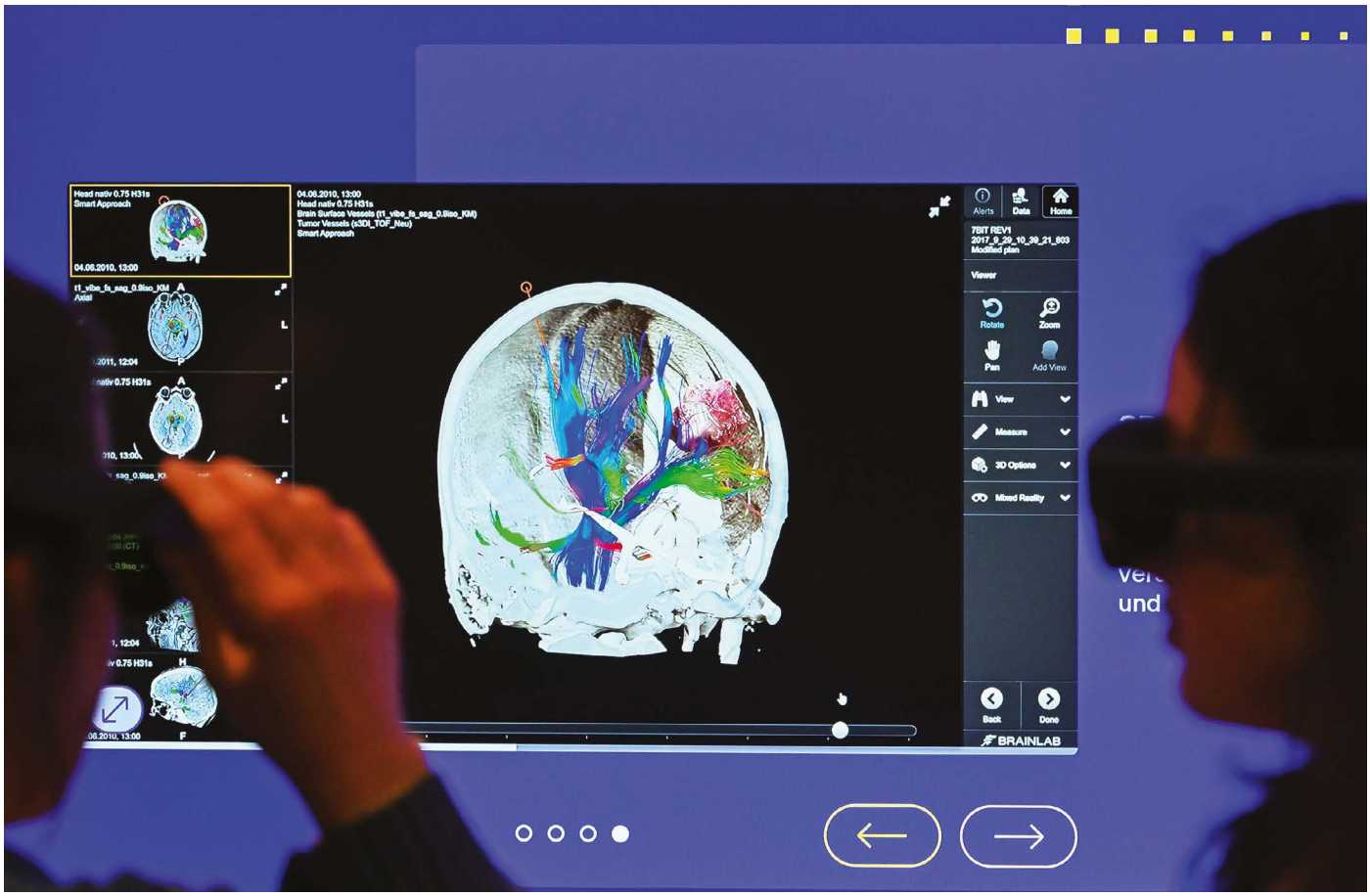
Comment



JENS KALAENE/DPA/ALAMY

Visitors to an interactive AI exhibition at the German Museum of Technology in Berlin use virtual-reality glasses to view an image of the brain.

