doctor can spontaneously weigh up seemingly incommensurable variables, some of which she will not know of until the moment at which the decision is being made, and in this way, she has flexibility to identify treatment goals in response to, and in collaboration with, a patient such as John. In this regard, John’s autonomy as a patient is respected.

2. Algorithms and goals of treatment

So how does this compare with treatment recommendations made by an AI algorithm? The first point to make is that an algorithm has a specific treatment goal programmed into it – a predefined utility. It cannot deviate from this, while in principle at least, the doctor can. Depending on what goal it has been programmed with, the AI may be looking for treatments that will maximise the patient’s number of disease-free years or their chance of achieving complete treatment response defined as lack of detectable evidence of the tumour, e.g. through magnetic resonance imaging or histopathology.¹¹ Let us suppose for the sake of argument that in this case, the AI’s goal is to maximise the probability of 5 years survival, post treatment.

¹¹ DELGADO, A.; GUDDATI, A. K. Clinical endpoints in oncology—a primer, in *American Journal of Cancer Research*, v. 11, n. 4, 2021, p. 1121-1131. PMID: 33948349; PMCID: PMC8085844.

Like the doctor, the AI may be choosing from among a predefined range of available treatments whose goals are also pre-fixed at the RCT phase. Like the doctor, the AI can take into account a number of factors: side effects, costs, life expectancy, etc., but only if these have been specifically incorporated into the design of its algorithm. What the AI cannot do, is to accommodate John's own preferences, with all his idiosyncrasies and foibles. It cannot move beyond the rigid confines of the pre-set goals embedded in its programming to identify a new goal incorporating John's particular needs. While algorithms can be trained to perform multiple optimisations simultaneously, to receive rewards for various objectives, these are programmed in advance, and the way these various objectives are weighted in the final algorithmic decision will be fixed for all patients.

The AI is static. Its recommendations may be good in terms of achieving the 5 year survival goal, but they are not necessarily good in terms of respecting John's autonomy or identifying or responding to his goals. In contrast, the doctor's decision-making process is dynamic, she can respond at every moment to different factors that emerge during the consultation. This is precisely what the AI cannot do.

3. Training, reliability and uncertainty

We have explored the process by which the doctor arrives at the recommendation of a particular drug and observed that it allows for a degree of respect for patient autonomy that does not feature in the AI's recommendation. However, the autonomy-respecting element of the doctor's recommendation is not all that valuable if she is wrong in her judgment about what drug will meet John's treatment goals. Identifying goals in collaboration is one element of respect for autonomy, but this is an empty achievement if in fact those goals do not influence the eventual outcome. So can John really trust the doctor's recommendation, or is the AI likely to be more reliable and consistent, even if it is less responsive to individual needs?

As noted, the AI is trained to seek a particular goal. In this case, we know that the AI's goal is to maximise the probability of 5 years survival for patients suffering from the same kind of cancer as John. John's personal wishes may not perfectly align with this goal, but at the same time, it is not just any random goal, but a specifically chosen goal that might be expected to be important for most patients. John might therefore weigh his options thus: the 5 year goal isn't quite what I want, but if I'm almost certain to achieve it, if I take the AI recommended drug, maybe that is a

better outcome than going for the doctor's recommendation, which after all may fail.¹²

John knows that doctors receive years of training. But he may know very little about how the AI comes to make its recommendation, and hence, how reliable it may be. During its training phase, the AI uses the information in its training database to identify the attributes most likely to lead to 5 or more years' survival. The training database may include perhaps 100 000 patients. For each patient, there is a record of which drugs they received, how long they survived, which side-effects they experienced, etc. The data also includes additional information for each patient's previous diseases and their treatment; where they live, if they have a family, their educational level, information about their genetic profile. The machine is trained by showing it the data, and by rewarding it when it makes successful predictions about which drugs, were most likely to result in the achievement of the goal: in this case, the patient surviving for at least five years.

