OPEN ACCESS

Primer on an ethics of AI-based decision support systems in the clinic

Matthias Braun ,[1] Patrik Hummel,[1] Susanne Beck,[2] Peter Dabrock[1]

[1]Insitute for Systematic Theology, Friedrich-Alexander University Erlangen-Nürnberg (FAU), Erlangen, Germany
[2]Institute for Criminal Law and Criminology, Leibniz University Hannover, Hannover, Germany

Correspondence to
Dr Matthias Braun, FAU, Erlangen 91054, Germany; matthias.braun@fau.de

Received 20 September 2019
Revised 23 December 2019
Accepted 4 February 2020
Published Online First 3 April 2020

ABSTRACT

Making good decisions in extremely complex and difficult processes and situations has always been both a key task as well as a challenge in the clinic and has led to a large amount of clinical, legal and ethical routines, protocols and reflections in order to guarantee fair, participatory and up-to-date pathways for clinical decision-making. Nevertheless, the complexity of processes and physical phenomena, time as well as economic constraints and not least further endeavours as well as achievements in medicine and healthcare continuously raise the need to evaluate and to improve clinical decision-making. This article scrutinises if and how clinical decision-making processes are challenged by the rise of so-called artificial intelligence-driven decision support systems (AI-DSS). In a first step, this article analyses how the rise of AI-DSS will affect and transform the modes of interaction between different agents in the clinic. In a second step, we point out how these changing modes of interaction also imply shifts in the conditions of trustworthiness, epistemic challenges regarding transparency, the underlying normative concepts of agency and its embedding into concrete contexts of deployment and, finally, the consequences for (possible) ascriptions of responsibility. Third, we draw first conclusions for further steps regarding a 'meaningful human control' of clinical AI-DSS.

INTRODUCTION

The tremendous potentials of computerised decision support tools within the medical sector have been propelled to new heights. They benefit from significant increases in computing power, the amounts of available data and progress in artificial intelligence (AI). AI can be understood as an umbrella term for technologies intended to mimic, approximate or even extend features and abilities of animals and human persons.[1] Particular success is being achieved in image-based diagnosis. Some convolutional neural networks have been shown to perform on par with[2] or even better than[3] dermatologists in classifying images of skin lesions and distinguishing benign and malignant moles. Paradigmatic for the involvement of big technology companies in these endeavours, Google has developed a deep learning algorithm that detects diabetic retinopathy and diabetic macular oedema[4] in retinal images with similar accuracy as ophthalmologists. Microsoft is involved in initiatives applying automated analysis of radiological images in order to ameliorate the time-consuming and error-prone delineation of tumours.[5] IBM's Watson for Oncology applies AI to the personalisation of cancer care.[6] These dynamics shape and gradually change healthcare and biomedical research.[7–10]

Besides these ongoing endeavours, decision-making especially in the clinic remains a complicated and critical task, as healthcare providers have to provide diagnoses and possible treatments according to the specific medical condition of the patient and within time constraints.[11] The primary goal of clinical decision support systems (DSS) is to provide tools to help the clinicians as well as the patients to make better decisions. AI-driven decision support systems (AI-DSS) take various patient data and information about clinical presentation as input, and provide diagnoses,[12] predictions[13] or treatment recommendations[14] as output. Overall, awareness of these correlation-based patterns may inform decision-making, contribute to more cost-effectiveness[15] and fundamentally ameliorate clinical care. While some of these prospects seem visionary or even vague, there are a rising number of concrete research endeavours[16] seeking to harness the increasing sophistication of AI-DSS.

Against this background, we put forward two hypotheses on the distinctive ethical challenges posed by clinical AI-DSS. First, they affect and transform the *modes of interaction* between different agents in the clinic. Second, these modes are entangled with *four normative notions* which are shifted and whose presuppositions are undercut by AI-DSS: (A) conditions of *trustworthiness*, (B) epistemic challenges regarding *transparency*, (C) alterations in the underlying concepts of *agency* and, finally, (D) the consequences for (possible) ascriptions of *responsibility*. While there is a lot of work on each of these individual normative notions, there is a lack of understanding of their entanglement and especially their significance for governance strategies towards shaping and modelling the current and future use of AI-DSS in the clinic. In order to tackle these challenges, we sketch the contours of 'meaningful human control' of clinical AI-DSS.

AI AND THE SHIFTING MODES OF CLINICAL INTERACTIONS

AI-based applications are capable of considering large amounts of data in which they discover and highlight correlations that might have otherwise escaped the attention of clinicians and researchers. One important preliminary observation is that AI-DSS stretch or sometimes even collapse the borders between the clinic and biomedical research.[17] One reason for this is that such AI-DSS merge a variety of different sets of data, which often have been gathered within the framing of research and are then transferred by these tools to clinical care settings. A second reason is implied in the way AI-DSS produce hypotheses. Fundamentally

To cite: Braun M, Hummel P, Beck S, *et al*. *J Med Ethics* 2021;**47**:e3.

