Artificial Intelligence is Changing the Ethics of Medicine: Reflections from the Australian Epilepsy Project

Mangor Pedersen[1,2*], Heath R. Pardoe[2], Anton de Weger[2], Donna Hutchison[2], David F. Abbott[2,4], Karin Verspoor[3] & Graeme D. Jackson[2,4,5]

1. Department of Psychology and Neuroscience, Auckland University of Technology (AUT), Auckland, New Zealand
2. The Florey Institute of Neuroscience and Mental Health, The University of Melbourne, Australia.
3. School of Computing Technologies, RMIT University, Melbourne, Australia
4. Department of Medicine, Austin Health, The University of Melbourne, Australia
5. Department of Neurology, Austin Health, Melbourne, Australia.

*Corresponding author: mangor.pedersen@aut.ac.nz

Abstract

Artificial intelligence (AI) is a multidisciplinary scientific field that uses machines to solve real-world problems and predict outcomes. Despite the current enthusiasm about AI's potential as a clinical support tool, there is also a growing awareness and concern about the potentially harmful effects of AI. Because AI will likely impact expert-based decision-making in medicine, it is critical to consider the ethical issues that AI raises in medical research. This paper outlines the AI ethics guidelines of the Australian Epilepsy Project. This large-scale platform aims to democratise specialist care in epilepsy and use AI for clinical decision support based on prospective multimodal datasets (MRI, genetic, clinical, and cognitive data) from thousands of people with epilepsy. As AI develops rapidly, we focus on key areas of medical AI ethics previously identified in the literature, including *Transparency, Justice and Fairness, Non-maleficence, Responsibility,* and *Sustainability.* We believe AI is changing the ethics of medicine, and it is imperative to advance and update ethical guidelines adaptably while preparing for an era of augmented-intelligence-based medicine.

Background

Artificial intelligence (AI) uses computing storage, processing, and knowledge in combination with interactions with environments to understand human intelligence [1]. AI is an exciting, rapidly evolving, and multidisciplinary field of science focused on problem-solving by machines. AI will most likely substantially impact expert-based decision-making in medicine and increasingly realise the challenge of deep phenotyping, which means understanding each individual's uniqueness through detailed characterization based on multimodal data [2]. To process complex data meaningfully, clinicians and AI scientists will benefit from working together to build and validate new models and demonstrate their utility to improve health outcomes [3].

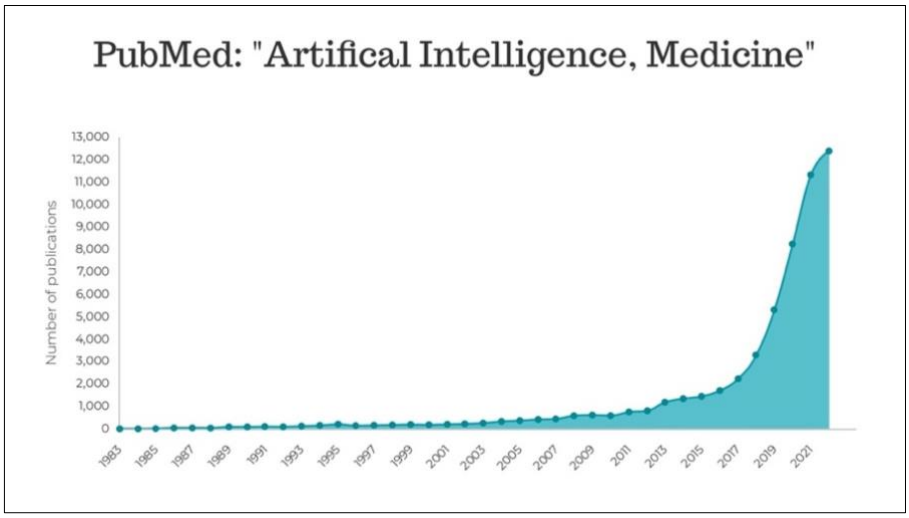

Figure 1: In this figure, we display the number of publications in PubMed (https://pubmed.ncbi.nlm.nih.gov/) for the keywords: Artificial Intelligence, Medicine. It shows a rapid increase in medical publications referring to AI from 2017, at a time of significant scaling of AI models and the computational feasibility of deep learning models. Note that this is a rough approach to estimating the number of studies, not including papers relating to AI (e.g., using terminologies such as machine learning and natural language processing).

