

<https://doi.org/10.1038/s41746-025-01725-9>

Artificial intelligence should genuinely support clinical reasoning and decision making to bridge the translational gap

Kacper Sokol¹ ✉, James Fackler^{2,3} & Julia E. Vogt¹

Artificial intelligence promises to revolutionise medicine, yet its impact remains limited because of the pervasive translational gap. We posit that the prevailing technology-centric approaches underpin this challenge, rendering such systems fundamentally incompatible with clinical practice, specifically diagnostic reasoning and decision making. Instead, we propose a novel sociotechnical conceptualisation of data-driven support tools designed to complement doctors' cognitive and epistemic activities. Crucially, it prioritises real-world impact over superhuman performance on inconsequential benchmarks.

Artificial intelligence (AI) is advancing at a breakneck pace with a promise to overcome numerous real-life challenges across many domains¹. Of particular relevance is medicine, where data-driven tools can lead to better quality of and access to healthcare—especially in resource-scarce regions—by helping with early detection and prevention of diseases as well as delivery of personalised treatments^{2–4}. Specifically, AI has the potential to increase the efficiency of healthcare institutions, abate shortages of medical professionals, aid with managing the demand for care in view of population ageing and lifestyle diseases, alleviate the economic burden of healthcare as well as reduce the recovery time, mortality and morbidity of devastating medical conditions, saving numerous lives on a global scale. However, even the most advanced AI models boasting state-of-the-art or superhuman predictive performance on benchmark tasks have negligible or non-existent benefit—setting aside the technical progress itself—if they are never integrated into clinical practice. While there have been some success stories in this regard, they remain scarce compared to the sheer number of such systems currently being developed^{5–7}.

This phenomenon is a stark manifestation of a *translational barrier* that is ubiquitous in AI for healthcare research^{8–11}. While significant focus remains on advancing predictive performance of such models^{2,10,12–15}, this approach does not appear to offer much progress in terms of AI adoption, except for a very limited range of clinical application domains^{5–7}. Some of the underlying reasons include technical misalignment and incompatibility of such systems with deployment requirements⁸, but frictions at the interface of users and technology as well as societal concerns are more prominent⁹. While research into AI fairness, accountability, robustness, interpretability and the like attempts to address these challenges^{16–18}, progress across these disciplines has thus far not managed to unequivocally overcome the translational barrier^{19–23}. Since such technological advancements on their own do not seem to deliver the anticipated real-life impact, an alternative

sociotechnical approach focused on seamless integration of AI-based systems into clinical practice and their overall acceptability may be more fruitful^{14,24–32}. After all, even rudimentary data-driven tools that offer modest improvements across a disease lifecycle can have bigger impact than strictly more powerful systems if only the former are adopted by clinicians while the latter remain purely a research feat.

In this *Perspective* we outline a promising sociotechnical research direction that could help medical AI models to overcome the translational barrier by realigning their operation with the needs and expectations of doctors as well as the intricate environments in which these systems operate. Achieving these goals requires an *interdisciplinary, human-centred* approach that abandons the *autonomous* view of (artificial or human) intelligence and acknowledges its *social* and *relational* nature^{33,34}. Instead of striving to replace clinicians with undesired, fallible and potentially harmful data-driven automation, we posit that AI systems ought to seamlessly integrate into and augment—as opposed to disrupt—well-established medical workflows as well as real-life reasoning and decision-making processes. Consequently, artificial intelligence can assist healthcare professionals in their everyday tasks, complement their abilities, boost their effectiveness and champion clinical best practice^{8,10,14,23–25,27,29,33–44}.

By recognising insights from *cognitive sciences* and embracing the *systems ecology* of clinical decision making—that is the complex interconnected network of its various facets—we can design a new generation of AI tools. One that supports fundamental cognitive (pertaining to conscious intellectual activity) and epistemic (relating to knowledge) functions of doctors—for example, reasoning under noise and uncertainty—therefore empowers them to make the best judgement given available information. Such systems could, among others, improve consistency of decisions (e.g., by eliminating decision noise), alleviate common reasoning limitations and faults (e.g., arising due to cognitive biases), reduce overall (clinical) errors

¹Department of Computer Science, ETH Zurich, Zurich, Switzerland. ²Department of Anesthesiology and Critical Care Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA. ³Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, MD, USA. ✉e-mail: kacper.sokol@inf.ethz.ch

and mistakes (e.g., resulting from a lapse of judgement) and generally make the underlying thought processes more principled^{2,4,25,27,28,45–52}. Our discussion throughout the rest of this *Perspective* is supported by observations and hurdles from (paediatric) sepsis since this disease offers a representative case study of commonplace reasoning and decision-making challenges in medicine; nonetheless, the arguments we present generalise to other areas of healthcare as well as different (high stakes) domains.

Sepsis is a life-threatening condition that arises when human body injures itself in response to an infection⁵³. It is the third leading cause of death (estimated at over ten million a year) and critical illness worldwide, it is the primary cause of mortality from infection and in hospitals, and its survivorship often entails long-term health problems^{54,55}, not to mention its considerable economic burden^{13,53}. However, since sepsis spans a diverse range of incompletely understood processes, it remains elusive⁵³, especially in children for whom it is just as much of a threat as for adults yet it is far less explored⁵⁶. This is particularly concerning given that many observations from the better-understood adult sepsis do not transfer or generalise to the paediatric population with its six clinically and physiologically distinct subgroups^{57,58}.

Building upon decades of research, paediatric sepsis is currently defined by the rigorous *Phoenix Criteria* established through an international consensus⁵⁶. Its diagnosis is based on suspected or confirmed infection in presence of potentially life-threatening organ dysfunction of the respiratory, cardiovascular, coagulation and/or neurological systems determined by a *Phoenix Sepsis Score* of two or more. Nonetheless, some aspects of this definition remain problematic due to their inherent ambiguity. Chief among them is predicating sepsis on *suspected* infection, which is interpreted as a physician placing an order for a microbiological test; but it is well known that such tests are overutilised⁵⁹, likely leading to sepsis overdiagnosis and consequently antibiotic overtreatment⁶⁰. Presupposing organ dysfunction is also a point of contention as this guideline can be compared to expecting cancer to only be diagnosed after discovering its metastases. Further recognition nuances include the existence of *culture-negative* sepsis, which lacks a generally accepted definition but refers to sepsis caused by an infection that is either undetectable by a bacteria culture test or simply when this result is *assumed* to be a false negative^{61,62}.

Sepsis is thus best described as poorly understood. Its true incidence is unknown, its best treatment strategy is uncertain and attempts over the past two decades to develop new treatments have been largely unsuccessful—we still lack rigorous clinical criteria, biological markers, imaging features and laboratory tests to identify this disease^{13,53}. As it stands, bedside clinicians often struggle to anticipate, identify and treat (paediatric) sepsis given variations in medical guidelines and poor predictive value of many current indicators, creating an urgent need for suitable diagnostic tools that could aid doctors in (less biased and more consistent) decision making and delivery of personalised care^{58,63}.

Medical artificial intelligence adoption challenges

Medicine is uniquely positioned to reap the benefits of the recent progress in artificial intelligence given the high impact of even minute improvements in clinical practice^{3,4}. AI tools are of particular importance to the field of paediatrics, where they have been largely underutilised in the recent past². Building these systems now is especially timely given the increased availability of high-quality, large-scale, real-life data as well as leaps in AI, opening this technology up for many real-world applications¹.

Success stories include automated analysis of medical imaging data—e.g., detection of diabetic retinopathy⁵, classification of skin cancer⁶ and detection of lymph node metastases from breast cancer⁷—enabled by recent advances in deep learning. However, such modelling problems are not necessarily representative of medical workflows since they deal with self-contained tasks whose broader context can often be disregarded. Additionally, many of these success stories pertain to the visual domain. But the accuracy of doctors' diagnostic abilities varies widely between disciplines and tends to be task-specific, with visual specialities—e.g., dermatology, radiology or anatomic pathology—exhibiting a far lower error rate (one to

two per cent) than many other areas of medicine (around fifteen per cent). Such disparities can, among other factors, be attributed to a disproportionate signal-to-noise ratio inherent to different specialities²⁶.

Looking at application of AI-based predictive models and decision-support tools in healthcare through the lens of (paediatric) sepsis offers a more comprehensive perspective. In addition to being prototypical, yet unlike other illnesses, this disease is multifaceted and provides enough depth and complexity to elicit real-life desiderata and requirements of such technologies. Specifically:

- sepsis *recognition* is problematic due to the ambiguity surrounding relevant definitions (as explained in the previous section);
- its *management* is impeded by the lack of well-established and universally accepted tools and techniques for gauging patient risk as well as anticipating the disease progression and its severity; and
- the *treatment* of this illness is inconsistent because of the underlying uncertainty as well as lack of reliable mechanisms to systematically monitor patients' response to therapy (i.e., antibiotics). For the paediatric population, its heterogeneity further compounds these issues as they need to be addressed independently for each age group.