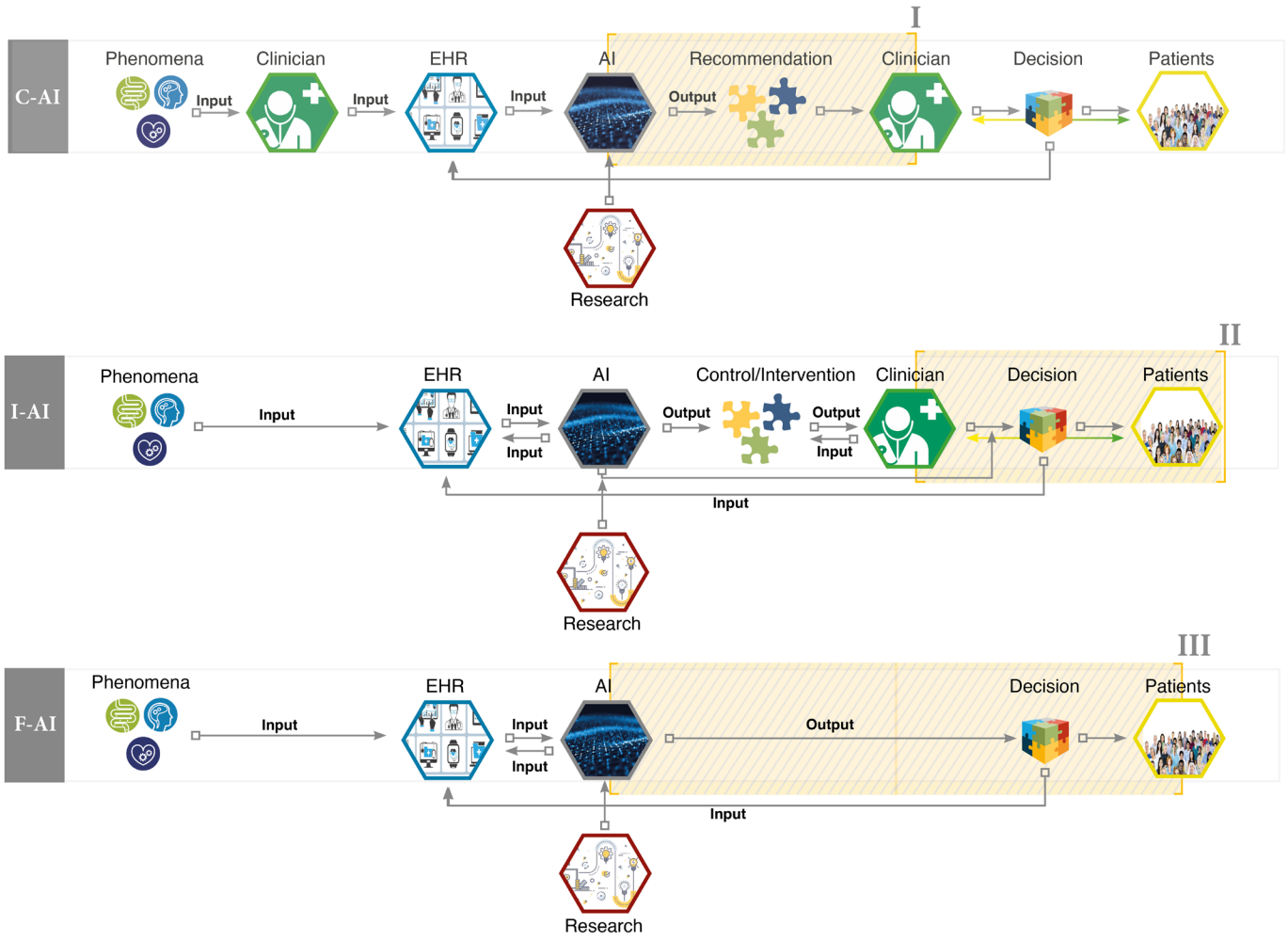

Figure 1 Interaction modes within AI-driven decision support systems. AI, artificial intelligence; C-AI, conventional AI-driven decision support system; EHR, electronic health record; F-AI, fully automated AI-driven decision support system; I-AI, integrative AI-driven decision support system.

new is that unlike hypothesis-driven research, data or model-driven decision-making proceeds on the basis of AI processing large amounts of data and providing classifications without a thorough understanding of the underlying mechanisms of the model.[8] This offers exciting opportunities, for example, to evaluate large amounts of cross-sectorial data, to discover unforeseen correlations and to feed them into clinical practice via DSS. The clinician is eventually deploying the tool, but her decisions and actions are intertwined with the research and development efforts of the individuals designing, calibrating and refining the support system.

This entanglement of research and clinic could be described as a first fundamental precondition for the use of AI-DSS in the clinical context because the whole process of collecting, gathering and interpreting different sets of data as well as developing the algorithms of AI systems is the result of strong interlinkage between the clinic and research. The second fundamental precondition is that there are several different ways in which AI-DSS could function in the clinic (see figure 1). Drawing on the work of Yu *et al*,[7] we can distinguish so-called conventional AI-DSS (C-AI) systems from integrative AI-DSS (I-AI) and fully automated AI-DSS (F-AI). In *conventional* DSS, an algorithm takes patient data as input and informs decision-making by delivering a statement for consideration to the clinician. In *integrative* DSS, the algorithm can request and gather patient data autonomously, present the result of data processing to the clinician and write it into the patient's electronic health records. Across these variants, the roles attributed to AI shift the modes of interaction among the clinical agents in different ways.

The *first* interaction mode affected by AI-DSS is between the clinicians and their machine(s). In ordinary clinical contexts, the clinician guides the patient through the spectrum of available care. The introduction of AI-driven clinical DSS can supplement the professional's experience and knowledge, and even alter her decisional authority by shaping expectations, verdicts, roles, and responsibilities. Across the AI-DSS, the clinician shares agency to different degrees with the AI tool. On the far end of the spectrum, the AI system is not merely a tool to *augment* the clinician's decision-making, but to some extent *replaces* human reasoning[18 19] by evaluating clinical presentations, arriving at decisions and updating health records autonomously.