Amongst the contemporary excitement around AI's potential, as well as its commercial success (the expected worth of the medical AI market is projected to be US$41.7 billion before 2027[*]) and research success (see the recent rise in medical AI publications in Figure

[*] https://www.reportlinker.com/p06325450/Global-AI-in-Healthcare-Market-Analysis-By-Component-By-Algorithm-By-Application-By-End-User-By-Region-Size-Forecast-with-Impact-Analysis-of-COVID-19-and-Forecast-up-to.html?utm_source=GNW

1), there is an increasing awareness and apprehension about the potentially harmful effects of AI. As new developments in AI are continuously emerging, there are concerns about the potential detrimental impacts of AI: mainly how AI solutions are developed, shared, monitored, and regulated in medicine [4–16].

Given that AI will lead to a change in medical practice, ethical guidelines need to advance to consider ethical risks when developing, implementing, and governing AI algorithms in clinical settings. Next, we outline our large-scale Australian Epilepsy Project that intends to implement novel AI solutions for thousands of people living with epilepsy in Australia.

The Australian Epilepsy Project and our approach to AI Ethics

Epilepsy is more than just seizures. Reduced educational and occupational attainment, decreased independence, lower quality of life, stigma, a higher risk of injury, cognitive deficiencies, mental illness, and suicide are all connected with it [17–22]. *The Australian Epilepsy Project* [23–26] aims to give epilepsy specialists the information they need to make a difference in people's lives, facilitate access to currently limited care, and answer questions that are presently unanswerable. Will this person continue to suffer seizures? Which drug works the best? Can brain surgery be a successful solution? These are fundamental concerns that current medical practice frequently cannot answer.

The Australian Epilepsy Project has several aims that AI can address by integrating multimodal data, including advanced brain imaging, cognitive, and genetics data (see Figure 2). The clinical purposes of the Australian Epilepsy Project include predicting whether a person who has recently suffered their first seizure is likely to experience a second seizure (epilepsy is diagnosed after more than one seizure). After people have been diagnosed with epilepsy, more critical clinical decisions exist. Typically, the first decision for the treating physician following diagnosis is the most appropriate antiseizure medication. In this context, AI could help decide which antiseizure medication is most useful for each patient by providing an individual-specific treatment plan. There is also a subset of patients where seizures arise from a circumscribed part of the brain, and in one third of people with epilepsy medication will not effectively control their seizures [27]. In these people, AI and advanced MRI images can be used to identify subtle epileptogenic lesions not otherwise seen in hospital-based brain imaging.

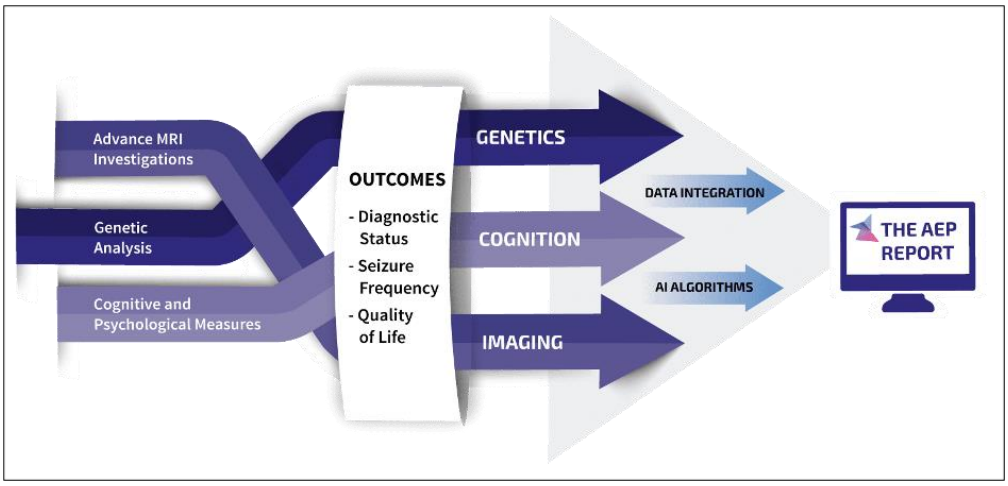

Figure 2: In this figure, we outline the structure of the Australian Epilepsy project, including the data collection and long-term outcomes integral to using AI to improve the lives of people with epilepsy.

