Cambridge Quarterly of Healthcare Ethics (2024), 33: 3, 380–389
doi:10.1017/S0963180122000718

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Reflection Machines: Supporting Effective Human Oversight Over Medical Decision Support Systems

Pim Haselager[1]*, Hanna Schraffenberger[2], Serge Thill[1], Simon Fischer[1], Pablo Lanillos[1], Sebastiaan van de Groes[3] and Miranda van Hooff[3,4]

[1]Donders Institute for Brain, Cognition and Behaviour, Department of AI, Radboud University, Nijmegen, The Netherlands
[2]Information Science, iHub, Radboud University, Nijmegen, The Netherlands
[3]Health Sciences, Radboud UMC, Nijmegen, The Netherlands
[4]St Maartenskliniek, Nijmegen, The Netherlands
*Corresponding author. Email: pim.haselager@donders.ru.nl

Abstract

Human decisions are increasingly supported by decision support systems (DSS). Humans are required to remain "on the loop," by monitoring and approving/rejecting machine recommendations. However, use of DSS can lead to overreliance on machines, reducing human oversight. This paper proposes "reflection machines" (RM) to increase meaningful human control. An RM provides a medical expert not with suggestions for a decision, but with questions that stimulate reflection about decisions. It can refer to data points or suggest counterarguments that are less compatible with the planned decision. RMs think against the proposed decision in order to increase human resistance against automation complacency. Building on preliminary research, this paper will (1) make a case for deriving a set of design requirements for RMs from EU regulations, (2) suggest a way how RMs could support decision-making, (3) describe the possibility of how a prototype of an RM could apply to the medical domain of chronic low back pain, and (4) highlight the importance of exploring an RM's functionality and the experiences of users working with it.

Keywords: human oversight; meaningful human control; decision-making; medical decision support system; human–computer interaction

Introduction

Increasingly, aspects of decision-making are aided or taken over by AI. In domains like warfare, finance, law, healthcare, insurance, and dating, algorithms inform, prepare, or generate hypotheses, diagnoses, behavioral options, or decisions. For instance, decision support systems (DSS) have been introduced to improve medical decisions, by matching the characteristics of an individual patient to a clinical knowledge base, and by providing patient-specific assessments or recommendations to support the clinician in reaching a decision. In terms of scale and speed, these systems can utilize data and observations outside human reach.[1,2]

Questions have been raised about how humans could remain in control over the overall process and be responsible for its outcomes.[3,4] As different people interact with an AI system (e.g., developing it, operating it, being subject to its outcome), each requires a different set of solutions to exercise control. For humans affected by decisions involving a DSS, Article 22(1) of the EU General Data Protection

Resubmission to the special issue "Ethical Implications of AI Use in Medicine" in Cambridge Quarterly of Healthcare Ethics (Guest Editors: Orsolyo Friedrich and Sebastian Schleidgen). The text is based on a PhD proposal by the same authors (except S.F.) for the Donders Centre of Cognition of the Radboud University in 2022.

Regulation (GDPR) states that a "data subject shall have the right not to be subject to a decision based solely on automated processing."[5] Further, Article 14.4 (b-d) of the AI Act states that humans using a DSS to make a decision should be empowered to "remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (i.e., 'automation bias')."[6] It goes on to specify that humans should be able to interpret, accept, disregard, override or reverse DSS outputs. Similarly, the EU High-Level Expert Group[7] has emphasized the importance of human oversight for trustworthy AI, for instance by having humans "in-the-loop" (i.e., humans monitoring the system's operation and intervening during the decision cycle).

It is, however, not clear whether humans can provide the type, amount, and consistency on supervision over longer periods of time, such that it would amount to effective human oversight of machine contributions to decision-making. Rather, due to general psychological reasons, human attention and concentration may wax and wane, such that effectively being in the loop might not be possible in practical.[8] Various factors such as tiredness, recklessness, boredom, or a lack of attention can play a role in automation bias. There is thus a considerable risk that humans become overly reliant on DSSs,[9,10] leading to deficiencies in "meaningful human control"[11,12] and potentially raising "responsibility gaps."[13,14] Overall, a human in the loop does not ensure that effective human oversight will be exerted to the extent required for moral and legal responsibility. Rather, humans might end up being "under" the loop,[15] merely playing a symbolic role by providing formal "stamps of approval" without genuine reflection.