The *second* changing mode of interaction is the one between the clinicians and the patients. In ordinary clinical contexts, it is already an oversimplification to regard the clinician as the sole decision-maker. According to the ideal and principles of shared decision-making in clinical practice, healthcare professionals and patients share the best available evidence, and patients are provided with support to consider options and to arrive at informed preferences.[20 21] AI-DSS provide a direct link between patient data (vital parameters and her digital health

diary) and the clinician's diagnostic toolbox, and thereby add to the evidence base at the centre of shared decision-making. Availability of this additional evidence requires clinician and patient to jointly evaluate its significance and to relate it to both the clinician's judgement based on her knowledge and experience as well as the patient's preferences, expectations and concerns. While additional pieces of evidence can be a welcome enrichment of decision-making processes, they can also stand in tension with previous assessments and intentions. The situation can be complicated by the possibility that the quality of such additional evidence is not immediately transparent to all involved and affected. When the support system's recommendations and predictions raise tensions and suspicions, shared decision-making requires clinician and patient to reassess and to render beliefs, preferences and intentions coherent.

Third, new forms of interactions between patients and the machines occur. By agreeing to consult the AI-DSS, the patient is typically feeding her data into the algorithmic tools on which the system is based, and thereby contributes to their training and refinement. While the patient might be motivated by the prospect of improved care and additional control mechanisms, attitudes of solidarity, giving and contributing to the common good can be in play, too. Once data are being shared, the machine shapes the care of the patient. One example is the use of AI-triggered self-medication in order to manage risks of non-adherence.[22] Here too, the concrete dimension of a possible shift in the interaction between patients and machine depends on the way the AI-DSS is constructed and equally important: how the interaction interface between the patients and the respective machine is constructed. While in a C-AI system the interaction between the clinician and the patient may still be led by the clinician, there is no clinician directly involved in collecting and gathering the data in the I-AI system. Furthermore, in the case of a possible F-AI system, no clinical expert is directly present even in the decision-making process. Hitherto there is a lot of discussion, if such a shift in the process of clinical decision-making would lead to new modes of concrete participation, or whether it will increase or even create new spaces of vulnerability. This would, for example, be the case if such vulnerable groups would have no concrete idea of the way the inferences or even recommendations of the AI-DSS have been processed, if there are biases in its training data and which influence they may have on the results.

THE NORMATIVE CHALLENGES

The described shifts in interaction modes transform established processes of arriving at clinical decisions, and have implications for a range of concepts and categories for *evaluating* deliberation processes and decisions from a normative perspective.

First, questions arise about the *conditions of trustworthiness* of such systems, and what it takes to advance '[t]owards trustable machine learning'[17] and the 'implementation and realization of Trustworthy AI'[23] in the clinic. Empirical work highlights that different stakeholders introduce distinctive expectations into the set-up. In order to gain the trust of clinicians, AI should be user-friendly and based on adequate risk-benefit analyses.[24] Patients expect that AI enhances the care they receive, preferably without removing the clinician from the decision-making loop in order to maintain human interaction and interpersonal communication[25] with a clinician who is in a position to evaluate the outputs of the system and to compare them with judgements arising from her own professional experience and training. As with other data-intensive applications, adherence to data protection and privacy requirements[26 27] such as the general data protection regulation (GDPR) will be essential. Moreover, it will be important that legally adequate levels of risk are being clarified beforehand in order to enable legal security for relevant actors and to give potential victims of damages the possibility to address transgressions of these risk levels. This is one distinctive challenge in regulating AI adequately.[28] Loans of trust from society to researchers, engineers and users of novel biomedical technologies[29] will be withdrawn if expectations like these are not met.

Obstacles include the fact that AI-driven clinical DSS, while becoming increasingly powerful, offer no guarantee for validity and effectiveness. The whole process of setting up a data-based system—from the collection of data, training of the model, up to its actual use in a social context—involves many actors. Every step is prone to certain errors and misconceptions,[30] and can result in damages to involved or even uninvolved parties. On the one hand, this is the reason why it is not sufficient for trustworthiness to define overarching ethical or legal principles. Such principles, for example, autonomy, justice or non-maleficence, can provide orientation. But they need to be embedded into a context-sensitive framework.[31] On the other hand, patients' openness to AI-driven health tools varies considerably across applications and countries.[32] There remains a need for further empirical and conceptual work on which conditions of trustworthiness—if any[33]—matter relative to the full range of AI-driven applications and their implementation in the clinic.

Second, one specific epistemic challenge of AI-driven clinical DSS is *transparency*. While clinical decision-making is and has always been challenged by scarce evidence, time or the complexity of diagnosis, the use of AI-DSS promises to enhance the decision-making process in the clinic. Using AI-DSS may process larger sets of data in a much shorter time and may be less susceptible to biases in individual experiences. But despite these possible benefits, there remains a fundamental challenge which is discussed under the term of (epistemic) opacity.[34] While the logic of simple algorithms can be fully comprehensible, the kinds of algorithms that tend to be relevant and useful towards practical applications are more complex. With increasing complexity, for example, when artificial neural networks are used, it might still be possible to state the general working principles of the different algorithms that are implemented in a system. However, the grounds on which the algorithm provides a particular classification or recommendation in a specific instance are bound to be opaque to designers and users,[35] especially since it also depends on training data and user interactions. The opacity can refer to the question of *how*, that is, by means of which statistical calculations, rules or parameters, an algorithmic system arrives at the output. Even if in principle available, grasping machine-generated rules and correlations can be challenging as they take into account a large number of diverse parameters and dimensions. In a different sense, opacity can refer to the question of *why* a given output was provided.[36] In this explanatory sense, opacity concerns the underlying causal relationship between input and output, which algorithms—by virtue of solely providing correlations—do not necessarily elicit. With regard to evaluating the outputs of AI-DSS, there are thus informational asymmetries between patients and clinicians, and between clinicians and AI-DSS. Such black box issues make it almost impossible to assess risks adequately *ex ante* as well as to prove e*x post* which wrongful action has led to specific damage, thus hindering legal evaluation.[37] This threatens to result in a form of 'black-box medicine'[38] in which the basis for a given output is not always sufficiently clear and thus complicates its evaluation in view of potential errors and biases of the system, arising, for example, from the quality and breadth of data it has been trained with.