Train clinical AI to reason like a team of doctors

Christopher R. S. Banerji, Tapabrata Chakraborti, Aya Abdelsalam Ismail, Florian Ostmann & Ben D. MacArthur

As the European Union's Artificial Intelligence Act takes effect, AI systems that mimic how human teams collaborate can improve trust in high-risk situations, such as clinical medicine.

Following a surge of excitement after the launch of the artificial-intelligence (AI) chatbot ChatGPT in November 2022, governments worldwide have been striving to craft policies that will foster AI development while ensuring the technology remains safe and trustworthy. In February, several provisions of the European Union's Artificial Intelligence Act — the world's first comprehensive AI regulation — took effect, prohibiting the deployment of certain applications, such as automated systems that claim to predict crime or infer emotions from facial features.

Most AI systems won't face an outright ban, but will instead be regulated using a risk-based scale, from high to low. Fierce debates are expected over the act's classification of 'high-risk' systems, which will have the strictest oversight. Clearer guidance from the EU will begin emerging in August, but many AI-driven clinical solutions are likely to attract scrutiny owing to the potential harm associated with biased or faulty predictions in a medical setting.

Clinical AI — if deployed with caution — could improve health-care access and outcomes by streamlining hospital management

processes (such as patient scheduling and doctors' note-taking), supporting diagnostics (such as identifying abnormalities in X-rays) and tailoring treatment plans to individual patients. But these benefits come with risks – for instance, the decisions of an AI-driven system cannot always be easily explained, limiting the scope for real-time human oversight.

This matters, because such oversight is explicitly mandated under the act. High-risk systems are required to be transparent and designed so that an overseer can understand their limitations and decide when they should be used (see go.nature.com/3dtgh4x).

By default, compliance will be evaluated using a set of harmonized AI standards, but these are still under development. (Meeting these standards will not be mandatory, but is expected to be the preferred way for most organizations to demonstrate compliance.) However, as yet, there are few established technological ways to fulfil these forthcoming legal requirements.

Here, we propose that new approaches to AI development – based on the standard practices of multidisciplinary medical teams, which communicate across disciplinary boundaries using broad, shared concepts – could support oversight. This dynamic offers a useful blueprint for the next generation of health-focused AI systems that are trusted by health professionals and meet the EU's regulatory expectations.

Collaborating with AI

Clinical decisions, particularly those concerning the management of people with complex conditions, typically take various sources of information into account – from electronic health records and lifestyle factors to blood tests, radiology scans and pathology results. Clinical training, by contrast, is highly specialized, and few individuals can accurately interpret multiple types of specialist medical data (such as both radiology and pathology). Treatment of individuals with complex conditions, such as cancer, is therefore typically managed through multidisciplinary team meetings (known as tumour boards in the United States) at which all of the relevant clinical fields are represented.

Because they involve clinicians from different specialities, multidisciplinary team meetings do not focus on the raw characteristics of each data type, because this knowledge is not shared by the full team. Instead, team members communicate with reference to intermediate 'concepts', which are widely understood. For example, when justifying a proposed treatment course for a tumour, team members are likely to refer to aspects of the disease, such as the tumour site, the cancer stage or grade and the presence of specific patterns of molecular markers. They will also discuss patient-associated features, including age, the presence of other diseases or conditions, body mass index and frailty.

These concepts, which represent interpretable, high-level summaries of the raw data, are the building blocks of human reasoning – the language of clinical debate. They also typically feature in national clinical guidelines for selecting treatments for patients.

> **"Use of the AI tool might place an extra cognitive burden on the clinician."**

Notably, this process of debate using the language of shared concepts is designed to facilitate transparency and collective oversight in a way that parallels the intentions of the EU AI Act. For clinical AI to comply with the act and gain the trust of clinicians, we think that it should mirror these established clinical decision-making processes. Clinical AI – much like clinicians in multidisciplinary teams – should make use of well-defined concepts to justify predictions, instead of just indicating their likelihood.