The algorithm identifies connections (interactions) among the complicated mix of training data, which enable it to make predictions. During this training phase, each time the algorithm correctly identifies a patient/drug combination that meets the goal of the patient's surviving 5 years, it receives a reward — say one point. If it fails to get this correct, it receives no points. The algorithm is thus striving to attain the highest reward possible, by identifying all the elements that contribute to a higher likelihood of a successful prediction.¹³

When the algorithm reaches an acceptable level of accuracy across the training data population, it can be put to work in real life treatment situations. For a new patient such as John, the algorithm will be given certain information, which again must fit certain predefined criteria. This might include the type of cancer John has, his age, weight, genetic profile, previous treatments received etc., even where he lives and whether he has a family. The AI will then process John's data and suggest a treatment option that is most likely to satisfy the goal which it is programmed for. In this case, a period of at least 5 years survival with no recurrence of the original cancer.

We can contrast this with the doctor's training. Both the algorithm and the doctor make use of existing data about other patients. But the algorithm uses data from all 100 000 patients included in its training data, whereas

¹² Note, however, that John can only undertake this evaluation if he knows that the AI's treatment goal is 5 year survival.

¹³ RAJPURKAR, P.; CHEN, E.; BANERJEE, O.; TOPOL, E. J. AI in health and medicine, in *Nature Medicine*, v. 28, n. 1, Jan. 2022, p. 31-38. DOI: 10.1038/s41591-021-01614-0.

the doctor may have only treated a few hundred patients with John's condition during her career. However, the doctor also has access to the scientific literature, whereas the algorithm's judgements are based solely on the data from the 100 000 patients that it has been 'fed'. Assuming that the scientific literature contains data from many hundreds of thousands of patients, the doctor may have learnt – indirectly – from more patients than the algorithm's large database provides. If so, it seems she has the advantage in this respect.

However, in the literature, only certain pre-identified hypotheses can be tested, while the algorithm, during its training phase, is incentivised to try a vast array of possible connections. This, after all, is the point of machine learning: it can discover connections that are not previously known, apparent or predictable for human beings. Furthermore, it is worth noting that in the very near future, algorithms may be able to learn directly from all scientific literature. Already today large language models are often trained on all existing digital texts, including essentially all the scientific literature.¹⁴ While the doctor's human brain is only capable of processing a limited amount of information; the algorithm will in principle be able to consider *everything that has ever been written* on a particular condition.

An AI of this sort, if trained to high accuracy, would have extraordinary power, well beyond what we could expect of any human. But even so, if John is looking for *certainty*, he might not be satisfied. The ability to read billions of texts is valuable to a patient only insofar as it gives rise to good treatment recommendations. An important limitation of contemporary AI is the lack of quantification of the level of uncertainty of a decision: with what confidence does the algorithm suggest drug X as opposed to drug Y?¹⁵ It is of course significant for John if this confidence is high (say 95% that drug X will be better than drug Y) or not (say that this probability is 55%). In the case of a patient like John, the absence of a possible explanation and quantification of confidence of the algorithm in its recommendation could make John feel inclined to follow the doctor's advice and take the drug that she recommends. At least she can provide a convincing argument as to why it is the right one and she can also explain how confident she is in her advice.

Thus it seems that simply knowing that the AI has access to enormous amounts of data may not be sufficient for John to feel confident that the AI recommendation is the right choice. Even if John autonomously deci-

¹⁴ JIN, Q.; YANG, Y.; CHEN, Q.; LU, Z. GeneGPT: augmenting large language models with domain tools for improved access to biomedical information. arXiv preprint arXiv:2304.09667v3, 2023. Updated in: *Bioinformatics*, 2024. PMID: 37131884; PMCID: PMC10153281.

¹⁵ BEGOLI, E.; BHATTACHARYA, T.; KUSNEZOV, D. The need for uncertainty quantification in machine-assisted medical decision making, in *Nature Machine Intelligence*, v. 1, n. 1, Jan. 2019, p. 20-23. DOI: 10.1038/s42256-018-0004-2.

des to accept 5 year survival as his treatment goal, he needs to know *why* this is the right drug to achieve that aim, and how certain it is to achieve it. And he needs to know *that* this is the goal that the AI has been programmed to aim for. Again, we seem to come back to a conflict between beneficence and autonomy. Even if not every patient can have medical treatment that is specially tailored to his specific goals, each patient can surely expect at least an explanation of what the proposed treatment is intended to achieve, and how it is expected to do so.