With its diverse open challenges and plentiful avenues for improvement, sepsis offers a perfect case study to stimulate and guide the development of novel medical AI systems^{3,4,58,64}. While artificial intelligence techniques have been applied to this disease before, real-life impact of such tools is limited. Data-driven models were used to predict mortality and learn personalised optimal treatment strategies for adults^{13,15,65}; the paediatric population, nonetheless, remains largely neglected with only a few studies modelling sepsis onset and mortality^{58,66–68}. More broadly, AI was used to predict infection as well as assess susceptibility to antibiotics, quantify exposure to them and optimise their choice^{69–71}.

Among these, as well as many other, medical AI systems, traditional supervised and unsupervised models for classification and regression tasks are the most prominent, e.g., answering questions like “Will the treatment be continued in five days?” or “How many more days will the antibiotics be given?” Such practice, however, inadvertently transposes common predictive paradigms onto healthcare applications without considering their suitability or adapting them to fit the underlying, well-established clinical reasoning and decision-making workflows and their broader institutional situatedness². For example, while the evolution of a patient pathway is an inherently continuous process, such a sequence of events is often converted to a classification problem that yields a collection of independent point-in-time predictions about a patient's state in fixed time intervals. To illustrate the pitfalls of this modelling approach consider two patients: one whose health is declining and another who is recovering; at some point in time their state may be captured by the same data point, therefore they will receive an identical, naïve prediction despite one being ready for discharge and the other requiring critical care in the near future.

AI systems that account for temporality—thus are able to answer questions such as “When is the best time to administer antibiotics?” or “In how many days will the patient require critical care?”—are more appropriate, yet broadly underutilised^{2,72,73}. While the core activity of healthcare professionals is to *manage patients' trajectories* by investigating, monitoring and intervening to palliate and cure medical conditions⁸, AI solutions that support such responsibilities, or simply model patient pathways, are largely missing^{74,75}. This is particularly problematic for (paediatric) sepsis since this disease is characterised by a change in the patient's condition rather than their absolute health state; in case of children, this process is represented by an increase in their *Phoenix Sepsis Score*, which reflects progressive organ dysfunction (as explained in the previous section).

Consequently, despite significant technological advancements, healthcare remains one of the least digitised spheres of life with many open challenges^{8,10,76,77}. While the number of technical solutions proposed in the literature has soared in recent years, such contributions largely focus on

developing or adapting general-purpose algorithms to solve narrowly-defined benchmark tasks—e.g., intensive care unit (ICU) mortality or length of stay—that are selected primarily based on data availability and evaluation ease. These systems are also predominantly optimised for predictive performance with the goal of matching or surpassing capabilities of expert clinicians, boasting impressive results—often in synthetic or unrealistic experimental settings—that do not necessarily translate to clinical efficacy or acceptability (among others, due to their inherent misalignment with medical practice)^{10,12,14}.

While artificial intelligence—as a technology—is agnostic of its application, it is fundamentally not a one-size-fits-all solution; applying generic data-driven algorithms to medical challenges simply because relevant data are available is thus unlikely to deliver useful tools⁸. Healthcare requires bespoke AI models tailored to each unique clinical application, especially since incorporating domain-specific knowledge into them tends to enhance their performance and improve their acceptability⁹. The lack of such considerations leads to a mismatch between development/validation and deployment contexts or desiderata, preventing AI tools from being integrated into clinical workflows or simply making them unusable in practice^{8,58}. In case of technical requirements, for example, it is common to presuppose access to biomarkers from medical test results at the time of acquiring a sample as opposed to receiving the corresponding lab report, thus allowing AI to rely on information from the future, which disqualifies it from real-time operation¹³.

When such ill-conceived AI systems—whose functioning is at odds with practical constraints—are deployed, they are often ignored or dismissed by clinicians because of general frustration as well as mistrust, apprehension or aversion towards their outputs^{22,23}. In part, these attitudes can be attributed to AI being overly time-consuming to use, entailing high cognitive burden, failing to deliver the necessary information, disrupting or not integrating well into existing decision-making workflows and the like^{28,29,78}. The lacklustre adoption of such tools is further compounded by pervasive reproducibility issues, history of unsafe systems being deployed prematurely, prevalence of false automation promises as well as scarce data that are inherently private, difficult to collect, store, access or share, and often riddled with numerous biases^{20,79,80}. Ethical concerns and negative societal consequences of deploying data-driven predictive models in healthcare—where they may cause direct harm on a large scale—also stymie their integration into clinical practice. While the algorithmic nature of such techniques streamlines decision-making processes and arguably makes them more objective and equitable by replacing biased and fallible humans⁸¹, these anticipated benefits have time and again not come to fruition or been overshadowed by unforeseen adverse societal impact, disproportionately affecting minorities and people of colour^{9,19,21,82–86}.

From historical biases and discrimination captured in data and perpetuated by automated decisions, through entire populations being underrepresented in training samples leading to predictive performance disparities and unfairness, to modelling reliant on spurious data patterns, building AI models without unintended consequences is a formidable challenge^{82,83,87,88}. Fairness considerations are particularly pertinent in medical applications where protected attributes such as income, gender or ethnicity may be good (proxy) clinical predictors. Healthcare is a high stakes domain that requires thoroughly validated, fair, privacy-preserving, interpretable, reliable, robust and accountable AI systems; they additionally must satisfy the necessary regulatory requirements as well as be compatible with clinical workflows and suitable for real-time operation^{2,16–18,89}. Deploying and maintaining data-driven models to keep them functional, usable and relevant can also be a hurdle given resource constraints like computational infrastructure requirements and running costs. All of these aspects contribute to the aforementioned *translational barrier*—a chasm between technical solutions and clinical applications—which while pervasive is sporadically documented because AI is rarely tested (prospectively) in real-life clinical settings^{8–11,13,15,69–71}.

Systems ecology and artificial intelligence

The factors outlined in the previous section often lead to data-driven tools that are impractical and lack real-world usability, which hampers their adoption^{28,29,58}. This phenomenon is further compounded by the challenge of defining the epistemic task that AI should solve; consequently, predictive models are primarily trained to mimic human decisions since this strategy offers a tractable proxy objective that allows for direct optimisation of predictive performance as well as evaluation and benchmarking against human experts. However, such a reductionist approach that (over)simplifies the role of technology in its designated real-world environment often leads to pervasive lack of *ecological validity*²⁶. The umbrella term of data-driven *decision support* tools used to describe these systems is thus a misnomer since they are rarely designed to genuinely *support* decision making and instead present users with ready-made conclusions that compete with and curtail their own judgement²⁷. The *support* aspect, if any, is mostly confined to justifying predictions with algorithmic explanations and embedding this process within a *human-in-the-loop* interaction protocol that seemingly blends human and machine decisions^{29,30,40}.

This fundamental misalignment between the operationalisation of AI and human decision making leads to undesired automation that biases perception, impedes cognition, limits independent reasoning, inhibits natural exploration and hinders sense making^{29,31,90}. These factors tend to contribute to unwarranted reliance on AI and automation bias, but more notably they erode the value of expertise, disrupt well-established workflows as well as disempower, disenfranchise and displace people instead of supporting them and augmenting their abilities. In the worst case, for predictive models integrated into clinical practice, eliciting such behavioural patterns can reinforce cognitive blunders of doctors, thus undermine the care they provide^{26,91}. This is especially worrying in the context of the *medical diagnostic error* being estimated as one of the most consequential, and often a leading cause of (preventable) death^{26,48}. Notably, the majority of such errors are not rooted in insufficient medical knowledge or inadequate expertise, but rather in *structural causes* that result in deficiencies of medical judgment, with misdiagnosis rate reaching twenty-three per cent in everyday practice for healthcare domains that necessarily rely on high levels of subjectivity, where interobserver variation in diagnosis is to be expected^{47,92,93}.

The aforementioned structural factors include, among others, time pressure, uncertainty and various cognitive biases, all of which lead to problematic synthesis of diagnostic information. The two most prevalent causes are *anchoring*—i.e., steadfastly sticking to an initial impression, thus possibly ignoring subsequent evidence that may be contradictory or disproving—and *premature closure*—i.e., jumping to a conclusion without considering all the available or necessary evidence. These biases seem to affect doctors regardless of their experience, but they appear more common in experts⁴⁷. While medical reasoning and diagnostic error is amply documented in retrospect, detecting and preventing it prospectively is inherently difficult, and largely underexplored in the literature, because it does not manifest as openly as practical mistakes. As a consequence, any data that capture such aspects of clinical work may implicitly encode outcomes of inconsistent or incorrect diagnostic reasoning, with the ensuing AI models inadvertently perpetuating these errors.