On the one hand this is problematic from a legal point of view, as this could potentially lead to inadequate discrimination,[39] for example, when AI-DSS lead to varying levels of service quality for different populations. On the other hand, there are also ethical difficulties regarding the way we should deal with the inherent opacity. Some scholars propose to supplement traditional bioethical principles by a principle of *explicability*.[23 40] One motivation is to enable assessments of potential biases in AI-driven decisions, which in clinical settings could affect the quality of care for certain populations as well as reinforce and aggravate pre-existing health inequities.[41 42] At the same time, calling for explicability can only be the beginning. First, it raises the question what it takes to arrive at explainable AI, that is, *how* a sufficient degree of transparency of AI-driven clinical decision-making could be achieved, what counts as a sufficient explanation from the patient perspective and how residual opacity should be handled. Second, an important challenge is to define the level, at which point of the process something should be explainable. For example, we could strive for *input*-related transparency, that is, explainability relative to the point in time at which one deliberates about whether and which data to feed into an AI-DSS, and trying to illuminate how such data *will* be processed. Not incompatible, but different in orientation would be the attempt to make *outputs* transparent, that is, to explain why and how an AI-DSS arrived at a particular result.Third, and often underestimated: even if the degree of explainability at the respective level has been clearly defined, it still remains an open question how to communicate AI-driven outputs to patients, how to enhance patient literacy with regard to such information and how AI-driven outputs could be introduced into processes of shared clinical decision-making. One reason why this matters is that only then can consent be informed and hence meaningful from a normative perspective. Another is that sometimes, precisely *because* AI-DSS aims to improve the evidence base for a particular clinical decision, residual indeterminacy and risks become apparent which leave it somewhat *unclear* which option is in the patient's best interest. In such cases, it will be difficult to justify privileging one of the options without involving the patient in the assessment.

Third, AI-DSS give rise to structures in which *agency* is *shared* (figure 1). Already now, there is a variety of individuals in the clinic whose reflection and decision-making are mutually intertwined and interdependent. The presence and deployment of AI-DSS gives rise to new forms of this phenomenon. Already prior to the limiting case of fully automated AI decisions, the system affects, shapes and can stand in tension with the clinician's judgement. This raises the question of who is guiding clinical decision-making, in which ways and on what grounds. In order to lay the groundwork for dealing with new forms of shared agency, we do need a refined account of agency. Such an account first has to illuminate the range of agents involved in applications of AI-DSS. Second, it has to be defined in which sense the machine affects and accelerates decisions, and maybe has the ability to decide on its own. Third, we have to rethink how informed individuals can be presumed to be about the processes and working principles in question. In legal contexts, one long-standing proposal is to ascribe agency onto these systems,[43 44] or to require a 'human in the loop' for specific decisions.[45] *Shared agency* has not yet been transferred into practical jurisprudence or legislation, although often it is proposed that 'meaningful human control' over the events is required.[46]

Fourth, shared agency raises a problem of many hands[47] for ascriptions of *responsibility*: since a plurality of agents contributes to decision-making guided by AI-DSS, it becomes less clear who is morally and legally answerable in which ways. With the involvement of autonomous, adaptive and learning systems, it becomes harder to ascribe individual responsibility and liability for singular decisions, especially those with adverse outcomes. The difficulties with proving who made a mistake, and the *telos* of—at least partly—transferring decision-making to the machine make it less justifiable to regard one of the parties involved as fully accountable for the decision.[48 49] Responsibility redistributes and diffuses across agential structures, and it becomes questionable what counts as sufficient proof of misconduct of one of the parties involved. In the clinical setting, this raises a need for frameworks on medical malpractice liability resulting from deploying AI-DSS.

Some argue that unless AI genuinely *replaces* clinicians, it merely *augments* decision-making, and clinicians retain final responsibility,[16] thus becoming the (legally) responsible 'human in the loop'. Difficulties with this reasoning become apparent once we realise that the system considers large and complex data sets and leverages computational power towards identifying correlations that are not immediately accessible to human inquiry. With increases in complexity, it becomes less plausible to expect that the clinician is in a position to query and, second, guess the system's output and its attunement to the intended task.[50] We then reach a pitchfork where *either* responsible clinicians refrain from using potentially beneficial, powerful but complex and somewhat opaque tools *or* we rethink attributions of responsibility and liability.