A concerning trend within medical AI, identified in a study by Wu et al. [28], highlights the need for large-scale clinical projects like the Australian Epilepsy Project. The authors reviewed the number of FDA-approved AI medical devices between 2017 and 2020. They found that the number of FDA-approved AI devices increased sharply in this period. However, the median sample size of the studies decreased over time, and single-site data was more common than multi-site studies. Most worrying was that most studies were also almost always retrospective. Obtaining quality annotated clinical information and outcome data is challenging. Nagendran et al. [29] also acknowledges that few prospective AI studies exist in medical imaging, reinforcing the idea that the field of AI should move towards larger and prospective studies to enable reliable and generalizable science.

The Australian Epilepsy Project is a *prospective* and *clinically-led* study with a target sample size of *thousands of people* living with epilepsy from *all over Australia*. Participants in the Australian Epilepsy Project will also be tracked two years after study inclusion to provide reliable and clinically valid outcome data for AI classification, such as medical and surgical outcomes. In addition to using AI to create predictions directly related to participant outcomes, the Australian Epilepsy Project intends to use AI in many aspects of the team's workflow and planning. This could include using natural language processing to extract key terms from medical history and radiologist reports, estimating the "Failure to Attend" likelihood to help optimize bookings and minimize dropouts, and rating case similarity between participants to detect trends. Even if AI models are not implemented in production

workflows, data analysis using AI techniques may inform decision-making processes. This use of pattern matching, analytics, and machine learning techniques will be undertaken with due consideration of the ethical AI principles we discuss below. All data used to support project workflow or additional research will be treated and secured like the core study data.

To develop an ethical AI framework for the Australian Epilepsy Project, we have used a data-driven approach, using findings from a systematic review conducted by Jobin et al. [30]. Jobin and others summarised information from 84 publications that covered AI ethics and showed convergence around the following topics in AI ethics, *transparency, justice and fairness, non-maleficence,* and *responsibility,* which we will review next.

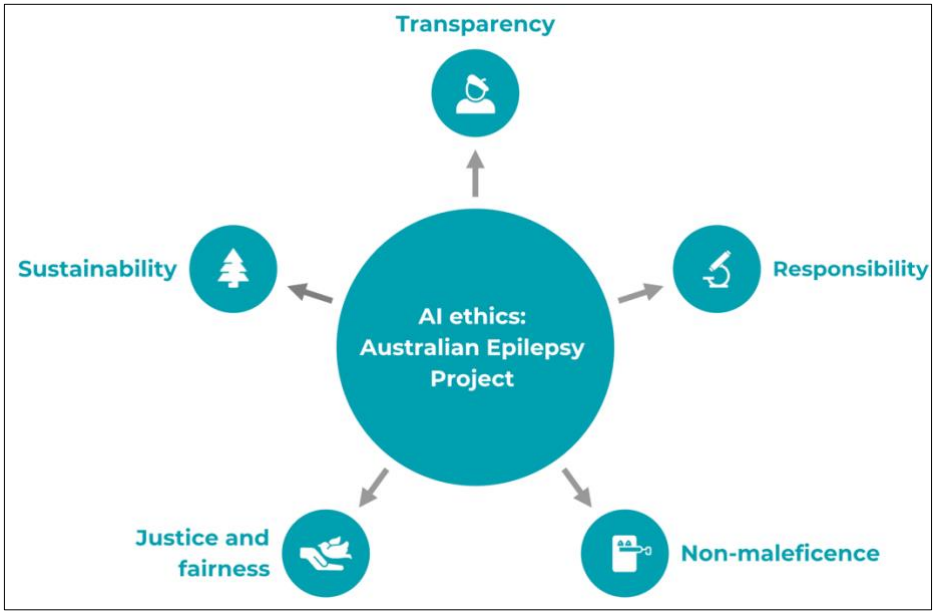

Figure 3: In this figure, we outline the five ethical AI domains that are important to the Australian Epilepsy Project, all of which we take conscious steps to improve iteratively as the field of AI advances.

Transparency:

Transparency was deemed a critical AI ethical topic in 73/84 (87%) previous publications with keywords such as explainability, interpretability, communication, and disclosure [31].