In view of these observations, overcoming the translational barrier is likely to require a fundamental change in the design and implementation of artificial intelligence systems. Specifically, their creation ought to be motivated by concrete needs, requirements and challenges faced by human decision makers, helping people to overcome their limitations while also eliciting their strengths^{8,29}. Additionally, the integration of these tools into decision-making workflows should not only be informed by viewing humans as independent agents, but also by recognising their role and placement in the broader context of the processes and environmental constraints in which they operate^{14,24,25,42,43,58}. Such AI systems have the potential to augment specific cognitive tasks, boost comprehension, promote active exploration, stimulate creative problem solving and facilitate (prospective) critical reasoning, thus truly aid and support evidence-driven decision making instead of attempting to “solve” it algorithmically through

disruptive automation. To this end, artificial intelligence could, for example, fill the gaps in people's *knowledge*, allow individuals to challenge automated decisions and then help them to consider and compare alternatives via AI-assisted prospective *mental simulation*, or aid people in progressively *updating their beliefs* to arrive at sound conclusions and decisions²⁴.

The *technological* translational barrier described in the previous section should therefore also be viewed as a *sociotechnical gap*—i.e., the differentiation of what must be supported socially and what can be supported technically^{43,94,95}—overcoming which requires an interdisciplinary approach that draws insights from social and cognitive sciences^{8,10,35,96}. The aforementioned (counterproductive) drive to match or exceed human-level performance in selected (often narrowly- or ill-defined) tasks with the aim of fully automating and replacing humans is thus a manifestation of AI systems being commonly misconstrued as “autonomous rather than social and relational”³³. This paradigm—sometimes referred to as *the race to the bottom* given its intention to remove agency and decision-making power from (individual) humans, shifting the authority to AI and its developers—is nonetheless challenged ever more frequently^{24,34}. Replacing humans with artificial intelligence may not yield the anticipated level of automation but instead shift people from (meaningful and engaging) decisive positions to (frustrating and dreadful) supervisory roles, by and large depriving them of any autonomy and (collective) bargaining power (by making them appear redundant)^{34,40,97}. Such a reconfiguration is particularly ironic for high stakes domains where lifting the perceived limitations and failures of humans with AI often requires those same flawed humans to monitor, interpret, vet and correct computers' output⁹⁸.

In addition to possible bias, discrimination, unfairness and displacement, full automation also raises ethical concerns given unclear attribution of responsibility when an algorithmic decision causes unintended harm. In contrast, the responsibility remains with humans when instead of replacing them, AI augments and aids their decision making by providing them with supporting information to be utilised within well-established frameworks like *evidence-based medicine*²⁷. Similarly, while ignoring or overriding decisions of AI whose performance is supposedly superior to that of human experts could be considered malpractice⁹⁹, such claims are often technically dubious (as argued earlier) and arise primarily from the autonomous view of intelligence⁴². This is not to say that (full automation based on) predictive systems should be discarded altogether, but rather that the adoption of data-driven tools ought to be based on sound justification and robust defence of the process and means via which scientifically grounded and rationally defensible decisions are reached¹⁰⁰.

Consequently, integration of AI cannot be premised on *trust*, e.g., established over prolonged interaction episodes¹⁰¹, as this concept is ill-suited for technology¹⁰²; instead, artificial intelligence should be judged in terms of *reliability* and *robustness*. After all, “if the outcome of a traditional machine becomes unpredictable, we do not think that it is creative or original—we think that it is broken”¹⁰³, and AI should be treated no different. These observations reinforce the relevance of *ante-hoc interpretable* artificial intelligence whose inherent soundness and human intelligibility are guaranteed by means of constraining the underlying model form, e.g., to account for application-specific requirements, making it suitable for high stakes domains like healthcare^{18,104}. This AI transparency paradigm is distinct from more prevalent *post-hoc explainability*, which simplifies opaque predictive systems to make them human-understandable by approximating their operation. The insights output by such methods, however, are not guaranteed to reflect the true behaviour of the underlying models; they may thus be misleading, hence unacceptable in some (safety-critical) applications¹⁸. Out of the two, it is *ante-hoc interpretability* that delivers a solid foundation for building human-centred predictive systems that are *acceptable*—as a result of appropriate social structures—*reliable*—because of sound technical practices—and *safe*—due to open management strategies¹⁰⁵.

In addition to recognising people as *independent agents* interacting with or being replaced by AI, they should also be viewed holistically as members of various societal and organisational structures that connect diverse stakeholders and facilitate their seamless communication and

collaboration⁴³. To avoid disrupting the fragile systems ecology, (interdisciplinary) AI development teams need to identify the best way of integrating this technology into such complex real-life settings; relevant considerations include ensuring its compatibility with environmental constraints, established communication protocols as well as institutional interdependencies, workflows, processes, objectives, desiderata, requirements, regulations, (industry) standards and best practice^{8,14,35–38,58}. Within this landscape, one must account not only for the relation between AI and individuals, but also their groups and the overarching organisations, striving to understand the roles, responsibilities and needs of each unit: how it works, interacts with other units, processes or exchanges information and makes decisions⁴¹. Many such interconnected systems are set up to provide operational frameworks that streamline task execution by division of labour and responsibility. This arrangement allows each unit to treat parts of the process as black boxes (which may only appear so while in principle being comprehensible with suitable expertise¹⁰⁶) that are robust and reliable given the existence of organisational mechanisms to ensure their proper functioning, manage risk and absorb contingencies⁴³.

By considering the broader societal context when building AI tools, an opportunity arises to assist, enhance, support and enable humans to flourish and excel at their work—a strategy that appears more promising than simply attempting to replace them. This vision can be realised not only with full automation of selected tasks, but crucially through human–machine symbiosis, collaboration, co-creation or hybrid intelligence^{8,10,14,25,33,39,40}. In view of the systems ecology, implementing AI within the *distributed cognition* paradigm—where distinct responsibilities are allocated to specialised, algorithmic or human, agents—can be highly beneficial, allowing to support people in their various cognitive and epistemic activities, e.g., comprehension, sense making, reasoning, problem solving, decision making and task execution^{24,42,107}.

As an example, consider doctors' reliance on and widespread integration into clinical workflows of laboratory test results or outputs of advanced medical devices like magnetic resonance imaging (MRI). While in principle these instruments are not black boxes, they can be safely used as such, without in-depth understanding of their inner workings or the underlying chemistry, physics or signal processing principles. This is possible because the responsibility for correct and reliable functioning of these tools has been shifted to appropriate entities: industries (e.g., tasked with device construction), government bodies (e.g., overlooking its certification processes) and professional staff (e.g., entrusted with its calibration and operation)⁴³. Consequently, the existence of these organisational structures streamlines the day-to-day work of doctors.

In medicine, but also elsewhere, AI is perceived no different to such tools²⁵. Its reception is more favourable when it is provided as a digital *partner* that complements, augments, amplifies and supports people's abilities and decision making—e.g., by adding more evidence, compensating for human weaknesses, preventing common biases and overcoming human limitations—rather than replacing human intelligence or reducing the role of people to accepting/rejecting algorithmic recommendations^{8,23}.

Human decision making and artificial intelligence

To overcome the challenges outlined in the previous section, it is first necessary to understand how experts make decisions in highly structured environments and place this process in the context of human–AI dynamics. Under the assumption that replacing people or competing with them is counterproductive at best and harmful at worst^{33,34}, we can identify different modes of artificial intelligence operating vis-à-vis humans². While many fine-grained taxonomies exist¹⁰⁸, distinguishing the following three levels of AI integration suffices for our purposes.

autonomy Tasks that do not require direct human input can be automated, with the resulting artefacts integrated into higher-level workflows as *additional sources of information* or treated as *prescriptive decisions* only to be monitored by people.

assistance (human-in-the-loop) Tasks that require human input can benefit from *descriptive* (summarising information and extracting insights

of interest) or *predictive* (forecasting quantities) *modelling*, with the resulting artefacts supporting problem understanding and decision making. To this end, the insights produced by AI are incorporated into the corresponding workflows as *clues*, *diagnoses* or *recommendations* to be reviewed and accepted or rejected by people.

augmentation (machine-in-the-loop) Tasks whose full or partial automation is undesired or infeasible, thus confining them to the purview of humans, can be streamlined by supporting the higher-level cognitive and epistemic functions of people responsible for their completion. To this end, descriptive and predictive modelling can be integrated into a collaborative co-creation process in which AI systems complement humans' abilities and help them to overcome personal shortcomings and limitations, e.g., by presenting people with a range of possible choices accompanied by their respective positive and negative consequences.