The diffusion of responsibility and liability can have problematic consequences: the victim might be left alone, the damages might remain unresolved and society might feel concerned about a technological development for which accountability for damages and violations of rights remains unclear. Fragile arrangements of trust can break, pre-existing reservations and unease about AI[25 51] be amplified, and calls for overly restrictive governance result if public attitudes, narratives and perceptions are not taken seriously and channelled into inclusive societal deliberations.[52]

The described transformations driven by AI-DSS bear the potential to (gradually) transform clinical interaction modes and in doing so, they transform normative concepts and standards of trustworthiness, transparency, agency and responsibility (see table 1). At the same time, it would be a much too simplistic approach to only analyse the impact of AI-DSS on each of the normative notions separately. Additional complexities arise from the fact that these normative categories are not only shifted individually, but due to mutual entanglements affect and change each other's meanings and connotations across agential configurations. For example, lack of transparency in system architecture and output explainability might leave the treating clinician somewhat in the dark about underlying causal relations on which the system picks up (clinician—AI-DSS), but she will still be much better placed to assess system outputs than the patient for whom such black box issues are aggravated (patient—AI-DSS) due to a lack in clinical and technical background. These transparency challenges for assessing the significance, quality and implications of outputs in turn change the ability to exercise well-informed agency in the context of shared clinical decision-making (clinician—patient), which then raises the need for new forms of counselling and communication pathways to maintain trustworthiness in the clinician–patient relationship. Mutual dependencies like these, coupled with the increasing sophistication of AI-DSS, do not make it easier to arrive at governance strategies that are sensitive to the rights, interests and expectations of those affected and allow us to move forward with harnessing

Table 1 Changing modes of interaction and their entanglement with different normative notions

Normative notions	Interaction modes		
	Clinician–AI-DSS	**Clinician–patient**	**Patient–AI-DSS**
Agency	*Description*: introduction of AI-DSS competence and authority alongside clinician competence. *Challenge*: spectrum ranges from augmentation to replacement. *Consequence:* need to prioritise and to mediate between clinician and AI-DSS judgements in case they diverge.	*Description:* introduction of additional evidence at the centre of shared decision-making. *Challenge:* finding context-sensitive modes of participation in collaborative evaluation of outputs. *Consequence:* need for enhanced and standardised sensitivity to clinician and patient attitudes and expectations about roles in decision-making informed by AI-DSS.	*Description*: interplay between AI-DSS ideally ameliorating individual patient care and patient data learning and refining AI-DSS. *Challenge*: sharing of large amounts of health data as a necessary condition for using medical AI. *Consequence:* need for mechanisms to exercise informational self-determination through awareness of available options, data flow controllability and privacy by design.
Trustworthiness	*Description:* room for errors across data collection, model training and implementation; significance of AI-DSS reliability, validity and attunement to task. *Challenge:* user-friendliness, safety and efficacy, potential burden of proof for clinician who deviates from AI-DSS recommendations. *Consequence:* need for clinician training to evaluate reliability, validity and attunement.	*Description:* perceived transformations to clinical decision-making processes. *Challenge:* shifts and extensions of conceptions of clinician competence and communication, expectation that human clinician remains in the loop. *Consequence:* need for new forms of counselling and communication pathways.	*Description*: emerging conditions of trustworthiness for deployment of AI-driven technology from the patient perspective. *Challenge*: varying degrees of background knowledge and openness for AI-driven clinical tools, possibility that contact to human clinicians becomes a luxury. *Consequence:* need for provision of clear information on opportunities, challenges, data protection, procedures and addressees for damage claims, and humans in the loop.
Transparency	*Description:* AI uncovering correlations without necessarily eliciting underlying causal relations. *Challenge:* lack of evidence on how and why AI-DSS arrives at a given output. *Consequence:* need for safety and efficacy validation and certification markers.	*Description:* black box issues coupled with information asymmetries between clinician and patient. *Challenge:* asymmetries in ability to assess and reflect on AI-DSS outputs, challenges in attributability of recommendations. *Consequence:* providing information and education about AI-driven clinical decision-making, clarifying the role and competence of the clinician relative to the AI-DSS.	*Description:* black box issues aggravated by lack of clinical and technical background. *Challenge:* making the meaning, quality and limitations of AI-DSS outputs transparent to patients. *Consequence:* need for provision of clear information on the nature of AI-DSS outputs, for example, through visualisations and innovative patient interfaces.
Responsibility	*Description:* partial shift of responsibility for diagnosis, recommendations and decision-making from clinician towards AI-DSS. *Challenge:* problem of many hands in AI development and implementation, complicating attributions of responsibility for malfunctions. *Consequence:* need for clear and context-sensitive principles for how ethical and legal responsibility distributes across multiagential structures.	*Description:* changing nature of clinician responsibility in view of informational asymmetries between clinician and patient. *Challenge:* collaboratively assessing risk-benefit ratios. *Consequences:* need for training and education for clinicians and patients towards responsible utilisation of AI.	*Description*: consideration of and reliance on AI-DSS outputs when patients make decisions on their own authority,[5] paternalism versus new forms of individual choices. *Challenge*: mediating between self-interest, solidaric participation and backlashes at the level of justice. *Consequence*: need for education and counselling.

AI-DSS, artificial intelligence-driven decision support system.

new technologies responsibly. Suitable governance strategies need to be mindful of how AI-DSS transform clinical interaction modes, and how relevant normative categories are mutually entangled and affected.

DEALING WITH CLINICAL AI-DSS: TOWARDS MEANINGFUL HUMAN CONTROL

The idea of meaningful human control is widely discussed as a possible framework to face challenges like the foregoing in dealing with AI-driven applications. Acknowledging that the concept of meaningful human control is still under discussion and currently remains a more or less fuzzy concept,[53–55] the underlying idea is clear enough: AI is nothing which just happens, but which should and can be controlled by humans. Even though a kind of shifting agency, a lack of transparency or even an erosion of control caused by AI-DSS is possible, the idea of meaningful human control articulates clear requirements for AI development and interactions: it is human agents who retain decisional authority. AI-DSS are auxiliary tools to enhance human decision-making. But they do not by themselves determine courses of action. Important clinical choices, for example, on treatments, resource allocation or the weighing of risks, require human supervision, reflexion and approval. In order to *meaningfully* control these choices, presumably sensitivity to and alignment with human concerns, needs and vulnerabilities is necessary throughout the process of system design, implementation and deployment. Besides this very fundamental basic line, there still remains the question how such an idea of meaningful human control can be developed and rolled out. We put forward three aspects of an account of meaningful control, and sketch some of its practical implications.