Therefore, this paper proposes and explores the concept of a "Reflection Machine" (RM): an additional computational system to support effective and meaningful human oversight over a DSS. Cornelissen et al.[16] have recently introduced a first technical proof-of-concept implementation. This paper therefore focuses on how RMs provide feedback on joint human–DSS decisions and urge the human to negotiate the proposed decision, thereby increasing human involvement in the decision-making process. RMs improve human oversight by asking questions about the reasoning behind accepting or rejecting a recommendation. In other words, whereas a DSS thinks "for" the human, the RM thinks "against" them. One way for the RM to do so would be to indicate data that support an alternative option other than the one recommended by the DSS. The questions raised by the RM add friction[17] and thereby prevent mindless decisions and instead promote deliberate and reflective decision-making.

Thinking "against" requires more time and effort of a human. Hence, an effective design of an RM and an appropriate balance between the activities of the DSS and RM will be important topics for research. The growing awareness of the potential risks of AI has led to a substantial increase in ethical, political, and legal codes and regulations within the EU. The increased attention, however, has not led to an unequivocal and practically precise set of instructions for the design and development of applications.[18,19] Besides, legislations such as the GDPR and the AI-Act are inspired by philosophical, political, and legal considerations, but do not explicitly take psychological mechanisms of decision-making into account. Hence, further analysis is required in order to determine what "effective oversight" in human–machine decision-making amounts to, and which RM features contribute to enabling human individuals to reflect on, disregard, or override the output of a machine.

Introducing Reflection Machines in Low Back Pain Medical Decision Support Systems

While specific RMs could be used in different contexts and by different people, we introduce a (currently hypothetical but soon to be implemented) RM used in the medical context by a physician who also uses a DSS to treat low back pain (LBP) in patients.

LBP accounts for more years lived with disability worldwide than any other health condition.[20] In the Netherlands, approximately 44% of the population experiences at least one episode of LBP in their lifetime, with one in five reporting persistent back pain lasting longer than three months (chronic low back pain, CLBP).[21] CLBP results in substantial limitations in activities and leads to high healthcare and socioeconomic costs.[22,23]

The treatment for CLBP is still debated and there is a wide variance in treatments available, ranging from standard physiotherapy to a combined physical/psychological (CPP) program and surgery. In the vast majority of patients with CLBP (85–90%), the etiology is unknown[24] and for medical specialists, it is challenging to identify patients who would benefit from surgical or non-surgical interventions. The etiology can be very different and the psychological coping strategies of the patient have a huge impact on the treatment outcome. The proposed treatment by the orthopedic surgeon is heavily dependent on the presentation of the patient and the experiences of the surgeon.

To reduce extreme variability in diagnosis or proposed treatment, a DSS has been developed, namely, the Nijmegen decision support tool for chronic low back pain (NDT-CLBP).[25] The example in Figure 1 shows the system's output for a patient likely to benefit from surgery, which potentially means the patient is referred to the spinal surgeon for consultation. The NDT-CLPB consists of (1) questionnaires that patients complete when they are referred to secondary care (in the Sint Maartenskliniek), (2) patient outcomes registry, and (3) formulae for calculating outcome predictions. The formulae are based on successful (responder) and disappointing (non-responder) outcomes one year after treatment. The tool supports shared decision-making between patient and physician based on patient profiles (patient characteristics related to treatment outcomes) and matches patients, based on questionnaires, to the treatment that they are most likely to benefit from.[26] The treatment options are spine surgery, conservative combined psychological and physical pain self-management program (CCP program) and no treatment in secondary care (meaning counseling during consultation and physiotherapy in primary care). Patients are referred to either spinal surgeon consultation or non-surgical consultation.

The current version of the NDT-CLBP is rated as very helpful by orthopedic surgeons as it gives extra information. However, most of them are aware that the result might also push toward tunnel vision. Poor outcome of surgery is seen frequently and there can be a bias towards non-surgical treatment in most cases. Moreover, a poor prediction of surgical treatment by the NDT-CLBP aggravates that bias, limiting an objective evaluation of the patient by the orthopedic surgeon. Eventually, the DSS will also limit the accessibility of orthopedic surgeons, since non-surgical consultations will be conducted by physician assistants, who have limited knowledge about surgical treatment indications. Therefore, once a patient is on a path of non-surgical treatment, it is difficult for them to be redirected. To overcome such problems, we introduce reflection support by means of an RM. This complementary system maintains or stimulates