4. Explainability and transparency

Here, it seems the AI is at a disadvantage relative to the doctor. The AI is not necessarily able to provide an easily understandable series of steps by which it arrived at its recommendation. Today, we do not always fully understand how AIs reach their conclusions, and this is a very important theme of research.^{16,17,18} As noted above, this is the essence of the notorious ‘black box’ problem.¹⁹ If John does not know either how the AI arrived at its decision, or what treatment goal it is aiming for, he is in a very dubious position. Even worse, John may be aware that a notorious aspect of AI large language models such as ChatGPT is that they occasionally ‘hallucinate’: that is, they give completely false information, that is they generate text, which — while being grammatically and stylistically flawless and sounding very convincing — is completely wrong factually. The reasons for this are not fully understood, but certainly as it stands, the risk of hallucinations might worry patients and doctors alike if they are relying on large language models. And indeed, the use of LLMs has been advocated recently in a variety of medical settings including gaining consent and providing information to patients pre-operatively, and for teaching ethics to medical students.^{20,21}

¹⁶ SINGHAL, K. et al. Large language models encode clinical knowledge, in *Nature*, v. 620, Aug. 2023, p. 172-180. DOI: 10.1038/s41586-023-06291-2.

¹⁷ PETRONI, F. et al. Language models as knowledge bases? arXiv preprint arXiv:1909.01066, 2019.

¹⁸ HU, X. et al. Do large language models know about facts? arXiv preprint arXiv:2310.05177, 2023.

¹⁹ DURÁN, J. M.; JONGSMA, K. R. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI, in *Journal of Medical Ethics*, v. 47, n. 5, May 2021, p. 329-335. DOI: 10.1136/medethics-2020-107096.

²⁰ ALLEN, J. W.; EARP, B. D.; KOPLIN, J.; WILKINSON, D. Consent-GPT: is it ethical to delegate procedural consent to conversational AI?, in *Journal of Medical Ethics*, v. 50, n. 2, Feb. 2024, p. 77-83. DOI: 10.1136/medethics-2023-107482. Available at: <https://jme.bmj.com/content/50/2/77>. Accessed on 26 Nov. 2025.

²¹ OKAMOTO, S. et al. AI-based medical ethics education: examining the potential of large language models as a tool for virtue cultivation, in *BMC Medical Education*, v. 25, n. 1, Feb. 2025, p. 185.

Of course, human beings also make mistakes, and are also occasionally subject to hallucinations — even doctors. It is important to note, however, that what we call a hallucination in a human being is not the same as what happens in an AI in terms of the underlying neurological processes. But both AI and humans can give incorrect answers or recommendations. As human beings, we are aware of each others' fallibility, and have evolved methods to identify motivations for certain behaviours as well as techniques for identifying when something is going wrong. A doctor might be motivated by the consensus in the literature, by regulatory guidelines, by the wish to save time, or any number of things. All these motivations make sense given certain contexts or pressures. A smart patient will be aware of this and may — given certain cues — wonder if the doctor is giving accurate or unbiased information. Moreover, if a doctor is genuinely hallucinating, they might talk incoherently, or stagger, or offer such wildly inappropriate suggestions that the patient would be alerted to the fact that something is wrong. A hallucinating doctor would be very unlikely to provide a lucid explanation for her suggestion. All of this would rightly make the patient concerned that in this instance the doctor's advice might not be reliable.

The particular risks associated with AI hallucinations or misinformation, however, cannot be mitigated in this way. A hallucinating AI, or one which has based its outputs on a flawed interpretation of the data will present its decision with just the same neutral confidence as it does at any other time, and its 'motivations' are not necessarily transparent to the patient. And because of the black box phenomenon, we will not necessarily be able to use explanation as a basis on which to establish that a mistake is being made. Perhaps ultimately, we will gain the skills to identify signs that suggest an AI is hallucinating, or providing inaccurate information just as we do with humans. In the meantime, the risk of errors cannot be ruled out. It is important to note here that there is considerable research ongoing to make AI explainable. In principle this would solve the issue of the black box, and offer a way of identifying cases of hallucination. However the methods being explored are themselves based on models and hypotheses, which one might accept or not. In fact, different methods that explain AI decisions might easily reach conflicting conclusions.²²

Finally, as we have noted, in addition to the black box problem, there is another transparency issue in relation to AI. That is the question of goals and rewards. We discuss this further in our section below. But for now the key issue is to note that transparency is an issue *both* in terms of explaining how the algorithm came to its decision *and* in terms of our awareness of the goal and reward system programmed in during training.