Notably, the most suitable automation paradigm should be selected independently for each separate (cognitive and epistemic) activity such as *data acquisition*, *information analysis* as well as *decision* or *action selection* and *implementation thereof*^{107,109}. The specific form in which AI can be safely and responsibly operated ought to be determined based on the *automation readiness level* of a particular task—a concept that can already be found in data science (*data readiness levels*¹¹⁰), autonomous driving (*levels of automation*¹¹¹) and digital healthcare (*levels of maturity* prescribed by the *analytics adoption model*¹¹²). Implementing data-driven tools in practice, however, remains challenging. This is because we generally lack the corresponding (technical) frameworks, guidelines and protocols that would formalise and lead the development and deployment of these systems. While true for many domains, including medicine²¹, a notable exception is data mining with its CRISP-DM process¹¹³.

This perspective on AI integration is compatible with cognitive and behavioural psychology research, which offers two complementary viewpoints on human decision making: *Naturalistic Decision Making* studies the success of expert judgement, whereas *Heuristics and Biases* (commonly known as the dual-process theory or System 1 and System 2 thinking¹¹⁴) deals with faults in basic reasoning⁸¹. Findings from these disciplines suggest that people's ability to become good decision makers as well as the viable level of automation of a given task depend largely on the properties of the domain in which such processes take place. Settings referred to as *wicked environments* either offer misleading signals or lack reliable cues, regularities and feedback for people to observe, learn from and develop correct and complete intuitions^{81,115}. But where humans flourish, their actions can be studied (through *Naturalistic Decision Making* frameworks) to possibly identify the source of their expertise and codify this knowledge in textbooks or predictive algorithms. In domains where people fail, AI may be able to learn and distil patterns that humans cannot, and use them to make decisions or present them to people in a digestible format^{109,115–117}.

Through such an approach we can recognise environments that offer *sufficient regularities* to be amenable to full or partial automation. This operationalisation of AI is best suited for procedural and repetitive tasks that both machines and humans can complete (on their own) as well as challenges that people struggle with but (data-driven) algorithms can streamline or outright solve. In the former case, the benefit comes from reducing (costly) human errors that arise due to a lapse of judgement, yielding improved (predictive) performance attributed to better decision consistency and efficiency. In the latter scenario, automation makes up for human cognitive deficiencies in tasks that are inherently incompatible with our reasoning capabilities or simply too complex for our minds.

Crucially, both of these AI deployment scenarios presuppose that the *stable-world principle*—sometimes referred to as the *closed-world principle*—holds and that the selected task is *structured*. The former premise implies that the decision-making environment is inherently predictable (given a set of observations) and that it does not evolve unexpectedly (over time), e.g., resulting in a data shift^{42,118,119}. Regarding the latter tenet, we can generally distinguish three decision categories¹⁰⁹:

- *structured* tasks—referred to before—come with a well-defined problem that has a *single* optimal, possibly analytical, solution, which in principle can be found;
- for *unstructured* challenges, *no* universally accepted solution exists given that it depends on individual preferences; and
- *semi-structured* problems have *multiple* viable solutions—each with its own pros and cons—determined according to some predefined criteria and ranking them requires analytical methods, which may include AI, as well as the decision maker's input.

However, evidence-based medical diagnostic reasoning often necessarily relies on incomplete and uncertain information, with some areas of healthcare requiring a high level of subjective judgement and providing outcomes that may not always be fully predictable^{93,120}. This is exactly why evidence-based medicine refrains from prescriptive rules and instead offers best practice guidelines; a decision-making environment as complex as healthcare cannot be easily distilled into rigid procedures that anticipate and encompass all the unique circumstances of individual patients¹⁰⁰. Consequently, semi-structured clinical tasks may not be amenable to (end-to-end) AI modelling as one “optimal” decision or solution that these systems tend to deliver is unlikely to exist in this context. Given the assertive nature of such algorithmic recommendations, they can also stymie human-driven exploration and discovery of alternatives that may prove better in the long term, possibly curtailing the progress of medicine¹⁰⁹. More generally, since predictive models usually optimise for past outcomes, their adoption inadvertently risks hampering scientific serendipity as well as impeding development of new and advancement of existing knowledge⁹¹.

In a stable-world setting, large data quantities and advanced learning algorithms tend to offer unparalleled performance for structured and semi-structured problems (e.g., in the game of chess or go). For open-world tasks, however, simple and inherently transparent models or high-level decision heuristics (both of which can be seen as forms of ante-hoc interpretable AI) can perform on a par with or better than complex data-driven systems (e.g., for predicting heart attack risk)¹¹⁹. Humans are known to rely on such straightforward heuristics and biases in their everyday decision making. When studied in a laboratory setting, these mental processes give rise to seemingly suboptimal, irrational or faulty reasoning as reported by the *Heuristics and Biases* community. Yet when viewed as evolutionary adaptation necessary to deal with complex and *unstable (open-world)* decision-making environments fraught not only with risk but also high degree of unpredictability and uncertainty, these reasoning patterns tend to manifest *ecological rationality* rather than universal defects of cognition¹¹⁹. While such aspects of human decision making are largely overlooked by AI research, they can inspire the design of predictive models—suitable for stable- and unstable-world problems—that offer better utility and acceptability than what is currently available. Among others, this new class of AI systems could help people to boost their comprehension and reasoning abilities in complex environments, increase decision consistency, overcome detrimental cognitive biases, reduce different types of errors and improve overall decision hygiene.

Instances of both reliable and wicked (e.g., due to their unstable and open-world nature) environments can be readily found in medicine. For example, nurses in neonatal ICUs were shown to correctly identify infants developing life-threatening infections (leading to paediatric sepsis) without knowing blood test results, yet they were unable to describe or explain their reasoning¹¹⁷. The *Naturalistic Decision Making* framework has been applied to study individual incidents and uncover the cues, patterns and observations that the nurses relied on, which led to the discovery of novel insights—including infection indicators opposite to those relevant for adults—validated across different hospitals and formalised into an instructional programme to help medical staff spot early signs of neonatal sepsis¹¹⁶. On the other hand, the prevalent uncertainty surrounding various aspects of this disease and its treatment strategy (discussed earlier) is a clear manifestation of the underlying environment being inherently wicked. These circumstances contribute to doctors taking different actions in similar situations

spread over time¹²¹—a phenomenon that in cognitive psychology is called decision *noise*, defined as “undesirable variability in judgments of the same problem”⁸⁹.

In sepsis, early intervention should significantly lower the risk of mortality and morbidity, especially so for the paediatric population in which infections can be extremely fulminant¹²². This belief, for example, motivates research efforts to swiftly identify *serious bacterial infections* so that clinicians can intervene *before* patients develop severe organ dysfunction¹²³. Rapid medical response is thus considered best practice, prompting doctors to administer the most effective treatment—antibiotics—both when the underlying infection is *confirmed* (i.e., clinically proven) or simply *suspected* (i.e., for individuals at risk but not necessarily septic or when *culture-negative sepsis* is surmised)^{63,124,125}. Such an approach, however, leads to many patients—more than ninety-eight per cent of neonates by some estimates—receiving an unnecessary or longer than required treatment, e.g., when the underlying illness is self-limited or infection is disproven¹²². Consequently, sepsis faces yet another challenge: antibiotic overtreatment and its plentiful long-term dangers¹²⁶.

Trading off outcomes and time by preferring immediate results—e.g., the perceived safety offered by antibiotics—in favour of potential future benefits—e.g., preventing antimicrobial resistance—is captured by the *time preference theory*^{127,128}. This “bias towards short-term rewards” together with confirmation bias, entrenchment, in-group favouritism, limited attention span and restricted short-term (working) memory are just some cognitive biases and heuristics that may contribute to humans inadvertently making suboptimal decisions^{25,47,49,50}. Supporting doctors’ cognitive and epistemic functions with AI can thus prove more beneficial than building predictive models optimised to (naïvely) mimic and improve upon their actions⁵⁸.

From benchmark to bedside

Investigating the complex and challenging environment posed by sepsis—in which clinicians struggle to consistently make optimal decisions—inspires a different use of artificial intelligence. One where its operation better aligns with the underlying clinical processes, human decision-making workflows and the broader institutional situatedness thereof; and one where its functioning respects the abilities of doctors and caters to the needs of their cognitive and epistemic activities. Adhering to these principles has the potential to deliver a technology that is readily accepted and adopted by clinicians, especially if it embraces, and does not disrupt, the overarching systems ecology, hence seamlessly blends into existing structures instead of being provided as a standalone tool^{58,129}. Since artificial and human intelligence each exhibits distinct capabilities, which tend to complement one another, the former can augment the latter (instead of replacing it) in the form of hybrid intelligence^{25,27}. This integration may span different stages of reasoning and decision making that arise throughout patient care, e.g., perception, comprehension, cognition and operation, addressing their respective unique desiderata^{2,48,107,130}. In particular, we can draw inspiration from insights, tools and techniques in cognitive sciences that help human experts make good and reliable decisions in highly structured environments such as clinical practice^{50,92}.