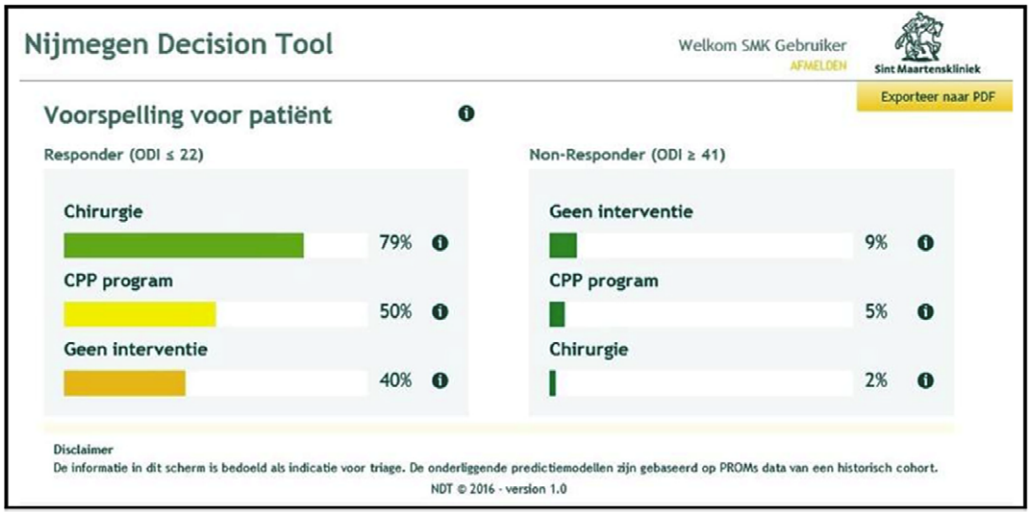

Figure 1. A screenshot from the NDT-CLBP as presented to the physician during the patient's visit (Voorspelling voor patient: Prediction for patient; CPP: Combined Physical and Psychological program; Chirurgie: surgery; Geen interventie: no intervention). The bars represent the likelihood that a specific patient will be "a responder" or a "non-responder" for each of the specified treatments. Responder is defined as a patient-acceptable symptom state and non-response as severe disability and persistence of LBP.[27]

active human involvement in machine-supported decision-making, such as required by the EU's ethical and legal codes.

The following case may serve as an illustration:

> *A male patient, age 40, suffers from LBP for over a year, physiotherapy was not effective. He is using morphine but cannot work, still needs help to get dressed, and feels depressed. The pain gets worse after 500 meters of walking, spreading to both legs. After some rest, the leg pain resolves. He has problems walking with an upward posture and tends to lean forward. Walking downhill is more problematic than walking uphill. Physical examination by the general practitioner reveals no neurological deficits and the GP refers the patient to an orthopedic surgeon. The treatment for CLBP is still under debate and there is a wide variance in treatments available ranging from standard physiotherapy to a combined physical/psychological (CPP) program and even surgery. The proposed treatment by the orthopedic surgeon heavily depends on the presentation of the patient and the experiences of the surgeon. The patient filled in all questionnaires and the NDT-CLBP estimated benefit from surgery as 28%. The orthopedic surgeon discussed all items with him, and he was sent to the CPP program. Although he became more positive, his LBP remained persistent. Fortunately for the patient, one of the physiotherapists in the CPP program thought of the option of the patient having a spinal stenosis. A new consultation by the orthopedic surgeon, with this specific question, resulted in an MRI scan which confirmed it. The patient was treated surgically and was relieved of all symptoms after three months.*

Based on historical patient data, the NDT-CLBP in the scenario described above led the orthopedic surgeon to non-surgical treatment for the patient. Thereby, the surgeon became less aware of the symptoms that fit the diagnosis of spinal stenosis, which can be successfully treated by surgery. If a RM were in place, it could have redirected the surgeon toward a non-psychological diagnosis, that is, the spinal stenosis, as reason for the LBP. Ideally, the DSS itself would have suggested surgical treatment. However, given the variety of different factors for CLBP, many of which cannot be adequately quantified, each case must be treated individually.[28] This is not to discount the usefulness of DSS, but the output should not be considered as universally applicable. In other words, lesser-known cases for which there is little or no data cannot be adequately addressed by the DSS. The RM thus urges the doctor to re-evaluate and re-consider the suggestion of the DSS. Figure 2 visualizes the basic contours of such a decision-making process (joint human–machine DSS & RM), in which the questions of the RM (in parallel to the