AI-based clinical decisions should be validated, explainable, and reproducible [32–34]. In other words, how does an algorithm reach its conclusion and prediction? Understanding how AI systems arrive at their decisions or recommendations likely requires a shift into a multidisciplinary area of inquiry. Many AI models, including deep learning, are often conceived as a 'black box', and their outputs or results can be challenging to interpret [35].

In epilepsy, identification of epileptogenic lesions increases the likelihood of obtaining a cure for many people with focal epilepsy via brain surgery [36]. Contemporary AI solutions such as deep learning have been explicitly developed for image-based applications and are well suited for detecting neuroanatomical abnormalities [37,38]. An example of a transparent and explainable lesion detection framework is proposed by Spitzer et al. [39], namely the Multi-centre Epilepsy Lesion Detection (MELD) framework. The MELD framework relies on detailed structural information about the brain to obtain a probability of a seizure focus based on structural brain imaging modalities commonly acquired in hospitals and research institutions. In addition to lesion probability, MELD provides interpretable outputs regarding the brain's cortical thickness, curvature, and white and grey matter imaging contrast explaining the model's output. Spitzer et al. [39] validated this approach with 1015 participants (epilepsy and controls) using a split-half training and testing paradigm. Patients with a visible brain lesion –focal cortical dysplasia– were detected with 85% accuracy. The sensitivity and specificity of the MELD approach were lower when the structural features of the seizure focus was unclear. However, we anticipate rapid developments in lesion detection in epilepsy over the next few years, given that AI and deep learning models are ideally suited to detect abnormalities in images. Large-scale neuroimaging data collection efforts such as the Australian Epilepsy Project will be critical to achieving this goal.

Another area in need of improvement in AI is research reporting. Several reporting checklists encourage detailed reporting of AI methodologies, facilitating transparency and replicability. One of the most common reporting frameworks in the field is the Transparent Reporting of studies on prediction models for Individual Prognosis Or Diagnosis (TRIPOD [40]). A systematic review by Nagendram et al. [29] summarised adherence to the TRIPOD framework

in AI studies. The authors found that in most studies reviewed, less than half of the data items in the TRIPOD framework were reported in peer-reviewed publications. These results show that reporting AI standards are inadequate, but a potential issue with the TRIPOD reporting is that it was made in an 'older era' of prediction modelling. The team is working on a TRIPOD-AI checklist for newer AI models [41]. In the meantime, other AI-specific reporting checklists have also been created [42–44]. One of the newer checklists is the Data, Optimization, Model, and Evaluation (DOME) framework [45]. The DOME framework is developed explicitly with contemporary AI and prediction models in mind. It includes a comprehensive list of items in AI research design, train/test/validation dataset split, and model parameter selection. Datasheets documenting key characteristics of datasets can improve transparency and accountability of models built on top [46].

In addition to reporting checklists, registering publications is a way to increase transparency in research. In a registered study, a manuscript submitted in two stages (before and after results are obtained) is pre-accepted for journal publication [47]. Emerging evidence suggests that study bias is reduced in pre-registered reports. Evidence indicates that five times more studies report no statistically significant findings if manuscripts are registered compared to manuscripts that are not registered. This practice also allows for the publication of negative findings [48], providing a more reliable and comprehensive picture of relevant evidence. This is in line with the benefits of clinical trial registration [49] and peer-reviewed publication of systematic review protocols via e.g., PROSPERO [50] or Cochrane (https://www.cochrane.org/), including to avoid selective outcome reporting [51,52].

Non-transparent models are a concerning trend amongst large companies. For example, at the time of writing, Open-AI's large language model, GPT, has not been released for open scientific exploration, nor have adequate details been provided about the resources and models on which it is based to enable meaningful evaluation. This contradicts the idea of safe and transparent AI. There is already substantial fear of AI in the community [53]. [54] has emphasized the critical need for developing adaptive guidelines that govern how AI technologies are used and developed. Legal frameworks will likely struggle to keep up with the pace of AI development. The European Union (EU) has proposed the 'AI-act' as a legal AI framework, as a unified regulatory and legal framework for AI. It will likely change the landscape of AI, ensuring that AI serves people and is a positive contribution to society (https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence).