On a high level, these approaches strive to increase the knowledge and expertise of doctors, offer them situational help as well as improve their critical thinking and reasoning processes⁹². Crucially, these techniques can be naturally embodied and enhanced by AI tools with the aim of providing advanced cognitive support, preventing common biases, alleviating decision-making challenges, overcoming human cognitive limitations and delivering contextual data-driven insights. Such intelligent systems promise to make up for the shortcomings of these processes that commonly arise due to various environmental factors and improve the overall decision hygiene on multiple levels^{25,28,48}. Rudimentary research in this direction has explored the influence of selected AI explanation types on human cognitive biases, constraints and reasoning faults, demonstrating that while some can be mitigated, others may inadvertently be exacerbated^{131–133}. However, this line of work by and large overlooks the broader systems ecology and remains confined to explainability of predictive models that mimic human decision making¹³⁴.

But in this context, AI integration possibilities are much broader, with explainability research poised to offer a viable implementation framework¹³⁴. Among others, artificial intelligence tools could^{24,27,45,46,48,49,51,52}:

- distil high-level human-comprehensible concepts from data, aid in pattern recognition as well as generate and test hypotheses;
- facilitate ideation and evaluation of ideas as well as forward projections and prospective reasoning by supporting mental simulation;
- prompt reasoning by analogies and counterexamples as well as identify incongruent, ambiguous or atypical manifestations of modelled phenomena (e.g., when evidence captured by data does not align);
- clearly indicate the context of each output by grounding it in domain knowledge, emphasising relevant clinical guidelines and offering pieces of knowledge that the target audience may be lacking (e.g., information about drug interactions); and
- stimulate metacognition through self-monitoring, self-critique as well as self-policing in the long term.

One such specific approach is *cognitive forcing*, which encompasses an array of (intervention) techniques intended to disrupt heuristic reasoning and trigger analytical thinking, thus prompt people to account for overlooked or disconfirming evidence, competing hypotheses and opposing ideas. These methods reduce overconfidence, decrease reliance on hunches and intuitions as well as improve reasoning quality and decision reliability^{48,51}. Another relevant strategy is to facilitate and encourage continuous, as opposed to one-off, decision making, where people explicitly account for the incidence of a given phenomenon (i.e., its base rate) and progressively *update their beliefs* by considering standalone insights along with their relevance/salience⁵². Moreover, humans tend to perform better over time when they are provided with (immediate) *feedback* or are prompted to perform post-factum evaluation of situations that they have encountered and decisions that they have taken. This approach together with the other aforementioned decision-making strategies facilitate and stimulate long-term learning and expertise development.

These mechanisms can, for example, be delivered as part of an AI toolkit that supports clinicians in *mental simulation*. Specifically, this type of technology could empower doctors to explicitly project possible future trajectories and states of patients, allowing them to test the hypothetical effects and implications of different scenarios (e.g., realisation of selected parameters) as well as aid them in comprehension and comparison of these pathways. Notably, such a simulation-oriented system is capable of embracing *missing* and *unknown patient information*—instead of “solving” this challenge via, potentially dubious, technical means—by outputting alternative trajectories (and their likelihood) conditioned on different values of a selected variable, e.g., a medical test result.

More broadly, this conceptualisation of artificial intelligence is able to support, and explicitly encourage, planning for multiple probable (future) outcomes rather than a single one chosen (possibly without robust justification) either based on an AI recommendation or a personal conviction. Recognising the contingency of different courses of action upon major critical junctures and branching points is beneficial as it promotes decisions that remain valid for the highest number of potential scenarios as well as solutions that can be easily adapted to unlikely situations^{135,136}. Such an AI tool could allow doctors to account for the uncertainty of the clinical environment as well as its dynamics, thus increase the robustness of their decision making by prompting them to explicitly consider the possible deviations, complications and unexpected events at each step of a patient pathway. This type of technological support has the potential to decrease the cognitive fatigue of clinicians, reducing their overconfidence and errors as well as improving their performance and decision consistency (thus overcoming bias and noise) in the demanding, high-paced and stressful healthcare setting.

In addition to directly benefiting clinical practice, the simulation capability of AI tools can also be used for *training* in a safe decision-making environment, offering plentiful chances for mental rehearsal, practice and learning, e.g., by replaying landmark case studies⁴⁷. This functionality may alternatively be used to provide access to *digital twins of individual clinicians*

that can be consulted for “second opinions”, compared across to identify points of disagreement leading to new insights, or used as benchmarks for junior doctors to learn from. This particular application of AI can prove especially potent when dealing with heterogeneous patient populations—as is the case for paediatric sepsis—where some clinicians may unknowingly exhibit better judgement for selected demographics. Combining many such independent “opinions” also facilitates *the wisdom of the crowd* approach, which is a strategy for estimating an answer to a question by polling diverse individuals whose aggregate response is a better approximation of the “correct” answer. A specific instantiation of this process called the *Delphi method* can be used by a group of people to arrive at a consensus; for example, clinicians rely on this technique to define and characterise complex medical conditions such as sepsis^{53,55–57,63}.

Conclusion

Our *Perspective* proposes a radical shift in how artificial intelligence should be operationalised, away from the narrow focus on pursuing superhuman or state-of-the-art performance—however defined—and instead towards supporting people’s cognitive and epistemic functions such as perception, reasoning and decision making. This view deprives AI of any special status or autonomy and intentionally decentres it, positioning this technology just as a supporting tool at the disposal of humans. Exactly thirty years after Collen concluded in his historical survey of medical informatics that “developing a comprehensive medical information system [appears to be] a more complex task than putting a man on the moon had been”¹³⁷ we are witnessing an explosion in AI’s capabilities, yet on many planes its current conceptualisation does not seem to bring us any closer to this goal. The vision put forth by this *Perspective*—far from exhaustive in itself—presents a different avenue for implementing artificial intelligence in medicine and beyond; one that unlocks capabilities and benefits unavailable when a data-driven model simply *decides* and its explainability *legitimises*.

Built atop insights from cognitive sciences, our *Perspective* seeks to seamlessly integrate AI tools into the broader systems ecology, aligning them with human needs and expectations as well as established real-life decision-making protocols and organisational workflows, taking care not to disrupt these (often fragile) structures. Instead of replacing people with undesired, fallible and potentially harmful automation, this novel approach empowers humans to make the best judgement given available information. In particular, it aims to promote best (clinical) decision-making practice and alleviate common reasoning shortcomings (e.g., arising due to cognitive biases), making these processes more factual, evidence-based and principled. It strives to achieve these goals by minimising decision errors as well as improving the consistency of and eliminating any undesired variability (i.e., noise) in human judgement.

To this end, our *Perspective* envisions AI assisting clinicians, and more generally the broader population, in everyday cognitive and epistemic tasks with which they commonly struggle or that exceed their capabilities. Providing doctors with timely and relevant (data-driven) insights can complement their expertise, boost their effectiveness and increase the likelihood of them reaching a correct diagnosis, leading to an overall improvement in decision making. To enhance and augment the abilities of clinicians, AI could, among others, illustrate the objective risk of various factors, present short- and long-term consequences of specific actions and communicate past (factual) decisions and their outcomes (e.g., to combat the time preference bias and undesired judgement variability). Artificial intelligence could also support clinicians in hypothesising about the effects and implications of different scenario—accounting for the uncertainty of the environment—including any unlikely complications that they ought to anticipate, yielding more robust diagnoses.

The main purpose of these algorithmic insights is to help doctors better understand what cues they need to look out for and what interventions they should consider implementing with the aim of alleviating the symptoms, improving the health state and managing the trajectories of their patients. Given the source and nature of this information, the task of interpreting and contextualising it remains strictly within the clinicians’ remit. Its safe

incorporation into diagnostic reasoning processes is therefore the responsibility of individual doctors, as is currently the case with non-AI medical devices. Facilitating such integration of systems based on artificial intelligence, however, requires the underlying predictive models to be sound, reliable and robust so that they may be used as tools whose inner workings can remain opaque to their operators¹⁰⁶; ante-hoc interpretability appears to offer a promising AI paradigm to achieve this goal^{18,104}. More prosaic technological considerations include application-specific technical and operational requirements, e.g., availability of clinical variables for real-time deployment.