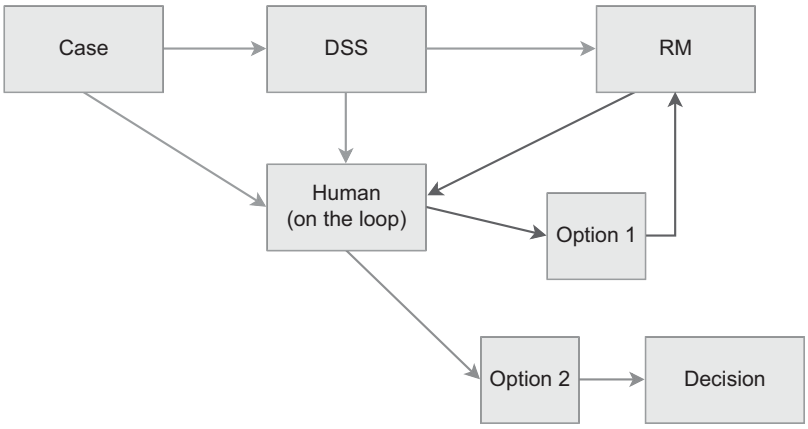

Figure 2. Simplified framework for a joint human–machine (DSS & RM) decision-making process. Case information & the DSS recommendation suggest the physician to proceed with option 1 (e.g., no surgical intervention), but upon RM questioning the physician may switch to option 2 (e.g., surgery) as the final decision, or re-affirm, with more trust based on increased reasoning, option 1.

physiotherapist in the case described above) increase the reflection of the orthopedic surgeon, which could influence the choice between the two options.

Reflection Machine

An RM can be described as a system that receives information about the medical situation, the NDT-CLBP's recommendation, and (optionally) the physician's behavior (e.g., reflection time, decision style, decision history, and preference for an option) as input. Based on this information, the RM can then produce output in the form of questions that prompt the physician to reflect on the decision more deeply. A core aspect is the identification of appropriate prompts to generate reasonable questions. This is similar to the problem of generating reasonable explanations in XAI[29] and often involves counterfactual reasoning. Counterfactuals feature prominently in human reasoning and communication about decisions.[30,31] Hence, an RM ideally fosters epistemic certainty about a suggestion by a DSS, which in turn promotes trust between doctor and patient.[32]

To propose relevant questions, the RM can explicitly take into account: (1) the weakest evidence for the accepted decision, (2) the strongest evidence for the alternative decision, and (3) missing evidence that would have had a significant impact on the recommended decision (based on unperformed but potentially useful tests). The relation between evidence and hypotheses will be formalized with a probabilistic approach,[33] distinguishing weak from strong evidence, given a decision, through model selection.[34] A variables significance analysis will identify which variables explain the decision made and select important evidence (focused on observations that have higher information gain).

RMs and Human Decision-Making

Although the focus of this paper is primarily to introduce and elucidate the basic idea of an RM, there is a significant amount of fundamental research in the cognitive neuroscience and psychology of human decision-making that will be useful to extend the simple framework provided in Figure 2. To illustrate the possibilities for extension of the framework, we will briefly review some important literature on four topics relevant to the implementation and/or application of an RM, viz. factors influencing decisions, evidence accumulation, neural mechanisms, and surrogate decision-making. Although this research on decision-making is clearly informative, it must be kept in mind that the contexts of such studies are most often social interaction or consumer choices, using for example, investment games with personal gains or losses as tasks. Hence, a direct translation to a medical context is to be treated with caution.

First, much is known about factors that influence (for better or worse) human decision-making. In a review paper, Saposnik et al.[35] indicated that overconfidence, lower tolerance to risk, anchoring effects, and information and availability biases are associated with diagnostic inaccuracies in 36.5 to 77% of case scenarios. It seems fair to say that at least part of the attractiveness of DSSs derives from their potential to remedy (avoid or compensate for) such flaws in human decision-making. Croskerry[36] suggests that six major clusters of factors can influence the clinical reasoning during the diagnostic process. These range from individual characteristics of the decision maker to factors in the work environment to factors associated with the patient. Although the emphasis of the paper is to provide a general introduction to the main ideas of an RM, knowledge about aspects influencing decision-making will be very informative in the implementation of the prototype.

Second, decision-making takes place over time and often involves the weighing of incoming information. Busemeyer et al.[37] focus on decisions involving multi-attribute, multi-alternative choices, to be made under risk and uncertainty, where, over time, evidence accumulation takes place. Noguchi and Stewart[38] present a sequential sampling version of a multi-alternative decision model. It provides a useful framework for understanding how incoming partial evidence (such as empirical evidence or test results) can be sequentially weighed and compared in terms of their support for alternative hypotheses or diagnoses. Moreover, models of evidence accumulation have been found to provide "remarkably

accurate" descriptions for the neural dynamics of decision-making (albeit currently mainly in animal studies).[39]

Third, knowledge about the neural mechanisms will be useful in the attempt to understand the why and how of the effects of the RM on the persons involved in making the decision. Rilling and Sanfey[40] review the neural mechanisms underlying social decision-making, for instance in relation to trust (see also van Baar et al.[41]). Park et al.[42] investigate the neural mechanisms involved in making group decisions where the outcome of one's decision can depend on the decisions of others.