Justice and fairness:

Justice and fairness were identified as essential topics in 68/84 (81%) publications in Jobin et al. [30] review on ethical AI, with keywords including bias, discrimination, diversity, and accessibility .

Study bias is a common and intrinsic issue in research studies. For instance, clinical trial participation is generally lower among women, people of colour, and older people [56]. Genome-wide association studies have a strong Eurocentric bias, with 79% of GWAS participants of European descent. Polygenic risk scores derived from this data have far lower predictive value than non-European populations [57]. Even in mega-sized community studies such as the UK biobank project, there is evidence of study bias. The UK biobank study aims to provide multimodal data from 500,000 people representing a 'snapshot' of the aging population in the UK between 49-70 years. Nevertheless, in the UK biobank study, the subjects are generally older, more often female, and have a higher socioeconomic status than the general population in the UK [58]. Subjects in the UK biobank study were also less likely to be obese, to smoke, to drink alcohol daily, and to have fewer self-reported health conditions when compared to the general population. The UK Biobank is an excellent initiative and an invaluable resource for extensive population research. Still, given the 'volunteer bias' evidence, it may not represent the sampling population.

Biases in data acquisition also mean that AI algorithms will inevitably be biased. Medical AI exhibits bias, too, often with insufficient generalisation and skewed samples regarding gender, age, racial and ethnic backgrounds, hospitals, and methods related to data collection [59]. Wang and colleagues demonstrated reduced bias if MRI-based AI models include multisource data, including demographic, clinical, genetic, and cognitive scores. The U.S. Food and Drug Administration (FDA) recently presented a draft document requiring overt plans for having diversity in clinical trials (https://www.fda.gov/media/106965/download) which signifies a step toward less biased and community-representative studies. This FDA document outlines how a researcher should describe and explain the rationale behind the anticipated enrolment of individuals from underrepresented groups.

The Australian population is ethnically diverse, which will likely benefit the generalisability of AI models derived from the Australian Epilepsy Project. The project will also collect demographic, gender, race, and other data often associated with biases in large datasets. It

will use this information to reduce potential acquisition bias when training and using AI models. A priority for the Australian Epilepsy Project is to democratise clinical access for people with epilepsy. This means bringing people to research centers with novel technologies and capabilities, such as the ability to acquire advanced imaging including high-resolution neuroanatomy, functional MRI and diffusion imaging scans to image the brain's white matter pathways. Other data collection can now be done in people's homes, including saliva samples for genetic tests which can be sent by mail, and tele-neuropsychology, which is delivered online [26].

Non-maleficence:

Non-maleficence was considered a topic of interest in 60/84 (71%) of previous studies [30]. Keywords related to non-maleficence were security, safety, harm, and protection [60].

A recent example of why governance and ethical regulations are needed in AI comes from a recent report from Urbina et al. [61]. The authors, part of a company developing drug compounds to combat human disease, were invited to a conference hosted by the Swiss government. This conference scrutinizes potential harm and adverse outcomes from AI technologies. Based on this conference, Urbina et al. inverted their original 'good' AI algorithm using chemical compounds to a new model designed to do 'harm'. By transforming the positive algorithm into a harmful one, the authors created 40,000 compounds that could be used in chemical warfare in less than six hours of data training. Many compounds were deemed potentially more dangerous than current warfare agents. The authors acknowledge that they had never previously considered that their AI models with chemical agents could be used to cause harm. This example is an extreme scenario, as working with chemical compounds is potentially risky. Nevertheless, this study highlights the need to consider the known and potentially unknown ethical implications of AI and modelling, especially in medicine, as these AI models influence people's lives and well-being.

A recent open letter signed by some key opinion leaders called for a six-month pause on giant AI experiments that are yielding "unpredictable black-box models with emergent capabilities", urging AI developers to "work with policymakers and regulators to dramatically accelerate the development of robust AI governance systems" (see https://futureoflife.org/open-letter/pause-giant-ai-experiments/). While the AI developments planned by the Australian Epilepsy Project are more limited in scope, the need for oversight

of the use of resultant AI models is well recognized, including the need for appropriate education of clinicians wishing to take advantage of AI-powered decision support tools. Sharing of raw data and AI models are a double edge sword when it comes to AI ethics. It enables transparency, which is good. But it also increases data safety and protection risks, especially disclosing people's identities. Although most people are happy for their data to be publicly shared [62], strategies are needed to ensure safe data sharing, as the re-identification of individuals is a risk when openly releasing data. For example, if a person's whole genome is freely available, re-identifying their identity is possible [63]. It is also possible to re-identify individuals based on MRI of people's faces. Without de-facing MRI images, automated face recognition was 97% accurate in matching participants' real photographs to the correct MRI, and software packages exist to remove MRI features of people's faces before publicly releasing data [64].