Clearly, advancing such systems along *all* their sociotechnical dimensions is most likely to help them overcome the pervasive translational barrier found in medical artificial intelligence research. Doing so within the framework introduced in this *Perspective* is particularly promising. First and foremost, the key tenets of our approach agree with core beliefs and views of practising clinicians about the use of artificial intelligence in medicine¹³⁸. Our conceptualisation of AI is also crafted to closely align with the aforementioned “social and relational rather than autonomous” view of (human) intelligence. Additionally, our envisaged operationalisation of AI positions it as a partner for people to collaborate with or a tool at their disposal rather than their competitor or replacement, suggesting its favourable reception. Lastly, our approach keeps the allocation of decision responsibility with humans and preserves their autonomy, attesting to its strong sociotechnical foundations. All of these principles—firmly grounded in decades of interdisciplinary research—promise to increase acceptability and facilitate seamless integration of artificial intelligence in clinical practice for real-world impact.

When deployed and used, these systems can reduce mortality and morbidity of various diseases (besides alleviating their economic burden) by personalising treatments as well as minimising diagnostic and decision-making errors. Regarding paediatric sepsis, they could⁵⁸:

- aid with its *detection* in view of this population’s heterogeneity and the uncertainty surrounding this disease;
- improve its *management* given the lack of suitable (data-driven predictive) tools to assess its severity and monitor its progression; and
- make its *treatment* more consistent to reduce unnecessary exposure to antibiotics.

To this end, such AI systems could, for example, be integrated into the fabric of multidisciplinary *sepsis teams*¹³⁹ and (bedside) *sepsis huddles*¹⁴⁰, supporting the epistemic and cognitive functions of their participants. More broadly, these tools have the potential to generate new (scientific) knowledge and insights in (bio)medicine and healthcare as well as other disciplines amenable to AI modelling.

Data availability

No datasets were generated or analysed in this study.

Received: 11 December 2024; Accepted: 15 May 2025;

Published online: 10 June 2025

References

1. Sarker, I. H. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Sci.* **2**, 160 (2021).
2. Chang, A. C. *Intelligence-based medicine: Artificial intelligence and human cognition in clinical medicine and healthcare* (Academic Press, Cambridge, MA, USA, 2020).
3. Bica, I., Alaa, A. M., Lambert, C. & van der Schaar, M. From real-world patient data to individualized treatment effects using machine learning: Current and future methods to address underlying challenges. *Clin. Pharmacol. Therapeutics* **109**, 87–100 (2021).
4. Johnson, M. et al. The potential and pitfalls of artificial intelligence in clinical pharmacology. *CPT: Pharmacomet. Syst. Pharmacol.* **12**, 279–284 (2023).
5. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).

6. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
7. Golden, J. A. Deep learning algorithms for detection of lymph node metastases from breast cancer: Helping artificial intelligence be seen. *JAMA* **318**, 2184–2186 (2017).
8. Berg, M. Patient care information systems and health care work: A sociotechnical approach. *Int. J. Med. Inform.* **55**, 87–101 (1999).
9. Wiens, J. et al. Do no harm: A roadmap for responsible machine learning for health care. *Nat. Med.* **25**, 1337–1340 (2019).
10. Wardi, G. et al. Bringing the promise of artificial intelligence to critical care: What the experience with sepsis analytics can teach us. *Crit. Care Med* **51**, 985–991 (2023).
11. Markowetz, F. All models are wrong and yours are useless: Making clinical prediction models impactful for patients. *npj Precis. Oncol.* **8**, 54 (2024).
12. Wang, Y. The theoretical framework of cognitive informatics. *Int. J. Cogn. Inform. Nat. Intell. (IJCINI)* **1**, 1–27 (2007).
13. Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C. & Faisal, A. A. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat. Med.* **24**, 1716–1720 (2018).
14. Topol, E. J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
15. Tsirtsis, S. & Rodriguez, M. Finding counterfactually optimal action sequences in continuous state spaces. *Adv. Neural. Inform. Process. Syst* **36**, 3220–3247 (2023).
16. Qayyum, A., Qadir, J., Bilal, M. & Al-Fuqaha, A. Secure and robust machine learning for healthcare: A survey. *IEEE Rev. Biomed. Eng.* **14**, 156–180 (2020).
17. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv. (CSUR)* **54**, 1–35 (2021).
18. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
19. Chakravorti, B. Why AI failed to live up to its potential during the pandemic. *Harvard Business Review* (2022).
20. Ghassemi, M. & Nsoesie, E. O. In medicine, how do we machine learn anything real? *Patterns* **3**, 100392 (2022).
21. Volovici, V., Syn, N. L., Ercole, A., Zhao, J. J. & Liu, N. Steps to avoid overuse and misuse of machine learning in clinical research. *Nat. Med.* **28**, 1996–1999 (2022).
22. Tricco, A. C. et al. Implemented machine learning tools to inform decision-making for patient care in hospital settings: A scoping review. *BMJ Open* **13**, e065845 (2023).
23. Sivaraman, V., Bukowski, L. A., Levin, J., Kahn, J. M. & Perer, A. Ignore, trust, or negotiate: Understanding clinician acceptance of AI-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18 (ACM, 2023).
24. Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A. & Klein, G. Explanation in human–AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. Tech. Rep. AD1073994, Florida Institute for Human and Machine Cognition (2019).
25. Akata, Z. et al. A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer* **53**, 18–28 (2020).
26. Croskerry, P. *The cognitive autopsy: A root cause analysis of medical decision making* (Oxford University Press, Oxford, UK, 2020).
27. van Baalen, S., Boon, M. & Verhoef, P. From clinical decision support to clinical reasoning support systems. *J. Evaluation Clin. Pract.* **27**, 520–528 (2021).
28. Simkute, A., Surana, A., Luger, E., Evans, M. & Jones, R. XAI for learning: Narrowing down the digital divide between “new” and “old” experts. In *Adjunct Proceedings of the 2022 Nordic Human–Computer Interaction Conference*, 1–6 (2022).
29. Anjum, B. A conversation with Ken Holstein: Fostering human–AI complementarity. *Ubiquity* **2023**, 1–6 (2023).
30. Susanto, A. P., Lyell, D., Widyantoro, B., Berkovsky, S. & Magrabi, F. Effects of machine learning–based clinical decision support systems on decision-making, care delivery, and patient outcomes: A scoping review. *J. Am. Med. Inform. Assoc.* **30**, 2050–2063 (2023).
31. Wosny, M., Strasser, L. M. & Hastings, J. Experience of health care professionals using digital tools in the hospital: Qualitative systematic review. *JMIR Hum. Factors* **10**, e50357 (2023).
32. Liao, Q. V., Vorvoreanu, M., Subramonyam, H. & Wilcox, L. UX matters: The critical role of UX in responsible AI. *Interactions* **31**, 22–27 (2024).
33. Siddarth, D. et al. How AI fails us. *Justice, Health, and Democracy Impact Initiative & Carr Center for Human Rights Policy*. (The Edmond & Lily Safra Center for Ethics, Harvard University, 2021).
34. Munn, L. *Automation is a myth* (Stanford University Press, Redwood City, CA, USA, 2022).
35. Sheehan, B. et al. Informing the design of clinical decision support services for evaluation of children with minor blunt head trauma in the emergency department: A sociotechnical analysis. *J. Biomed. Inform.* **46**, 905–913 (2013).
36. Winby, S. & Mohrman, S. A. Digital sociotechnical system design. *J. Appl. Behav. Sci.* **54**, 399–423 (2018).
37. Pasmore, W., Winby, S., Mohrman, S. A. & Vanasse, R. Reflections: Sociotechnical systems design and organization change. *J. Change Manag.* **19**, 67–85 (2019).
38. Seeber, I. et al. Machines as teammates: A research agenda on AI in team collaboration. *Inf. Manag.* **57**, 103174 (2020).
39. Shneiderman, B. *Human-centered AI* (Oxford University Press, Oxford, UK, 2022).
40. Miller, T. Explainable AI is dead, long live explainable AI! Hypothesis-driven decision support using evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 333–342 (2023).
41. Herrmann, T. & Pfeiffer, S. Keeping the organization in the loop: A socio-technical extension of human-centered artificial intelligence. *AI Soc.* **38**, 1523–1542 (2023).
42. Ferrario, A., Facchini, A. & Termine, A. Experts or authorities? The strange case of the presumed epistemic superiority of artificial intelligence systems. *Minds Mach.* **34**, 30 (2024).
43. Keenan, B. & Sokol, K. Mind the gap! Bridging explainable artificial intelligence and human understanding with Luhmann’s functional theory of communication. *arXiv preprint arXiv:2302.03460* (2023).
44. Zhou, L. et al. From artificial intelligence (AI) to intelligence augmentation (IA): Design principles, potential risks, and emerging issues. *AIS Trans. Hum.–Computer Interact.* **15**, 111–135 (2023).
45. Patel, V. L. & Groen, G. J. Knowledge based solution strategies in medical reasoning. *Cogn. Sci.* **10**, 91–116 (1986).
46. Ramoni, M., Stefanelli, M., Magnani, L. & Barosi, G. An epistemological framework for medical knowledge-based systems. *IEEE Trans. Syst., Man, Cybern.* **22**, 1361–1375 (1992).
47. Kuhn, G. J. Diagnostic errors. *Academic Emerg. Med.* **9**, 740–750 (2002).
48. Croskerry, P. Cognitive forcing strategies in clinical decisionmaking. *Ann. Emerg. Med.* **41**, 110–120 (2003).
49. Klein, J. G. Five pitfalls in decisions about diagnosis and prescribing. *BMJ* **330**, 781–783 (2005).
50. Groopman, J. E. & Prichard, M. *How doctors think* (Houghton Mifflin, Boston, MA, USA, 2007).
51. Croskerry, P. & Norman, G. Overconfidence in clinical decision making. *Am. J. Med.* **121**, S24–S29 (2008).
52. Tetlock, P. E. & Gardner, D. *Superforecasting: The art and science of prediction* (Random House, New York, NY, USA, 2016).