²² MITTELSTADT, B.; RUSSELL, C.; WACHTER, S. Explaining explanations in AI. In: *Proceedings of Conference on Fairness, Accountability, and Transparency (FAT '19)*. New York: ACM, 2019, p. 279-288.

Accordingly, it seems again that the human doctor is preferable. Not only can she take John's own goals into account, but she can also explain *how* the drug works, in what respects it is better than the alternatives, and ultimately, how confident she is that it will have the desired effect. On all three of these measures, it seems that the AI is the loser.

But is it really the case that the doctor can provide a better explanation for her recommendation than the AI can? As discussed above, any doctor will have several criteria in mind, when taking her decision. The doctor may feel reluctant to admit some of these criteria to the patient. For example, if the health service regards drug A as being preferable, but drug B is cheaper. Or if drug A is preferred because of a deal made with a drug company. We do not usually like to think our doctors are influenced by these things, but it is by no means unfair or unrealistic to suppose that they are: this is normal life for a doctor. Moreover, perhaps not all doctors themselves recognise the degree to which they are influenced by heuristics, by hearsay, pressure from the health service, trends in medical practice, etc.

The doctor may not understand her own modes of reasoning very clearly. She may arrive at different conclusions on different occasions even if the salient facts remain the same. If John asks her to explain her recommendation, she will present a narrative that accounts for her decision, but whether that narrative is a reliable representation of the process by which she got there is another question. Moreover, the doctor may not *want* to share all of her reasoning with John. Unlike the machine, the doctor is a human, who may feel conflicted about the need to incorporate cost issues in her recommendation, for example. She may feel – rightly, perhaps – that to do so would undermine the trust that John has in her.

The ways in which doctors – or other human beings – juggle different options, and ascribe weights to particular outcomes is far from simple. In theory, the fact that we can ask a doctor for an explanation helps to avoid several kinds of risk: that a person relies on mere intuition or prejudice or habit; that they jump to conclusions, or that they make mistakes. But humans are astonishingly good at what is called “confabulation”²³ – in this context, making up a narrative to explain a decision that has already been made.²⁴ Thus, many decisions are not actually the result of a rational weighing up of factors; they are simply a gut feeling accompanied by a confabulation.

²³ There are similarities between confabulation and “post hoc reasoning”, but confabulation is a broader phenomenon based more on the creation of narrative explanations, rather than on faulty causal inferences per se.

²⁴ HAIDT, J. The emotional dog and its rational tail: a social intuitionist approach to moral judgment, in *Psychological Review*, v. 108, n. 4, Oct. 2001, p. 814-834. DOI: 10.1037/0033-295X.108.4.814.

This confabulation is a narrative that contributes to the suspension of disbelief we have when it comes to our self-perception as rational decision-makers. And it is this self-perception that seems to differentiate us from computers and AI algorithms. Perhaps we might hope that highly educated people such as doctors are less inclined to rely on confabulation, but experiments suggest that highly-educated individuals do not confabulate less than others; they are just particularly good at sounding plausible.²⁵ Accordingly, the doctor's explanation to John about why drug X is right for him, may be more akin to a pronouncement by a large language model such as chatGPT than to an accurate and rational explanation of how her conclusion was reached.

If this is right, the black box nature of medical algorithms is not essentially more of a threat to patient autonomy than the opacity of the human mind itself. Earlier, we noted that attempts to describe what is happening in the AI black box are unsatisfactory in various ways, including the possibility of arriving at different explanations of the same phenomena. At first glimpse this sounds unacceptable. But given the human tendency to confabulation, perhaps what we are seeing in the AI context is not so very different!