Another way to safely release data to the public is to generate synthetic data. This is because synthetic data has no identity and can cause similar features to the original, and real, data. Latent Diffusion Models, such as stable diffusion [65], have overtaken Generative Adversarial Networks [66] in the last year as the primary method for synthetic data generation. Pinaya et al. [67] recently showed that latent diffusion models could generate realistic MRI data using stable diffusion algorithms and released 100,000 synthetic MRI datasets with varying ages and ventricular volumes that can boost existing datasets, enabling large-scale AI approaches [68].

The Australian Epilepsy Project is committed to publicly releasing data to the best standards in the field. When sharing data, we will consider the intended use of the data and the scope of the disclosure to determine how the data will be shared.

Responsibility:

Responsibility in AI was featured in 60/84 (71%) publications reviewed by Jobin et al [30]. Terms around the topic of responsibility included accountability, liability, and acting with integrity [69–71]

In medicine, we work with people's most private information and potentially influence life-changing decisions. The argument of "who is responsible" for AI models and their decision-making is not trivial. Consider a scenario where a novel (and nominally validated) deep learning algorithm has detected a potential brain 'lesion' in a patient with refractory focal

epilepsy. The decision in this scenario was to use the AI-based lesion detection result as a target for brain surgery which is often curative in epilepsy [72]. But, in this case, the surgery on the brain area targeted by the AI algorithm was unsuccessful, and the patient still experienced recurrent seizures. Who is responsible for this outcome? The medical professional or the AI algorithm who 'blindly' identified a brain region that, in hindsight, was the wrong target for surgery? Ultimately, treating physicians are responsible for patients, but it highlights an important consideration as AI-based solutions become more integrated into clinical workflows. Therefore, a pertinent current issue in AI is to determine how much we can trust AI to correctly advise us of the ground truth.

Sand et al. [15] highlights physicians' responsibility to understand AI models used for clinical decision support. They list a set of criteria physicians should consider when employing AI tools in clinical practice (building on previous work by Deitte et al. [73])

In the case of radiologists, required human knowledge includes[*]:

1. Reporting and informing about sensitivity rates and experimental performance
2. Understanding reasonable output
3. Understanding input data (e.g., relationship between image quality and accuracy rate)
4. Awareness of impact of utilizing medical AIs on one's own skills and capacities
5. Awareness of task specificity of the medical AI
6. Assessing, monitoring and reporting of outputs over time

These guidelines empower human decision-making when developing, using, and interpreting AI models and ensure that domain knowledge is imperative to successfully implementing AI in medicine. This argument leads us onto the topic of augmented intelligence [74,75], an emerging field in AI where people and computers symbiotically work together for the best possible and explainable AI outcomes. Augmented intelligence leverages the best aspects of human knowledge and computer algorithms. This includes the responsibility of those who create and implement AI algorithms, as well as policymakers and regulators [76]. It also implies design of human-in-the-loop uses of AI in medicine, where AI is deployed to *support* clinician decision making, rather than *replace* or *emulate* it [77].

The Australian Epilepsy Project is a patient-focused project led by clinicians, and clinical expertise plays an integral role in AI model development and use. We hope our approach

[*] This list is verbatim from Sand et al. [15].

toward augmented intelligence in the Australian Epilepsy Project can serve as a blueprint for other projects that leverage human expertise to generate augmented intelligence-based medicine. We envisage an environment in the Australian Epilepsy Project where AI provides predictions and suggestions, and humans make the decisions.