53. Singer, M. et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA* **315**, 801–810 (2016).
54. Fleischmann-Struzek, C. et al. The global burden of paediatric and neonatal sepsis: A systematic review. *Lancet Respiratory Med.* **6**, 223–230 (2018).
55. Morin, L. et al. The current and future state of pediatric sepsis definitions: An international survey. *Pediatrics* **149**, e2021052565 (2022).
56. Schlapbach, L. J. et al. International consensus criteria for pediatric sepsis and septic shock. *JAMA* **331**, 665–674 (2024).
57. Goldstein, B., Giroir, B. & Randolph, A. International pediatric sepsis consensus conference: Definitions for sepsis and organ dysfunction in pediatrics. *Pediatr. Crit. Care Med.* **6**, 2–8 (2005).
58. Tennant, R. et al. A scoping review on pediatric sepsis prediction technologies in healthcare. *npj Digital Med.* **7**, 1–16 (2024).
59. Woods-Hill, C. Z. et al. Diagnostic stewardship for blood cultures in the pediatric intensive care unit: Lessons in implementation from the BrighT STAR Collaborative. *Antimicrobial Stewardship Healthc. Epidemiol.* **4**, e148 (2024).
60. Chiotos, K. et al. Antibiotic indications and appropriateness in the pediatric intensive care unit: A 10-center point prevalence study. *Clin. Infect. Dis.* **76**, e1021–e1030 (2023).
61. Klingenberg, C., Kornelisse, R. F., Buonocore, G., Maier, R. F. & Stocker, M. Culture-negative early-onset neonatal sepsis – At the crossroad between efficient sepsis care and antimicrobial stewardship. *Front. Pediatrics* **6**, 285 (2018).
62. Lee, R., Al Rifaie, R., Subedi, K. & Coletti, C. Comparative analysis of bacteremic and non-bacteremic sepsis: A retrospective study. *Cureus* **16**, e76418 (2024).
63. Schlapbach, L. J. & Kisson, N. Defining pediatric sepsis. *JAMA Pediatrics* **172**, 313–314 (2018).
64. Elish, M. C. The stakes of uncertainty: Developing and integrating machine learning in clinical care. In *Ethnographic Praxis in Industry Conference Proceedings*, Vol. 2018, 364–380 (Wiley Online Library, 2018).
65. Banerjee, S., Mohammed, A., Wong, H. R., Palaniyar, N. & Kamaleswaran, R. Machine learning identifies complicated sepsis course and subsequent mortality based on 20 genes in peripheral blood immune cells at 24 h post-ICU admission. *Front. Immunol.* **12**, 592303 (2021).
66. Griffin, M. P. & Moorman, J. R. Toward the early diagnosis of neonatal sepsis and sepsis-like illness using novel heart rate analysis. *Pediatrics* **107**, 97–104 (2001).
67. Schlapbach, L. J. et al. Prediction of pediatric sepsis mortality within 1 h of intensive care admission. *Intensive Care Med.* **43**, 1085–1096 (2017).
68. Joshi, R. et al. Predicting neonatal sepsis using features of heart rate variability, respiratory characteristics, and ECG-derived estimates of infant motion. *IEEE J. Biomed. Health Inform.* **24**, 681–692 (2019).
69. Kanjilal, S. et al. A decision algorithm to promote outpatient antimicrobial stewardship for uncomplicated urinary tract infection. *Sci. Transl. Med.* **12**, eaay5067 (2020).
70. Moehring, R. W. et al. Development of a machine learning model using electronic health record data to identify antibiotic use among hospitalized patients. *JAMA Netw. Open* **4**, e213460 (2021).
71. Adams, R. et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nat. Med.* **28**, 1455–1460 (2022).
72. Chicco, D. & Jurman, G. Survival prediction of patients with sepsis from age, sex, and septic episode number alone. *Sci. Rep.* **10**, 17156 (2020).
73. Liu, R., Hunold, K. M., Caterino, J. M. & Zhang, P. Estimating treatment effects for time-to-treatment antibiotic stewardship in sepsis. *Nat. Mach. Intell.* **5**, 421–431 (2023).
74. Beck, M. K. et al. Diagnosis trajectories of prior multi-morbidity predict sepsis mortality. *Sci. Rep.* **6**, 36624 (2016).
75. Allam, A., Feuerriegel, S., Rebhan, M. & Krauthammer, M. Analyzing patient trajectories with artificial intelligence. *J. Med. Internet Res.* **23**, e29812 (2021).
76. Capobianco, E. Data-driven clinical decision processes: It's time. *J. Transl. Med.* **17**, 1–2 (2019).
77. Spatharou, A., Hieronimus, S. & Jenkins, J. Transforming healthcare with AI: The impact on the workforce and organizations. *McKinsey & Company* **10**, 1–131 (2020).
78. Bansal, G. et al. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16 (ACM, 2021).
79. Gibney, E. Could machine learning fuel a reproducibility crisis in science? *Nature* **608**, 250–251 (2022).
80. Sohn, E. The reproducibility issues that haunt health-care AI. *Nature* **613**, 402–403 (2023).
81. Kahneman, D. & Klein, G. Conditions for intuitive expertise: A failure to disagree. *Am. Psychol.* **64**, 515–526 (2009).
82. O'Neil, C. *Weapons of math destruction: How big data increases inequality and threatens democracy* (Crown, New York, NY, USA, 2016).
83. Angwin, J., Larson, J., Mattu, S. & Kirchner, L. Machine bias. In *Ethics of Data and Analytics*, 254–264 (Auerbach Publications, New York, NY, USA, 2022).
84. Destin, J. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of Data and Analytics*, 296–299 (Auerbach Publications, New York, NY, USA, 2022).
85. Carney, T. Robo-debt illegality: The seven veils of failed guarantees of the rule of law? *Alternative Law J.* **44**, 4–10 (2019).
86. Geiger, G. et al. Suspicion machines: Unprecedented experiment on welfare surveillance algorithm reveals discrimination. *Lighthouse Reports* (2023).
87. Barocas, S., Hardt, M. & Narayanan, A. *Fairness and machine learning: Limitations and opportunities* (MIT Press, Cambridge, MA, USA, 2023).
88. Buolamwini, J. & Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 2018 ACM Conference on Fairness, Accountability and Transparency*, 77–91 (PMLR, 2018).
89. Kahneman, D., Sibony, O. & Sunstein, C. R. *Noise: A flaw in human judgment* (Hachette, London, UK, 2021).
90. Byrne, R. M. Good explanations in explainable artificial intelligence (XAI): Evidence from human explanatory reasoning. In *IJCAI*, 6536–6544 (2023).
91. Jin, W., Li, X. & Hamarneh, G. Why is plausibility surprisingly problematic as an XAI criterion? *arXiv preprint arXiv:2303.17707* (2025).
92. Graber, M. L. et al. Cognitive interventions to reduce diagnostic error: A narrative review. *BMJ Qual. Saf.* **21**, 535–557 (2012).
93. Battefeld, D. & Kopp, S. Formalizing cognitive biases in medical diagnostic reasoning. *Proc. 8th Workshop Form. Cogn. Reasoning* 3242, 102–118 (2022).
94. Ackerman, M. S. The intellectual challenge of CSCW: The gap between social requirements and technical feasibility. *Hum.-Computer Interact.* **15**, 179–203 (2000).
95. Ehsan, U., Saha, K., De Choudhury, M. & Riedl, M. O. Charting the sociotechnical gap in explainable AI: A framework to address the gap in XAI. *Proc. ACM Hum.-Computer Interact.* **7**, 1–32 (2023).
96. Patel, V. L. & Cohen, T. A. Clinical cognition and AI: From emulation to symbiosis. In *Intelligent Systems in Medicine and Health: The Role of AI*, 109–133 (Springer, Cham, Switzerland, 2022).
97. Brynjolfsson, E. The Turing trap: The promise & peril of human-like artificial intelligence. In *Augmented Education in the Global Age*, 103–116 (Routledge, New York, NY, USA, 2023).
98. Bainbridge, L. Ironies of automation. *Automatica* **19**, 775–779 (1983).