Explainability crisis

There are two typical approaches to explainable AI[1] – a system that explains its decision-making process. One involves designing the model so it has built-in rules, ensuring transparency from the start. For example, a tool for detecting pneumonia from chest X-rays could assess lung opacity, assign a severity score and classify the case on the basis of predefined thresholds, making its reasoning clear to physicians. The second approach involves analysing the model's decision after it has been made ('post hoc'). This can be done through techniques such as saliency mapping, which highlights the regions of the X-ray that influenced the model's prediction.

However, both approaches have serious limitations[2]. To see why, consider an AI tool that has been trained to help dermatologists to decide whether a mole on the skin is benign or malignant. For each new patient, a post-hoc explainability approach might highlight pixels in the image of the mole that were most important for the model's prediction. This can identify reasoning that is obviously incorrect – for instance, by highlighting pixels in the image that are not related to the mole (such as pen marks or other annotations by clinicians)[3].

When the mole is highlighted, however, it might be difficult[2,4] for an overseeing clinician – even a highly experienced one – to know whether the set of highlighted pixels is clinically meaningful, or simply spuriously associated with diagnosis. In this case, use of the AI tool might place an extra cognitive burden on the clinician.

A rules-based design, however, constrains an AI model's learning to conform rigidly to known principles or causal mechanisms. Yet the tasks for which AI is most likely to be clinically useful do not always conform to simple decision-making processes, or might involve causal mechanisms that combine in inherently complex or counter-intuitive ways. Such rules-based models will not perform well in precisely the cases in which a physician might need the most assistance.

In contrast to these approaches, when a dermatologist explains their diagnosis to a colleague or patient, they tend not to speak about pixels or causal structures. Instead, they make use of easily understood high-level concepts, such as mole asymmetry, border irregularity and colour, to support their diagnosis. Clinicians using AI tools that present such high-level concepts have reported increased trust in the tools' recommendations[5].

In recent years, approaches to explainable AI have been developed that could encode such conceptual reasoning and help to support group decisions. Concept bottleneck models (CBMs) are a promising example[6]. These are trained not only to learn outcomes of interest (such as prognosis or treatment course), but also to include important intermediate concepts (such as tumour stage or grade) that are meaningful to human overseers. These models can thereby provide both an overall prediction and a set of understandable concepts, learnt from the data, that justify model recommendations and support debate among decision makers.

This kind of explainable AI could be particularly useful when addressing complex problems that require harmonization of distinct data types. Moreover, they are well suited to regulatory compliance under the EU AI Act, because they provide transparency in a way that is specifically designed to facilitate human oversight. For example, if a CBM incorrectly assigns an important clinical concept to a given patient (such as predicting an incorrect

tumour stage), then the overseeing clinical team immediately knows not to rely on the AI prediction.

Moreover, because of how CBMs are trained, such concept-level mistakes can also immediately be corrected by the clinical team, allowing the model to 'receive help'[7] and revise its overall prediction and justification with the aid of clinician input. Indeed, CBMs can be trained to expect such human interventions and use them to improve model performance over time.

CBMs have also been developed to take account of 'unknown concepts' – that is, sources of important variation in the data related to outcomes of interest that are not accounted for by known concepts (see go.nature.com/3xtepne). Doing so might improve model accuracy[8] and allow the overseer to assess the extent to which the model's predictions are based on information that cannot be explained by the concepts it was trained to learn.

Predictions that rely heavily on unknown concepts can then be flagged for further investigation or discarded entirely – as is required by the EU AI Act. In the context of a clinical multidisciplinary team meeting, this might signal to the team not only that more data are required to make an effective decision, but also what type of data are needed.

The way forward

Despite the importance of collective decision-making to clinical medicine, group decisions are often fallible. Human factors such as time pressure, decision fatigue and group politics can contribute significantly to decision-making efficacy[9–11]. AI has the potential to ameliorate such human factors. However, to do so in a safe, trustworthy and legally compliant way, it should be attuned to the social, psychological and procedural facets of group debate.