So far, we have seen that there are several differences between AI treatment recommendations, and those made by a human doctor. These differences may be significant in terms of their implications for patient autonomy. AI cannot incorporate in-the-moment updates, and therefore setting individualised goals for patients is not possible. Yet many of the goals embedded in RCTs and in systems such as QALYs are also inflexible to the individual. AI cannot give reliable explanations for its decisions. Yet one might also call into question the degree to which doctors' explanations are genuine accounts of a decision-making process, or a construction of hastily assembled reasons that seem to weigh in favour of a decision, but may not have been involved in the making of that decision.

5. Motivation and rewards

There is another respect in which medical AI differs from the doctor. That is, even if we agree that we may not always get a rational explanation for a doctor's recommendation, we are justified in trusting the doctor because they are well-motivated. Doctors aim to *benefit* their patients. They care about their them: empathy is a specific attribute that is looked for, and tested

²⁵ SCHWITZGEBEL, E.; CUSHMAN, F. Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers, in *Mind & Language*, v. 27, n. 2, Apr. 2012, p. 135-153. DOI: 10.1111/j.1468-0017.2012.01438.x.

for among medical students, and is regarded as a defining feature of the doctor/patient relationship.²⁶ We may not know precisely how our doctors' decisions are reached, and they might not know either — but perhaps this is not really what matters. If we know they have our welfare at heart, we can be satisfied in knowing that is the goal the doctor is aiming for.

In contrast, the AI algorithm is motivated only by the rewards it receives when correctly fulfilling the optimisation tasks programmed into it during the training phase. It does not know or care about John as a patient.²⁷ The AI may be rewarded for identifying the cheapest treatment, the treatment most likely to achieve 5 years' survival, the treatment most likely to yield the lowest number of circulating cancer cells, or the treatment least likely to cause nausea, pain or insomnia. In many cases, as noted above, human users of AI are not aware of what reward it has been programmed with, or even that rewards are a feature of medical AI in this respect. Because of this, it seems that we have reason to be extremely cautious about the role of medical AI, especially insofar as it relates to patient autonomy. If John does not know whether the AI is prioritising survival, cost-effectiveness, avoidance of surgery, reduction of circulating cancer cells, etc, he cannot know if its recommendation is right *for him*. More worryingly, neither can the doctor, if she also lacks access to this information.

But rewards are at play in the case of the human doctor too in her recommendation. At the most basic level, she gets paid for what she does. In some cases, she will be influenced by additional financial rewards. For example, she might get 'freebies' from drug companies, designed to influence her likelihood to recommend their product. Alternatively, she might get rewards from the healthcare system for providing certain kinds of treatment. In the UK, for example, there are financial incentives to encourage GPs to undertake vaccinations, measure patients' blood pressure, prescribe statins, etc.²⁸ None of these are straightforwardly about doing what is best for 'this patient'; they are designed to reward the doctor for meeting certain targets, just as the algorithm is rewarded for meeting its targets.

Of course a good doctor will *also* be motivated by benefitting her patient. But these rewards may pull in different directions; the pharmaceutical company one way, the government incentives another, patient benefit in a third direction. It may be difficult, if not impossible to sift through all

²⁶ BEAUCHAMP, T. The principle of beneficence in applied ethics. In: ZALTA, E. N. (Ed.). *The Stanford Encyclopedia of Philosophy*. Spring 2019 ed. Stanford: Metaphysics Research Lab, Stanford University, 2019. Available at: <https://plato.stanford.edu/archives/spr2019/entries/principle-beneficence/>. Accessed on 26 Nov. 2025.

²⁷ ZHUANG, S.; HADFIELD-MENELL, D. Consequences of misaligned AI. In: *Proceedings of Advances in Neural Information Processing Systems* 33. 2020, p. 15763-15773.

²⁸ SLOWTHER, A.; FORD, S.; SCHOFIELD, T. Ethics of evidence-based medicine in the primary care setting, in *Journal of Medical Ethics*, v. 30, n. 2, Apr. 2004, p. 151-155. DOI: 10.1136/jme.2003.004150.

these. On top of all this, it is precisely because she is a human, with feelings, that the doctor is subject to more negative emotional influences. She may find John irritating, in ways that influence her decision-making. Even if the machine cannot care about John's wellbeing in the way that a good doctor might, the upshot of this is that it cannot find John annoying or dislike him, in ways that impact its functioning. (This is not to deny that the machine can, of course be biased in another sense, if it was trained on data that is not adequately representative of John and his particular needs.^{29,30} Equally, John's doctor may be biased against him because of his ethnicity or religion, or simply because he is a troublesome patient.