Finally, in clinical decision-making there is a difference between the decision makers (physicians) and the ones undergoing the consequences of the decision (patients). Füllbrun et al.[43] edited a special issue focusing on decision-making for others. They suggest that it is important to consider the "psychological distance" between decision maker and recipient. They refer to a model proposed by Tunney and Ziegler[44] that captures four different perspectives that together can influence decision makers in their ultimate choice. The final decision is the weighted result of a combination of egocentric (what do I want), projected (what would I do), benevolent (what should I do), and simulated (what would you do) perspectives. Several factors (intent, significance, accountability, calibration, and empathy) determine the weights these perspectives are given in a concrete case, thereby increasing or decreasing the psychological distance between decision maker and recipient. Taken together, these insights can help to identify optimal points of RM intervention in the decision-making process.

From an ethical perspective, the implications of the RM can best be understood along the lines indicated by the High-Level Expert Group on AI of the EU.[45] In their ethical guidelines for trustworthy AI, the first of seven key requirements concerns human agency and oversight. As they say:

> Users should be able to make informed autonomous decisions regarding AI systems. They should be given the knowledge and tools to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system. AI systems should support individuals in making better, more informed choices in accordance with their goals.[46]

We suggest that the RM precisely captures the idea of a tool that makes it possible to "reasonably self-assess or challenge" a DSS. In so doing, the ethical requirement of effective human oversight is fulfilled to, at least to a higher degree, in that an RM "helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects."[47]

Clearly, the overall framework of a joint human–machine decision-making process is complicated. First of all, there is information about the patient, historical cases and general medical knowledge. Second, there are the features of the human involved in decision-making (decision style, expertise level). Third, there are two AI systems (a DSS and an RM). The timing and frequency of information exchange between these elements will need to be specified and the standard flow of information processing and decision-making will need to be analyzed, using the above insights derived from psychology and cognitive neuroscience regarding decision-making in order to derive optimal points for RM intervention. Finally, the usefulness of an RM needs to be explored. Preliminary investigations[48,49] suggest that usefulness needs to be investigated in at least three ways. First, does the RM increase effective human oversight over the machine-supported decision process? This involves measuring potential overreliance on the DSS.[50,51] Overreliance is worrisome insofar as it leads to culpability gaps should an erroneous suggestion be followed.[52,53] The RM tries to counter this problem by urging the human expert to make a more intentional decision, ideally foreseeing possible outcomes. Nevertheless, an RM is itself a new actor in the decision-making process, and in its wake new human actors are introduced in the chain of decision-making––namely, the designers of the RM. Consequently, the danger of new culpability gaps emerges. It is thus important, when building an RM, to avoid an overly complex model that leads to another "black box." Instead, the aim of the RM is to ideally provide or assist in the creation of epistemic insights into the workings of the DSS by reducing its opacity, and by strengthening the explicit arguments used by the human in making the final decision. So, although the RM requires a certain level of domain knowledge, a balance between simplicity and complexity must be found.

Second, does the RM increase the quality of decision-making, in terms of accuracy and efficiency? This requires an analysis of false positives and false negatives, as well as decision time measurements.[54] Relatedly, the question of whether the RM contributes to the patient's well-being arises. Successful treatment of CLBP relieves suffering, and patients have presumably more trust in doctors if they can explain their decision in ways that go beyond repeating the DSS recommendation. Patient interviews will be necessary to establish this.

Third, how does the RM affect user (i.e., the medical expert) experience?[55] Here, the focus will be on the understandability of RM interventions, interference with workflow, user confidence in selected options, sense of agency regarding decision-making, and satisfaction levels regarding the overall decision-making process.[56] A variety of conditions needs to be controlled for, for example, level of medical expertise, automation experience, and automation expectancy.[57,58] Recent cases of clinical practice with CLBP will be used. In the end, an RM should not simply be a "technological fix" (i.e., solutionism). Rather, context-specific knowledge from involved stakeholders must inform the design of the RM, which is part of a broader socio-technical system, to also anticipate potential harms.