First, it is necessary to analyse the *legal dimensions* of the challenges and problems in depth, and to look for potential solutions in close cooperation with other disciplines (for a historical perspective on the debate see Bench-Capon and colleagues[56]). Some examples for discussed legal regimes are regulation via strict liability, the creation of the e-Person,[57] the introduction of obligatory insurances for the usage of AI and mandating a human in the loop who then also would be accountable for the decision. In order to promote meaningful human control, these ideas will need to be complemented with mechanisms of validation and certification for algorithms and developers as 'hallmarks of careful development'[58] which clinicians and facilities should take into account before deploying AI-driven tools. For example, regulatory approval of AI-DSS could be tied to evidence that the system reliably improves patient outcomes, is based on proper risk assessments and is ethically trained to mitigate bias.[59] Some even demand that AI systems can be genuine bearers of responsibility, and call for a distinctive legal status resembling the legal personhood of collective entities. This suggestion would involve transformations of present societal understandings of autonomy, personhood and responsibility,[60] and could lead to reconceptions of fundamental legal concepts such as action, attribution, liability and responsibility. This is not the place to conclusively evaluate such wide-ranging proposals. We merely highlight that meaningful human control will require new legal regimes, which need to be assessed based on whether or not they promote such control.

Second and with regard to the governance of data that is necessary for developing and refining AI-driven system, the ideal of individual *data sovereignty* is gaining traction. The concept relates to issues of control about who can access and process data.[61 62] It is driven by the conviction that claims to informational self-determination can only be realised against the backdrop of social contexts and structures in which they are articulated, recognised and respected.[63 64] In this respect, digitisation has the potential to transform the social core in which articulations of these claims are always embedded. This is why it is inadequate to insist on rigid, input-oriented data protection principles like data minimisation and purpose limitation.[65] As one example, Wachter and Mittelstadt maintain that a full-fledged data protection law also needs to encompass rights that concern the *inferences* that are being drawn on the basis of data-driven analytics.[66] These rights shall cover high-risk inferences and require disclosure of information that allows to determine why the considered data are acceptable bases for the inferences, why the inferences are acceptable for a given purpose and whether these inferences are accurate. While succeeding in going beyond mere input orientation, the concrete content of a right to reasonable inferences substantially depends on the criterion of 'reasonableness'. Thus, a right to reasonable inferences in a sense shifts the problem to another level: the elaboration and societal negotiation of what counts as reasonable in a given context, and why. That is, in order to develop this right into a comprehensive approach to meaningful control of AI-DSS, the focus must shift to the social transformations[67] that are being brought about by digitisation. In these settings, individuals should be put in a position to exercise informational self-determination reliably and robustly by being put in a position to control the flow of their data. For example, rather than regarding patients as mere data subjects whose personal health data can be analysed under the GDPR on the basis of *broad* and potentially even *no* consent mechanisms, the ideal of meaningful control calls for concrete modes for individual control. Such modes of control could, for example, be implemented by envisioning patients as *comanagers* of their data and of the processes into which such information is channelled.

Indeed, AI-driven tools need training data to provide useful outputs, and so one essential condition for their success in the clinic is the willingness of individuals to share health data[16 68] and thereby to contribute towards applications that will benefit them and enhance the common good. In order for these acts of sharing to be the result of data sovereignty,[69] individual decision-making must be informed about the working principles of the system, the consequences of data processing and the availability of alternative methods of diagnosis and care.

Third, and in view of the agential configurations surrounding AI-DSS, similar questions about sovereignty arise with regard to the *role and decisional authority of the clinician*. Observers are torn between highlighting putative skills of clinicians that machines cannot emulate[70] and cautioning against romanticising human judgement.[13 18] Soundbites like '[c]ould artificial intelligence make doctors obsolete?'[71] or '[t]he practice of medicine will never disappear, but our role in it as clinicians hinges on what we do next [after AI]'[72] illustrate that public perceptions and self-understandings of clinicians are being transformed. On the one hand, opacity and uncertainty about the validity and error-proneness of AI-driven systems frame interpretative processes, derivations of appropriate actions and already the design and debugging of the system itself. On the other hand, heightened anticipations and perceived potentials of these sophisticated systems raise the question under which conditions clinicians can actually *refrain* from deploying such systems or, once they are deployed, make decisions that *contrast* with what the system's outputs suggest. The burden of proof might shift towards the deviating clinician. Whether this is a problematic development or could be a part of a responsible dealing with AI-DSS will then

depend on two factors. First, whether there is sound evidence that in the particular context at hand, reliance on the AI-DSS addresses the patient's need better than alternative courses of actions. Second, the idea of meaningful human control would require that any remaining risks and uncertainties about the foregoing are deliberated on by humans, in particular by the clinician(s) together with the patient. Strictly speaking, *deviance* becomes a misnomer: the description presupposes that there is a determinate right course of action. Even with the most sophisticated AI-DSS, complexity and uncertainty will most likely remain part of medical practice. AI-DSS might help navigate them, but will not resolve them. It remains a critical task of the medical profession more than ever to provide the competence and resources for assessing, avoiding and taking risks responsibly, and to involve and to counsel the patient throughout this process.