Doctors may be irritable, tired, bored, rushed, and any number of other things, in addition to being prejudiced and biased. This is what it is to be human, and it is both unreasonable and unwise to suppose that doctors are above this. Interestingly, however, there is some evidence that medical AI can be perceived as being more empathetic in some contexts, than real doctors.³¹ Clearly there is much more research to be done here. However it is perhaps precisely because AIs do not get irritable, tired, or bored, or feel challenged, or defensive, that they are able to achieve a stable, cheery mode of communication that simply is not possible for human beings in real world contexts.

Yet part of what makes the doctor patient relationship special is the idealised vision of pure benevolence that we tend to invest doctors with. And when we think about the role of the placebo effect in the doctor/patient relationship, it seems clear that the ideal we have of the therapeutic relationship can be therapeutic *in itself* independent of any additional scientific factor, and perhaps independent of what the doctor actually feels or intends. It is a fascinating question for future research, to consider whether the same placebo effect could arise in conjunction with AI medicine.

Conclusion

Given the challenges we have discussed, it seems unlikely that relying solely on AI recommendations would be an effective way of respecting patient autonomy, at least at the present time. We could, perhaps in the

²⁹ GAONKAR, B.; COOK, K.; MACYSZYN, L. Ethical issues arising due to bias in training AI algorithms in healthcare and data sharing as a potential solution, in *AI Ethics Journal*, v. 1, n. 1, 2020, p. 1-13.

³⁰ GIOVANOLA, B.; TIRIBELLI, S. Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms, in *AI & Society*, v. 38, n. 2, Apr. 2023, p. 549-563. DOI: 10.1007/s00146-021-01298-5.

³¹ AYERS, J. W.; POLIAK, A.; DREDZE, M. et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum, in *JAMA Internal Medicine*, v. 183, n. 6, Jun. 2023, p. 589-596. DOI: 10.1001/jamainternmed.2023.1838.

future attempt to create an algorithm that maximally supports patient autonomy, rather than – say – maximising the period of remission. This would require a representation of potential individual patients' wishes during the training phase, which the algorithm can take into consideration. Ultimately, the development and training of one AI algorithm per patient, each patient with a private AI counsellor. Training of the private AI algorithms would then reflect the wishes (translated in rewards when satisfied by the algorithm) of John. This would be pushing the principle of autonomy as far as it can go.

Yet such an endeavour would be hugely costly and complex. Moreover, it would seem to require an idealised respect for autonomy that is neither expected nor delivered in human doctor/patient interactions today. In all the contexts we have discussed where the human seems to have the advantage, it is an *idealised* conception of the doctor: one who has time and inclination to attend to John's individual wishes and needs. In practice, most health services do not facilitate this kind of attentive care. Perhaps they should, but this would come at the cost of other goods that we perceive as valuable in the health context. In effect, health services are already struggling with many of the difficulties in regard to patient autonomy that are being characterised as specific to medical AI. RCTs cannot cater to every individual's preferences in terms of treatment goals. Opacity is built into the system in many respects.

However, with the advent of medical AI we can at least try to identify those aspects that pose ethical challenges in advance, and try to ensure they do not become fixed and embedded in the system. The most serious aspects of the ethical work relating to medical AI involve assumptions that go into the design and programming of the systems well before they have an impact on any specific patient. Not only this, but the goals and rewards involved are not commonly recognised as raising significant ethical issues. They are identified and implemented by people whose expertise and motivation may be utterly different from those of the people who will be affected by them. John does not take part in the production of the AI algorithm that might decide his treatment. What this means is that AI is a very blunt tool in terms of facilitating patient autonomy or ethical decision making in medicine.