Other ethical considerations in AI

Less discussed ethical issues in AI, according to Jobin et al. [30] review –highlighted in <50% of AI ethics publications– included *privacy, beneficence, freedom and autonomy, trust, sustainability, dignity, and solidarity.* We found it surprising that sustainability was only considered an important AI ethics topic in 14/84 publications (17%). Although the hope is that AI will ultimately play a significant role in reducing the effects of climate change and other environmental impacts [78], currently, AI leaves a large carbon footprint on the planet [79–81]. For example, ~15% of Google's energy usage was dedicated to AI processing over the last 3 years – and the large language model used to train the GPT-3 model (the model behind ChatGPT), with approximately 175 billion parameters, used 552.1 tons of CO_2-equivalent emissions to train [82]. It is not only training AI algorithms that use large amounts of energy. The majority of Facebook's carbon emission is used for AI model inference –i.e., model adaptation to new data– of their AI algorithms [83] and lifecycle of the Amazon Echo incurs a massive toll on non-renewable materials, labor, and data [84]. From an ethics perspective, is also important to factor in that environmental impacts have disproportionately large effects on marginalised communities [85].

There are several ways to reduce carbon emissions in AI. One of these includes accessing computing resources in regions with cleaner energy. Accessing computing resources in regions powered by more renewable energy sources can reduce emissions by up to 30 times [86]. Another way to reduce carbon emission footprints is to explore whether a research question can be answered with smaller AI models (as large AI models require the most energy [82]). Reporting the sensitivity of models to hyperparameters also allows for understanding training resource requirements [79]. It is furthermore good practice to use frequent checkpoints when training AI models. Frequent checkpoints pause the training and allow the researcher to estimate the model's performance. With checkpointing, errors or sub-optimal performance can be detected early in the training of models, potentially saving time and energy. It is worth noting that checkpointing can require a large amount of storage, and consequent energy consumption, depending on the amount of information to evaluate. A way

to track the carbon footprint of AI models is the CodeCarbon Python package (https://github.com/mlco2/codecarbon) [87,88]. CodeCarbon calculates the electricity consumption from GPU, CPU, and RAM, into a summary statistics of carbon footprint for the AI model.

Table 1. AI ethics goals in the Australian Epilepsy Project

Category	Goals for AI ethics in the Australian Epilepsy Project
Transparency	• Encourage using explainable AI algorithms (e.g., MELD lesion detection approach [39]). • Encourage using AI reporting checklists and registered reports to enhance transparency and reproducibility.
Justice and fairness	• Democratise clinical access for people with epilepsy in Australia. • Aim to monitor for, and remove, unacceptable demographic-related biases in AI models, including multimodal data.
Non-maleficence	• Sharing data to the best standards, with minimal probability of re-identification (e.g., operate within The Five Safes Framework [89])
Sustainability	• Encourage AI checkpoints to evaluate model performance. • Encourage tracking the carbon footprint of AI models (e.g., CodeCarbon Python package [87,88])

Flexibility and adaptability in the fast-changing field of medical AI

The metrics that we use to judge and evaluate AI systems need careful consideration. While model accuracy, calibration, and robustness are common intrinsic evaluation metrics, other dimensions of AI ethics goals, including transparency, efficiency, fairness, and bias, should be quantified and reported. Various frameworks for AI model reporting have been proposed, such as model cards [90] to document the benchmark performance of models under various conditions. Proposed conditions include many directly relevant to medicine (e.g. cultural, demographic, or phenotypic groups). Medical applications of AI may require new metrics, and possibly additional dimensions – e.g. for generative language models, researchers have proposed that toxicity should be added to the list [91]. An additional concern for metric-based evaluation of AI models in a biomedical context is to balance the need for short-term

decisions about optimal model performance with the need for improved long-term health outcomes in patients. Here the adage known as Goodhart's law may be relevant, which is often phrased as "when a measure becomes a target, it ceases to be a good measure". Care must be taken to avoid optimizing AI models for short-term metrics, such as performance on an imaging-based lesion detection task, which may not reflect the ultimate patient-centred goal of achieving improved quality of life. We must continuously mature our approach to evaluation.

In 2023, AI is changing fast, and it sometimes seems futile to predict the next 5 and 10 years of AI development and its impact on medicine. Therefore, we need to have a flexible and adaptable approach to AI ethics that can help address the concerns of various stakeholders. Viewpoints on AI ethics from patients, healthcare providers, regulators, and technology developers are needed to generate adaptive frameworks that support safe AI to benefit health and society. This paper is a starting point toward interpretable, secure, and sustainable AI in the Australian Epilepsy Project (see Table 1). As AI is changing the ethics of medicine, we aim to update ethics guidelines adaptively as we collect data from thousands of people with epilepsy in the following years.