Twitter Matthias Braun @brau_matt

Acknowledgements The authors are gratefully thankful for the comments and critiques by the reviewers as well especially (in alphabetical order) Hannah Bleher, Eva Hille, Stephanie Siewert and Max Tretter.

Contributors MB and PH are the main authors of the article. SB and PD commented on all parts of the article.

Funding This work is part of the research project DABIGO (ZMV/1–2517 FSB 013), which has been funded by the German Ministry for Health, as well as the research project vALID (01GP1903A), which has been funded by the German Ministry of Education and Research.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iD

Matthias Braun http://orcid.org/0000-0002-6687-6027

REFERENCES

1 Bringsjord SG, Sundar N, Zalta EN. *Artificial intelligence, in the Stanford encyclopedia of philosophy*, 2018.
2 Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542(7639):115–8.
3 Haenssle HA, Fink C, Schneiderbauer R, *et al*. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol* 2018;29(8):1836–42.
4 Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus Photographs. *JAMA* 2016;316(22):2402–10.
5 Microsoft. Project InnerEye - Medical Imaging AI to Empower Clinicians. Microsoft Research, 2018. Available: https://www.microsoft.com/en-us/research/project/medical-image-analysis/ [Accessed 28 Feb 2018].
6 Somashekhar SP, Sepúlveda M-J, Puglielli S, *et al*. Watson for oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. *Ann Oncol* 2018;29(2):418–23.
7 Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng* 2018;2(10):719–31.
8 Zhu L, Zheng WJ. Informatics, data science, and artificial intelligence. *JAMA* 2018;320(11):1103–4.
9 Topol EJ. High-Performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25(1):44–56.
10 Esteva A, Robicquet A, Ramsundar B, *et al*. A guide to deep learning in healthcare. *Nat Med* 2019;25(1):24–9.
11 Shoshtarian Malak J*et al*. Neonatal intensive care decision support systems using artificial intelligence techniques: a systematic review. *Artificial Intelligence Review* 2018.
12 Castaneda C, Nalley K, Mannion C, *et al*. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. *J Clin Bioinforma* 2015;5(1):4.
13 Chen JH, Asch SM. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. *N Engl J Med* 2017;376(26):2507–9.
14 Komorowski M, Celi LA, Badawi O, *et al*. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med* 2018;24(11):1716–20.
15 Beam AL, Kohane IS. Translating artificial intelligence into clinical care. *JAMA* 2016;316(22):2368–9.
16 Fenech M, Strukelj N, Buston O. *Ethical, Social, and Political Challenges of Artificial Intelligence in Health*. Future Advocacy, Wellcome Trust, 2018.
17 Nature Biomedical Engineering. Towards trustable machine learning. *Nat Biomed Eng* 2018;2(10):709–10.
18 London AJ. Groundhog day for medical artificial intelligence. *Hastings Cent Rep* 2018;48(3).
19 Nabi J. How bioethics can shape artificial intelligence and machine learning. *Hastings Cent Rep* 2018;48(5):10–13.
20 Makoul G, Clayman ML. An integrative model of shared decision making in medical encounters. *Patient Educ Couns* 2006;60(3):301–12.
21 Elwyn G, Frosch D, Thomson R, *et al*. Shared decision making: a model for clinical practice. *J Gen Intern Med* 2012;27(10):1361–7.
22 Labovitz DL, Shafner L, Reyes Gil M, *et al*. Using artificial intelligence to reduce the risk of nonadherence in patients on anticoagulation therapy. *Stroke* 2017;48(5):1416–9.
23 High-Level Expert Group on Artificial Intelligence, Draft Ethics Guidelines for Trustworthy AI. *Working Document for stakeholders' consultation*. Brussels: The European Commission, 2018.
24 Harwich E, Laycock K. *Thinking on its own: AI in the NHS*, 2018.
25 Castell S*et al*. *Public views of machine learning. findings from public research engagement conducted on behalf of the Royal Society*. London: Ipsos MORI, The Royal Society, 2017.
26 Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. *PLoS Med* 2018;15(11):e1002689.
27 Wachter S. Data protection in the age of big data. *Nat Electron* 2019;2(1):6–7.
28 Scherer MU. Regulating artificial intelligence systems: risks, challenges, competencies, and strategies. *Harvard Journal of Law & Technology* 2016;29.
29 Braun M, Dabrock P. "I bet you won't": The science-society wager on gene editing techniques. *EMBO Rep* 2016;17(3):279–80.
30 Zweig KA, Fischer S, Lischka K. *Wo Maschinen irren können. Fehlerquellen und Verantwortlichkeiten in Prozessen algorithmischer Entscheidungsfindung*. Bertelsmann Stiftung: Gütersloh, 2018.
31 Mittelstadt B. Principles alone cannot guarantee ethical AI. *Nat Mach Intell* 2019;1(11):501–7.
32 Arnold D, Wilson T. *What doctor? why AI and robotics will define new health*, 2017.
33 Bryson J. *AI & Global Governance: No One Should Trust AI*, 2019.
34 McDougall RJ. Computer knows best? the need for value-flexibility in medical AI. *J Med Ethics* 2019;45(3):156–60.
35 Burrell J. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data Soc* 2016;3(1).
36 Ferretti A, Schneider M, Blasimme A. Machine learning in medicine: opening the new data protection black box. *European Data Protection Law Review* 2018;4(3):320–32.
37 Gleß S, Weigend T. Intelligente Agenten und das Strafrecht. *Zeitschrift für die gesamte Strafrechtswissenschaft* 2014;126(3):561–91.
38 Price N. Black-Box medicine. *Harvard Journal of Law & Technology* 2015;28(2):419–67.
39 Hacker P. Teaching Fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review* 2018;55(4):1143–85.
40 Floridi L, Cowls J, Beltrametti M, *et al*. AI4People-An ethical framework for a good AI Society: opportunities, risks, principles, and recommendations. *Minds Mach* 2018;28(4):689–707.
41 Adamson AS, Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatol* 2018;154(11):1247–8.
42 Yu K-H, Kohane IS. Framing the challenges of artificial intelligence in medicine. *BMJ Qual Saf* 2019;28(3).
43 Solum LB. Legal personhood for artificial intelligences. *North Carolina Law Review* 1992;70(4):1231–87.
44 Teubner G. *Digital personhood? the status of autonomous software agents in private law*. Ancilla Iuris, 2018: 36–78.
45 Sharkey N. Staying in the loop: human supervisory control of weapons. In: Bhuta NC, ed. *Autonomous weapons systems: law, ethics, policy*. Cambridge: Cambridge University Press:, 2016: 23–38.
46 Chengeta T. Defining the Emerging Notion of „Meaningful Human Control" in Weapon Systems. *NYU Journal of International Law, Forthcoming* 2017;49:833–90.
47 Nissenbaum H. Accountability in a computerized Society. *Sci Eng Ethics* 1996;2(1):25–42.
48 Beck S. Jenseits von mensch und Maschine: ethische und rechtliche Fragen zum Umgang MIT Robotern, künstlicher Intelligenz und Cyborgs 2012.
49 Beck S. Intelligent agents and criminal law—Negligence, diffusion of liability and electronic personhood. *Rob Auton Syst* 2016;86:138–43.
50 Goodman KW. *Ethics, medicine, and information technology*. Cambridge University Press, 2016.
51 YouGov Künstliche Intelligenz: Deutsche sehen eher die Risiken als die Nutzen. 2018.