But we need not ditch the possibility of ethical AI in medicine altogether. One way in which patient autonomy *can* be respected, in the context of AI, is to ensure that patients are aware of the ways in which algorithms have contributed to the decision-making process. Autonomy is not simply a question of doing what the patient wants. It is a question of ensuring that the patient is fully informed. Thus, an ethically responsible way of using an algorithm would involve a discussion with the patient and the doctor, taking into account the algorithm's recommendation, in the context of all

the other relevant factors, and while understanding its inner workings. That is to say, for John to decide whether to accept the AI's treatment recommendation, he would need to know that the AI is only looking for 5 year survival. Once he is aware of this, he can make an autonomous decision about whether this aligns with his own goals.

To enable algorithms to contribute to a fully-informed decision making process, it is necessary for designers to have this aim in mind when creating the systems. It is necessary for doctors to be aware of the way that rewards are used to create the incentives for the algorithm during training, and to be able to communicate this to the patient. These challenges are not simple. But there is an opportunity to bring clarity and transparency to a field that is currently in development. We can at least design our tools and algorithms in such a way that their workings are compatible with the most important ethical principles associated with medical practice. However, to achieve this, we need to recognise the limitations of AI algorithms as well as their strengths, and ensure that the goals we programme them with are carefully considered, and transparent to those who are affected by them.

Clinical trial number: not applicable

Declarations

- Ethics approval and consent to participate: not applicable
- Consent for publication: not applicable
- Availability of data and materials: not applicable
- Competing interests: no competing interests
- Funding: Norges Forskningsråd, 332645
- Authors' contributions: AS wrote the first draft, MZ and AF contributed thereafter with ideas and arguments and amendments, and all authors reviewed the manuscript.
- Acknowledgements: not applicable

References

ALLEN, J. W.; EARP, B. D.; KOPLIN, J.; WILKINSON, D. Consent-GPT: is it ethical to delegate procedural consent to conversational AI?, in *Journal of Medical Ethics*, v. 50, n. 2, Feb. 2024, p. 77-83. DOI: 10.1136/medethics-2023-107482. Available at: <https://jme.bmj.com/content/50/2/77>. Accessed on 26 Nov. 2025.

AYERS, J. W.; POLIAK, A.; DREDZE, M. et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, v. 183, n. 6, Jun. 2023, p. 589-596. DOI: 10.1001/jamainternmed.2023.1838.

BEAUCHAMP, T. The principle of beneficence in applied ethics. In: ZALTA, E. N. (Ed.). *The Stanford Encyclopedia of Philosophy*. Spring 2019 ed. Stanford: Metaphysics Research Lab, Stanford University, 2019. Available at: <https://plato.stanford.edu/archives/spr2019/entries/principle-beneficence/>. Accessed on 26 Nov. 2025.

BEGOLI, E.; BHATTACHARYA, T.; KUSNEZOV, D. The need for uncertainty quantification in machine-assisted medical decision making, in *Nature Machine Intelligence*, v. 1, n. 1, Jan. 2019, p. 20-23. DOI: 10.1038/s42256-018-0004-2.

DELGADO, A.; GUDDATI, A. K. Clinical endpoints in oncology—a primer, in *American Journal of Cancer Research*, v. 11, n. 4, 2021, p. 1121-1131. PMID: 33948349; PMCID: PMC8085844.

DURÁN, J. M.; JONGSMA, K. R. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI, in *Journal of Medical Ethics*, v. 47, n. 5, May 2021, p. 329-335. DOI: 10.1136/medethics-2020-107096.

FERRARIO, A. Design publicity of black box algorithms: a support to the epistemic and ethical justifications of medical AI systems, in *Journal of Medical Ethics*, v. 48, n. 7, Jul. 2022, p. 492-494. DOI: 10.1136/medethics-2021-107482.

FLEMING, T. R.; POWERS, J. H. Biomarkers and surrogate endpoints in clinical trials, in *Statistics in Medicine*, v. 31, n. 25, Sept. 2012, p. 2973-2984. DOI: 10.1002/sim.5403.

GAONKAR, B.; COOK, K.; MACYSZYN, L. Ethical issues arising due to bias in training AI algorithms in healthcare and data sharing as a potential solution, in *AI Ethics Journal*, v. 1, n. 1, 2020, p. 1-13.

GIOVANOLA, B.; TIRIBELLI, S. Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms, in *AI & Society*, v. 38, n. 2, Apr. 2023, p. 549-563. DOI: 10.1007/s00146-021-01298-5.