Conclusion

Effective human oversight of DSS is one of the major societal challenges posed by AI. The need for maintaining or increasing meaningful human control over machine recommendations, decisions, or actions has been expressed in several recent EU codes, regulations, and acts. In addition, DSSs bring the risk of "hollowing out" of professional skills.[59,60] By taking over much of the knowledge consultation and inferencing involved in decision-making, human professionals are at risk of losing their high-level skills, and moreover may experience (and resist) being side-lined. RMs offer the possibility to develop a "best of both worlds" approach, increasing opportunities for responsible decision-making without accountability gaps.

An RM will be relevant for the assessment of work disability for social security benefits.[61] In the Netherlands, the Dutch social security institute, UWV, performed 155,900 such assessments ("invaliditeitskeuringen") in 2020.[62] A recurring problem is the consistency of assessments over time, doctors, and patients. The RM could help physicians make consistent and thoughtful diagnoses in a complex domain while maintaining human authority and accountability and ensuring the efficient use of AI resources in joint decision-making. The RM would thus also improve patient well-being.

Finally, the idea of an RM is applicable to other domains. The application of DSSs and the various issues surrounding human oversight, accountability, and professional expertise is not restricted to the medical domain, but encompasses legal, financial, and policing areas as well. The significant growth in the usage of DSS emphasizes the urgent need for effective human oversight. Overall, RMs could mitigate the harms of over-relying on DSS, as, for example, is the case with wrongful arrests through DSS in law enforcement. The development of RM architectures could therefore stimulate the development of trustworthy intelligent systems that is currently at the core of the EU approach to AI.

Notes

1. Sim I, Gorman P, Greenes RA, Haynes RB, Kaplan B, Lehmann H, et al. Clinical decision support systems for the practice of evidence-based medicine. *JAMA* 2001;**8**(6):527–34. doi:10.1136/jamia.2001.0080527.
2. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: Benefits, risks, and strategies for success. *NPJ Digital Medicine* 2020;**3**(17):1–10. doi:10.1038/s41746-020-0221-y.
3. Danaher J. The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology* 2016;**29**(3):245–68. doi:10.1007/s13347-015-0211-1.
4. von Eschenbach WJ. Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology* 2021;**34**(4):1607–22. doi:10.1007/s13347-021-00477-0.

5. European Union. *General Data Protection Regulation of the European Union*; 2018; available at https://gdpr-info.eu/ (last accessed 20 December 2022).
6. European Union. *Proposal for a Regulation of the European Parliament and the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligece Act) and Amending Certain Union Legislative Acts. ACTS. COM/2021/206 Final*; 2021; available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (last accessed 20 December 2022).
7. EU High Level Expert Group on AI. *Ethics Guidelines for Trustworthy AI*; 2019. doi:10.1017/9781108936040.022.
8. Merritt SM, Ako-Brew A, Bryant WJ, Staley A, McKenna M, Leone A, et al. Automation-induced complacency potential: Development and validation of a new scale. *Frontiers in Psychology* 2019;**10**:1–13. doi:10.3389/fpsyg.2019.00225.
9. Van der Stigchel B, van Diggelen J, Haselager WFG, van den Bosch K. Intelligent decision support in medical triage: Are people robust to biased advice? *Journal of Public Health*, forthcoming.
10. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety* 2019;**28**(3):231–7. doi:10.1136/bmjqs-2018-008370.
11. de Sio FS, van den Hoven J. Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI* 2018;**5**:15. doi:10.3389/frobt.2018.00015.
12. Mecacci G, Santoni de Sio F. Meaningful human control as reason-responsiveness: The case of dual-mode vehicles. *Ethics and Information Technology* 2020;**22**(2):103–15. doi:10.1007/s10676-019-09519-w.
13. Santoni de Sio F, Mecacci G. Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology* 2021;**34**:1057–84. doi:10.1007/s13347-021-00450-x.
14. Matthias A. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 2004;**6**(3):175–83. doi:10.1007/s10676-004-3422-1.
15. Liu HY. The power structure of artificial intelligence. *Law, Innovation and Technology* 2018;**10**(2):197–229. doi:10.1080/17579961.2018.1527480.
16. Cornelissen NAJ, van Eerdt RJM, Schraffenberger HK, Haselager WFG. Reflection machines: Increasing meaningful human control over decision support systems. *Ethics and Information Technology* 2022;**24**:19. doi:10.1007/s10676-022-09645-y.
17. Cox AL, Gould SJ, Cecchinato ME, Iacovides I, Renfree I. Design frictions for mindful interactions: The case for microboundaries. In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM: New York, NY; 2016:1389–97.
18. Roig A. Safeguards for the right not to be subject to a decision based solely on automated processing (Article 22 GDPR). *European Journal of Law and Technology* 2017;**8**(3):1–17; available at https://ejlt.org/index.php/ejlt/article/view/570 (last accessed 20 December 2022).
19. Wachter S, Mittelstadt B, Floridi L. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law* 2017;1–47. doi:10.2139/ssrn.2903469.
20. Hoy D, March L, Brooks P, Blyth F, Woolf A, Bain C, et al. The global burden of low back pain: Estimates from the Global Burden of Diseases 2010 study. *Annals of the Rheumatic Diseases* 2014;**73**:968–74.
21. Picavet H, Schouten J. Musculoskeletal pain in the Netherlands: Prevalences, consequences and risk groups, the DMC3-study. *Pain* 2003;**102**:167–78.
22. van Tulder M. Health technology assessment (HTA) increasingly important in spine research. *European Spine Journal* 2011;**20**:999–1000.
23. Lambeek LC, van Tulder MW, Swinkels IC, Koppes LL, Anema JR, van Mechelen W. The trend in total cost of back pain in the Netherlands in the period 2002 to 2007. *Spine (Phila Pa 1976)* 2011;**36**:1050–8.
24. Haldeman S, Kopansky-Giles D, Hurwitz EL, Hoy D, Mark Erwin W, Dagenais S, et al. Advancements in the management of spine disorders. *Best Practice & Research. Clinical Rheumatology* 2010;**14**:167–79.