Tools that do so can be built only by strongly collaborative multidisciplinary teams that gather input from the full range of stakeholders and practitioners, alongside technical AI expertise. The teams must also work together throughout the development pipeline, not just at deployment.

Although the need for such multidisciplinary teams in AI development is well known[12], it is not enough to simply recognize this fact and place the burden of building such teams on individual developers. These systemic issues cannot be properly addressed in an ad hoc way; rather, they require coordinated commitment by academic, clinical, regulatory and governmental stakeholders to establish new approaches to education, training, infrastructure development and working culture.

"Institutions should establish career pathways that encourage cross-disciplinary expertise."

Two issues are immediately relevant. First, multidisciplinary communities by their nature require active management and maintenance. Roles such as research community managers – individuals who are highly trained in communication, strategic planning, stakeholder mapping and engagement[13] – are therefore crucial. They should be prioritized when building AI development teams and valued in organizations that do so.

Second, truly sustainable integration of disciplines also requires individuals with cross-disciplinary interests and expertise who can act as connectors between different fields. However, most stakeholder communities, such as academia and clinical medicine, are highly specialized at senior levels – indeed, specialization is often a prerequisite for career progression.

Institutions that are interested in developing successful clinical AI tools should establish career pathways that encourage the development of cross-disciplinary expertise and incentivize individuals to work between conventionally distinct technical and clinical areas.

The EU AI Act provides the impetus to establish these positions and practices. Overcoming the challenges that emerge along the way will require new ways of working. These must recognize the place of AI technologies in the wider clinical context, and provide support for clinical teams so that they can make better-informed decisions for the benefit of all patients.

The European Parliament in Brussels adopted the Artificial Intelligence Act last March.

GEERT VANDEN WIJNGAERT/AP PHOTO/ALAMY

The authors

Christopher R. S. Banerji is theme lead for clinical AI at the Alan Turing Institute, London, UK. **Tapabrata Chakraborti** is theme lead for transparent AI at the Alan Turing Institute, London, UK. **Aya Abdelsalam Ismail** is the chief scientist officer at Guide Labs, San Francisco, California, USA. **Florian Ostmann** is director of AI governance and regulatory innovation at the Alan Turing Institute, London, UK. **Ben D. MacArthur** is a director of science and innovation at the Alan Turing Institute, London, UK.
e-mails: cbanerji@turing.ac.uk; bmacarthur@turing.ac.uk

1. Saranya, A. & Subhashini, R. *Decis. Anal. J.* **7**, 100230 (2023).
2. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. *Lancet Digit. Health* **3**, E745–E750 (2021).
3. Narla, A., Kuprel, B., Sarin, K., Novoa, R. & Ko, J. *J. Invest. Dermatol.* **138**, 2108–2110 (2018).
4. Kindermans, P.-J. *et al.* In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (eds Samek, W. *et al.*) 267–280 (Springer, 2019).
5. Chanda, T. *et al. Nature Commun.* **15**, 524 (2024).
6. Koh, P. W. *et al. Proc. 37th Int. Conf. Mach. Learn.* **119**, 5338–5348 (2020).
7. Zarlenga, M. E. *et al.* Preprint at arXiv https://doi.org/10.48550/arXiv.2309.16928 (2023).
8. Wang, H., Hou, J. & Chen, H. Preprint at arXiv https://doi.org/10.48550/arXiv.2410.15446 (2024).
9. Rosell, L., Alexandersson, N., Hagberg, O. & Nilbert, M. *BMC Health Serv. Res.* **18**, 249 (2018).
10. Soukup, T., Gandamihardja, T. A. K., McInerney, S., Green, J. S. A. & Sevdalis, N. *BMJ Open* **9**, e027303 (2019).
11. Walraven, J. E. W. *et al. BMC Health Serv. Res.* **22**, 829 (2022).
12. Piorkowski, D. *et al. Proc. ACM Hum. Comput. Interact.* **5**, 131 (2021).
13. Sharan, M. *et al.* Preprint at arXiv https://doi.org/10.48550/arXiv.2409.00108 (2024).

The authors declare no competing interests.