GROTE, T. Trustworthy medical AI systems need to know when they don't know, in *Journal of Medical Ethics*, v. 47, n. 5, May 2021, p. 337-338. DOI: 10.1136/medethics-2021-107463.

HAIDT, J. The emotional dog and its rational tail: a social intuitionist approach to moral judgment, in *Psychological Review*, v. 108, n. 4, Oct. 2001, p. 814-834. DOI: 10.1037/0033-295X.108.4.814.

HU, X. et al. Do large language models know about facts? arXiv preprint arXiv:2310.05177, 2023.

JIN, Q.; YANG, Y.; CHEN, Q.; LU, Z. GeneGPT: augmenting large language models with domain tools for improved access to biomedical information. arXiv preprint arXiv:2304.09667v3, 2023. Atualizado em: *Bioinformatics*, 2024. PMID: 37131884; PMCID: PMC10153281.

KERASIDOU, C.; KERASIDOU, A.; BUSCHER, M.; WILKINSON, S. Before and beyond trust: reliance in medical AI, in *Journal of Medical Ethics*, v. 48, n. 11, Nov. 2022, p. 852-856. DOI: 10.1136/medethics-2020-107095.

KIRKDALE, R. et al. The cost of a QALY, in *QJM: An International Journal of Medicine*, v. 103, n. 9, Sept. 2010, p. 715-720. DOI: 10.1093/qjmed/hcq081.

LUXTON, D. D. AI decision-support: a dystopian future of machine paternalism?, in *Journal of Medical Ethics*, v. 48, n. 4, Apr. 2022, p. 232-233. DOI: 10.1136/medethics-2022-108243.

MAINZ, J. T. Medical AI: is trust really the issue?, in *Journal of Medical Ethics*, 2023 (Online first, 26 jul. 2023). DOI: 10.1136/jme-2023-109414. Available at: <https://jme.bmj.com/content/early/2023/07/26/jme-2023-109414>. Accessed on 26 Nov. 2025.

MISHRA, A. Transparent AI: reliabilist and proud, in *Journal of Medical Ethics*, v. 47, n. 5, May 2021, p. 341-342. DOI: 10.1136/medethics-2021-107352.

MITTELSTADT, B.; RUSSELL, C.; WACHTER, S. Explaining explanations in AI. In: *Proceedings of Conference on Fairness, Accountability, and Transparency (FAT '19)*. New York: ACM, 2019, p. 279-288.

OKAMOTO, S. et al. AI-based medical ethics education: examining the potential of large language models as a tool for virtue cultivation, in *BMC Medical Education*, v. 25, n. 1, Feb. 2025, p. 185.

PETTITT, D. A. et al. The limitations of QALY: a literature review, in *Journal of Stem Cell Research and Therapy*, v. 6, n. 4, 2016, p. 334.

PETRONI, F. et al. Language models as knowledge bases? arXiv preprint arXiv:1909.01066, 2019.

RAJPURKAR, P.; CHEN, E.; BANERJEE, O.; TOPOL, E. J. AI in health and medicine, in *Nature Medicine*, v. 28, n. 1, Jan. 2022, p. 31-38. DOI: 10.1038/s41591-021-01614-0.

SCHWITZGEBEL, E.; CUSHMAN, F. Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers, in *Mind & Language*, v. 27, n. 2, Apr. 2012, p. 135-153. DOI: 10.1111/j.1468-0017.2012.01438.x.

SINGHAL, K. et al. Large language models encode clinical knowledge, in *Nature*, v. 620, Aug. 2023, p. 172-180. DOI: 10.1038/s41586-023-06291-2.

SLOWTHER, A.; FORD, S.; SCHOFIELD, T. Ethics of evidence-based medicine in the primary care setting, in *Journal of Medical Ethics*, v. 30, n. 2, Apr. 2004, p. 151-155. DOI: 10.1136/jme.2003.004150.

ZHUANG, S.; HADFIELD-MENELL, D. Consequences of misaligned AI. In: *Proceedings of Advances in Neural Information Processing Systems 33*. 2020, p. 15763-15773.