25. van Hooff ML, van Loon J, van Limbeek J, De Kleuver M. The Nijmegen decision tool for chronic low back pain. Development of a clinical decision tool for secondary or tertiary spine care specialists. *PLoS One* 2014;**9**(8):1–12. doi:10.1371/journal.pone.0104226.
26. van Hooff ML, van Dongen JM, Coupé VM, Spruit M, Ostelo RWJG, de Kleuver M. Can patient-reported profiles avoid unnecessary referral to a spine surgeon? An observational study to further develop the Nijmegen decision tool for chronic low back pain. *PLoS One* 2018;**13**(9):1–18. doi:10.1371/journal.pone.0203518.
27. Coupé V, van Hooff M, de Kleuver M, Steyerberg EW, Ostelo RWJG. Decision support tools in low back pain. *Best Practice & Research. Clinical Rheumatology* 2016;**30**(6):1084–97. doi:10.1016/j.berh.2017.07.002.
28. van Baalen S, Boon M. An epistemological shift: From evidence-based medicine to epistemological responsibility. *Journal of Evaluation in Clinical Practice* 2015;**21**(3):433–9. doi:10.1111/jep.12282.
29. Schemmer M, Kühl N, Satzger G. Intelligent decision assistance versus automated decision-making: Enhancing knowledge workers through explainable artificial intelligence. In: *55th Hawaii International Conference on System Sciences*. University of Hawaii, Manoa. 2022. doi:10.24251/hicss.2022.185.
30. Barocas S, Selbst AD, Raghavan M. The hidden assumptions behind counterfactual explanations and principal reasons. In: *FAT* 2020 - Proc 2020 Conf Fairness, Accountability, Transparency*, ACM, Barcelona, Spain; 2020:80–9. doi:10.1145/3351095.3372830.
31. Miller T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 2019;**267**:1–38. doi:10.1016/j.artint.2018.07.007.
32. Bjerring JC, Busch J. *Artificial intelligence and patient-centered decision-making*. Philosophy & Technology 2021;**34**(2):349–71. doi:10.1007/s13347-019-00391-6.
33. Jaynes ET. *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press; 2003.
34. Claeskens G, Hjort NL. *Model Selection and Model Averaging*. Cambridge: Cambridge University Press; 2008.
35. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: A systematic review. *BMC Medical Informatics and Decision Making* 2016;**16**(1):1–14. doi:10.1186/s12911-016-0377-1.
36. Croskerry P. Adaptive expertise in medical decision making. *Medical Teacher* 2018;**40**(8):803–8. doi:10.1080/0142159X.2018.1484898.
37. Busemeyer JR, Gluth S, Rieskamp J, Turner BM. Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. *Trends in Cognitive Sciences* 2019;**23**(3):251–63. doi:10.1016/j.tics.2018.12.003.
38. Noguchi T, Stewart N. Multialternative decision by sampling: A model of decision making constrained by process data. *Psychological Review* 2018;**125**(4):512–44. doi:10.1037/rev0000102.
39. See note 37, Busemeyer et al. 2019, at 251–63.
40. Rilling JK, Sanfey AG. The neuroscience of social decision-making. *Annual Review of Psychology* 2011;**62**:23–48. doi:10.1146/annurev.psych.121208.131647.
41. van Baar JM, Chang LJ, Sanfey AG. The computational and neural substrates of moral strategies in social decision-making. *Nature Communications* 2019;**10**(1):1483. doi:10.1038/s41467-019-09161-6.
42. Park SA, Sestito M, Boorman ED, Dreher JC. Neural computations underlying strategic social decision-making in groups. *Nature Communications* 2019;**10**(1):1–12. doi:10.1038/s41467-019-12937-5.
43. Füllbrunn S, Luhan W, Sanfey A. Current issues in decision making for others. *Journal of Economic Psychology* 2020;**77**:102250. doi:10.1016/j.joep.2020.102250.
44. Tunney RJ, Ziegler FV. Toward a psychology of surrogate decision making. *Perspectives on Psychological Science* 2015;**10**(6):880–5. doi:10.1177/1745691615598508.
45. See note 7, EU High Level Expert Group on AI 2019.
46. See note 7, EU High Level Expert Group on AI 2019, at 19.
47. See note 7, EU High Level Expert Group on AI 2019, at 19.