99. Hacker, P., Krestel, R., Grundmann, S. & Naumann, F. Explainable AI under contract and tort law: Legal incentives and technical challenges. *Artif. Intell. Law* **28**, 415–439 (2020).
100. van Baalen, S. & Boon, M. An epistemological shift: From evidence-based medicine to epistemological responsibility. *J. Evaluation Clin. Pract.* **21**, 433–439 (2015).
101. Tomsett, R. et al. Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns* **1** (2020).
102. Ryan, M. In AI we trust: Ethics, artificial intelligence, and reliability. *Sci. Eng. Ethics* **26**, 2749–2767 (2020).
103. Esposito, E. *Artificial communication: How algorithms produce social intelligence* (MIT Press, Cambridge, MA, USA, 2022).
104. Rudin, C. et al. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surv.* **16**, 1–85 (2022).
105. Shneiderman, B. Human-centered artificial intelligence: Reliable, safe & trustworthy. *Int. J. Hum.-Computer Interact.* **36**, 495–504 (2020).
106. Sokol, K. & Vogt, J. E. (Un)reasonable allure of ante-hoc interpretability for high-stakes domains: Transparency is necessary but insufficient for comprehensibility. In *3rd ICML Workshop on Interpretable Machine Learning in Healthcare* (2023).
107. Parasuraman, R., Sheridan, T. B. & Wickens, C. D. A model for types and levels of human interaction with automation. *IEEE Trans. Syst., Man, Cybern. – Part A: Syst. Hum.* **30**, 286–297 (2000).
108. Sheridan, T. B. & Verplank, W. L. Human and computer control of undersea teleoperators. Tech. Rep. ADA057655, Massachusetts Institute of Technology Cambridge Man–Machine Systems Laboratory (1978).
109. Phillips-Wren, G. Intelligent decision support systems. *Multicriteria Decision Aid and Artificial Intelligence: Links, Theory and Applications* 25–44 (2013).
110. Lawrence, N. D. Data readiness levels. *arXiv preprint arXiv:1705.02245* (2017).
111. United States Department of Transportation: National Highway Traffic Safety Administration (NHTSA). Levels of automation (2022). <https://www.nhtsa.gov/document/levels-automation>.
112. Healthcare Information and Management Systems Society (HIMSS). Adoption model for analytics maturity (AMAM). <https://www.himss.org/what-we-do-solutions/digital-health-transformation/maturity-models/adoption-model-analytics-maturity-amam> (2021).
113. Martínez-Plumed, F. et al. CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering* (2019).
114. Kahneman, D. *Thinking, fast and slow* (Macmillan, London, UK, 2011).
115. Hogarth, R. M. *Educating intuition* (University of Chicago Press, Chicago, IL, USA, 2001).
116. Crandall, B. & Gamblian, V. Guide to early sepsis assessment in the NICU. *Instruction Manual Prepared for the Ohio Department of Development Under the Ohio SBIR Bridge Grant Program by Klein Associates Inc* (1991).
117. Crandall, B. & Getchell-Reiter, K. Critical decision method: A technique for eliciting concrete assessment indicators from the intuition of NICU nurses. *Adv. Nurs. Sci.* **16**, 42–51 (1993).
118. Katsikopoulos, K. V., Simsek, O., Buckmann, M. & Gigerenzer, G. *Classification in the wild: The science and art of transparent decision making* (MIT Press, Cambridge, MA, USA, 2021).
119. Gigerenzer, G. Psychological AI: Designing algorithms informed by human psychology. *Perspect. Psychol. Sci.* **19**, 839–848 (2023).
120. Dowding, D. & Thompson, C. Evidence-based decisions: The role of decision analysis. In *Essential Decision Making and Clinical Judgement for Nurses*, 173–195 (Churchill Livingstone, London, UK, 2009).
121. Al-Azzawi, R., Halvorsen, P. A. & Risør, T. Context and general practitioner decision-making – A scoping review of contextual influence on antibiotic prescribing. *BMC Fam. Pract.* **22**, 1–10 (2021).
122. Stocker, M. et al. Less is more: Antibiotics at the beginning of life. *Nat. Commun.* **14**, 2423 (2023).
123. Martin, B., DeWitt, P. E., Scott, H. F., Parker, S. & Bennett, T. D. Machine learning approach to predicting absence of serious bacterial infection at PICU admission. *Hospital Pediatrics* **12**, 590–603 (2022).
124. Cabral, C., Lucas, P. J., Ingram, J., Hay, A. D. & Horwood, J. “It’s safer to.” Parent consulting and clinician antibiotic prescribing decisions for children with respiratory tract infections: An analysis across four qualitative studies. *Soc. Sci. Med.* **136**, 156–164 (2015).
125. Fontela, P. S. et al. Clinical reasoning behind antibiotic use in PICUs: A qualitative study. *Pediatr. Crit. Care Med.* **23**, e126–e135 (2022).
126. Llor, C. & Bjerrum, L. Antimicrobial resistance: Risk associated with antibiotic overuse and initiatives to reduce the problem. *Therapeutic Adv. Drug Saf.* **5**, 229–241 (2014).
127. Frederick, S., Loewenstein, G. & O’Donoghue, T. Time discounting and time preference: A critical review. *J. Economic Lit.* **40**, 351–401 (2002).
128. Mahboub-Ahari, A., Pourreza, A., Akbari Sari, A., Rahimi Foroushani, A. & Heydari, H. Stated time preferences for health: A systematic review and meta analysis of private and social discount rates. *J. Res. Health Sci.* **14**, 181–186 (2014).
129. Cook, M. B. & Smallman, H. S. Human factors of the confirmation bias in intelligence analysis: Decision support from graphical evidence landscapes. *Hum. Factors* **50**, 745–754 (2008).
130. Corti, L. et al. “It is a moving process”: Understanding the evolution of explainability needs of clinicians in pulmonary medicine. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–21 (ACM, 2024).
131. Wang, D., Yang, Q., Abdul, A. & Lim, B. Y. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15 (ACM, 2019).
132. Buçinca, Z., Malaya, M. B. & Gajos, K. Z. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proc. ACM Hum.-Computer Interact.* **5**, 1–21 (2021).
133. Bertrand, A., Belloum, R., Eagan, J. R. & Maxwell, W. How cognitive biases affect XAI-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 78–91 (2022).
134. Sokol, K. & Vogt, J. E. NeXAI: Naturalistic explainable artificial intelligence for data-driven reasoning and decision support. *arXiv preprint* (2025).
135. Helmer, O. & Rescher, N. On the epistemology of the inexact sciences. *Manag. Sci.* **6**, 25–52 (1959).
136. Sokol, K., Small, E. & Xuan, Y. Navigating explanatory multiverse through counterfactual path geometry. *Mach. Learn.* **114**, 1–33 (2025).
137. Collen, M. F. *A history of medical informatics in the United States, 1950 to 1990* (American Medical Informatics Association, Indianapolis, IN, USA, 1995).
138. Röber, T. E., Goedhart, R. & Birbil, Ş. I. Clinicians’ voice: Fundamental considerations for XAI in healthcare. *arXiv preprint arXiv:2411.04855* (2024).
139. Simon, E. L. et al. Improved hospital mortality rates after the implementation of emergency department sepsis teams. *Am. J. Emerg. Med.* **51**, 218–222 (2022).
140. Currie, K. E., Barry, H., Scanlan, J. M. & Harvey, E. M. Impact of a multidisciplinary sepsis huddle in the emergency department. *Am. J. Emerg. Med.* **64**, 150–154 (2023).

Acknowledgements

This research was supported by the Hasler Foundation, grant number 23082.

Author contributions

Conceptualisation: K.S. Methodology: K.S. Formal analysis: K.S. Investigation: K.S. Writing—original draft: K.S. Writing—review and editing: K.S., J.F., J.E.V. Supervision: J.F., J.E.V. Funding acquisition: K.S. All authors have read and approved the manuscript.

Funding

Open access funding provided by Swiss Federal Institute of Technology Zurich.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Kacper Sokol.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025