52 The Royal Society. *Portrayals and perceptions of AI and why they matter*. London, 2018.
53 Ernst C. *Beyond Meaningful Human Control? – Interfaces und die Imagination menschlicher Kontrolle in der zeitgenössischen Diskussion um autonome Waffensysteme (AWS), in Die Maschine: Freund oder Feind? Mensch und Technologie im digitalen Zeitalter, C.B. Thimm, Thomas Christian, Editor*. Wiesbaden: Springer Fachmedien Wiesbaden, 2019: 261–99.
54 Santoni de Sio F, van den Hoven J. *Meaningful human control over autonomous systems: a philosophical account*. Frontiers in Robotics and AI, 2018.
55 Crootof R. A meaningful floor for meaningful human control autonomous legal Reasoning: legal and ethical issues in the technologies in conflict. *Temple International Comparative Law Journal* 2016;30(1):53–62.
56 Bench-Capon T, Araszkiewicz M, Ashley K, *et al*. A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law. *Artificial Intelligence and Law* 2012;20(3):215–319.
57 Günther J. *Issues of privacy and electronic personhood in robotics. 2012 IEEE RO-MAN: the 21st IEEE International Symposium on robot and human interactive communication*, 2012: 815–20.
58 Price N. Medical malpractice and black-box medicine, in big data, health law, and bioethics. In: Cohen IG, ed. Cambridge: Cambridge University Press, 2018: 295–305.
59 Daniel G*et al*. *Current state and near-term priorities for AI-Enabled diagnostic support software in health care*. Duke Margolis Center for Health Policy., 2019.
60 Beck S. Dealing with the diffusion of legal responsibility: the case of robotics, in Rethinking responsibility in science and technology. In: Battaglia F, Mukerji N, Nida-Rümelin J, eds. Pisa: Pisa University Press, 2014: 167–81.
61 German Ethics Council. *Big data und Gesundheit. Datensouveränität als informationelle Freiheitsgestaltung*. Berlin: German Ethics Council, 2017.
62 German Ethics Council. *Big data and health — data Sovereignty as the shaping of informational freedom*. Berlin: German Ethics Council, 2018.
63 Braun M, Dabrock P. *Ethische Herausforderungen einer sogenannten Big-Data basierten Medizin*. Zeitschrift für medizinische Ethik,, 2016.
64 Hummel P*et al*. *Sovereignty and data sharing*. 2. ITU Journal: ICT Discoveries, 2018.
65 Dabrock P. Die Würde des Menschen IST granularisierbar. Muss die Grundlage unseres Gemeinwesens neu gedacht werden? *epd-Dokumentation* 2018;22:8–16.
66 Wachter S, Mittelstadt B. A right to reasonable inferences. *Columbia Business Law Review* 2019;2:494–620.
67 Dabrock PRichter M, Hurrelmann K, eds. *Soziale Folgen Der Biomarker-basierten und Big-Data-getriebenen Medizin, in Soziologie von Gesundheit und Krankheit*. Wiesbaden: Springer VS, 2016: 287–300.
68 Castell S, Evans H. *The one-way mirror: public attitudes to commercial access to health data*. London: Ipsos MORI, 2016.
69 Hummel P, Braun M, Dabrock P. Data Donations as Exercises of Sovereignty, in The Ethics of Medical Data Donation. In: Krutzinna J, Floridi L, eds. Springer International Publishing Cham, 2019: 23–54.
70 Shapshay SM. Artificial intelligence: the future of medicine? *JAMA Otolaryngology–Head & Neck Surgery* 2014;140(3).
71 Goldhahn J, Rampton V, Spinas GA. Could artificial intelligence make doctors obsolete? *BMJ* 2018;363:k4563.
72 Coiera E. The fate of medicine in the time of AI. *Lancet* 2018;392(10162):2331–2.