48. Alhashime Z. The effects of falsification machines on the quality of clinical decision-making. Bachelor Thesis, Radboud University, Nijmegen. 2021. Available at https://theses.ubn.ru.nl/handle/123456789/12753 (last accessed 20 December 2022).
49. van Eerdt R. Falsification machines in medical decision making. Radboud University, Nijmegen. 2021.
50. Goddard K, Roudsari A, Wyatt J. Automation bias - A hidden issue for clinical decision support system use. *Studies in Health Technology and Informatics* 2011;**164**:17–22. doi:10.3233/978-1-60750-709-3-17.
51. Grissinger M. Understanding human over-reliance on technology. *Pharmacology & Therapeutics* 2019;**44**(6):320–1.
52. . See note 13, Santoni de Sio, Mecacci 2021.
53. See note 14, Matthias 2004.
54. Friedman CP, Elstein AS, Wolf FM, Murphy GC, Franz TM, Heckerling PS, et al. Enhancement of clinicians's diagnostic reasoning by computer-based consultation: A multisite study of 2 systems. JAMA. 1999;**282**(19):1851–6. doi:10.1001/jama.282.19.1851.
55. Khairat S, Marc D, Crosby W, Al Sanousi A. Reasons for physicians not adopting clinical decision support systems: Critical analysis. *Journal of Medical Internet Research Medical Informatics* 2018;**6**(2):e24. doi:10.2196/medinform.8912.
56. Van de Velde S, Heselmans A, Delvaux N, Brandt L, Marco-Ruiz L, Spitaels D, et al. A systematic review of trials evaluating success factors of interventions with computerised clinical decision support. *Implementation Science* 2018;**13**(1):1–11. doi:10.1186/s13012-018-0790-1.
57. See note 31, Miller 2019.
58. Riveiro M, Thill S. "That's (not) the output I expected!" On the role of end user expectations in creating explanations of AI systems. *Artificial Intelligence* 2021;**298**:103507. doi:10.1016/j.artint.2021.103507.
59. Noordegraaf M. Protective or connective professionalism? How connected professionals can (still) act as autonomous and authoritative experts. *Journal of Professions and Organization* 2020;**7**(2):205–23. doi:10.1093/jpo/joaa011.
60. Shiohira K. *Understanding the Impact of Artificial Intelligence on Skills Development*. Education 2030; 2021; available at https://unesdoc.unesco.org/ark:/48223/pf0000376162 (last accessed 20 December 2022).
61. Baumberg Geiger B, Garthwaite K, Warren J, Bambra C. Assessing work disability for social security benefits: International models for the direct assessment of work capacity. *Disability and Rehabilitation* 2018;**40**(24):2962–70. doi:10.1080/09638288.2017.1366556.
62. UWV. *UWV Jaarverslag*; 2020; available at https://jaarverslag.uwv.nl/.

Cite this article: Haselager P, Schraffenberger H, Thill S, Fischer S, Lanillos P, van de Groes S and van Hooff M (2024). Reflection Machines: Supporting Effective Human Oversight Over Medical Decision Support Systems. *Cambridge Quarterly of Healthcare Ethics* 33: 380–389, doi:10.1017/S0963180122000718