Acknowledgements

This work was supported by an Australian Government Medical Research Future Fund Frontier Health Grant (RFRHPSI000008). MP acknowledges funding from Health Research Council (HRC), New Zealand, Emerging Researcher Grant. D.F.A. acknowledges fellowship funding from the Australian National Imaging Facility.

The Florey Institute of Neuroscience and Mental Health acknowledges the strong support from the Victorian Government and, in particular, the funding from the Operational Infrastructure Support Grant. The authors acknowledge the facilities and scientific and technical assistance of the National Imaging Facility, a National Collaborative Research Infrastructure Strategy (NCRIS) capability, at the Florey Institute of Neuroscience and Mental Health.

References

1. McCarthy, J. What is Artificial Intelligence? (2004).
2. Weng, C., Shah, N. H. & Hripcsak, G. Deep phenotyping: Embracing complexity and temporality—Towards scalability, portability, and interoperability. *Journal of Biomedical Informatics* **105**, 103433 (2020).
3. Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical AI. *Nat Med* **28**, 1773–1784 (2022).
4. Braun, M., Hummel, P., Beck, S. & Dabrock, P. Primer on an ethics of AI-based decision support systems in the clinic. *Journal of Medical Ethics* **47**, e3–e3 (2021).
5. Geis, J. R. *et al.* Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement. *Radiology* **293**, 436–440 (2019).
6. Goldsmith, J. & Burton, E. Why Teaching Ethics to AI Practitioners Is Important. *Proceedings of the AAAI Conference on Artificial Intelligence* **31**, (2017).
7. Gundersen, T. & Bærøe, K. The Future Ethics of Artificial Intelligence in Medicine: Making Sense of Collaborative Models. *Sci Eng Ethics* **28**, 17 (2022).
8. Hamet, P. & Tremblay, J. Artificial intelligence in medicine. *Metabolism* **69**, S36–S40 (2017).
9. Martinho, A., Kroesen, M. & Chorus, C. A healthy debate: Exploring the views of medical doctors on the ethics of artificial intelligence. *Artificial Intelligence in Medicine* **121**, 102190 (2021).
10. McLennan, S. *et al.* Embedded ethics: a proposal for integrating ethics into the development of medical AI. *BMC Med Ethics* **23**, 6 (2022).
11. Morley, J. *et al.* The ethics of AI in health care: A mapping review. *Social Science & Medicine* **260**, 113172 (2020).
12. Muller, H., Mayrhofer, M. T., Van Veen, E.-B. & Holzinger, A. The Ten Commandments of Ethical Medical AI. *Computer* **54**, 119–123 (2021).
13. Naik, N. *et al.* Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility? *Front Surg* **9**, 862322 (2022).
14. Saheb, T., Saheb, T. & Carpenter, D. O. Mapping research strands of ethics of artificial intelligence in healthcare: A bibliometric and content analysis. *Computers in Biology and Medicine* **135**, 104660 (2021).
15. Sand, M., Durán, J. M. & Jongsma, K. R. Responsibility beyond design: Physicians' requirements for ethical medical AI. *Bioethics* **36**, 162–169 (2022).

16. Vayena, E., Blasimme, A. & Cohen, I. G. Machine learning in medicine: Addressing ethical challenges. *PLOS Medicine* **15**, e1002689 (2018).
17. Baker, G. A. Depression and suicide in adolescents with epilepsy. *Neurology* **66**, S5–S12 (2006).
18. Berg, A. T. Epilepsy, Cognition, and Behavior: The clinical picture. *Epilepsia* **52**, 7–12 (2011).
19. Elger, C. E., Helmstaedter, C. & Kurthen, M. Chronic epilepsy and cognition. *The Lancet Neurology* **3**, 663–672 (2004).
20. Kerr, M. P. The impact of epilepsy on patients' lives. *Acta Neurologica Scandinavica* **126**, 1–9 (2012).
21. Motamedi, G. & Meador, K. Epilepsy and cognition. *Epilepsy & Behavior* **4**, 25–38 (2003).
22. Smeets, V. M. J., van Lierop, B. A. G., Vanhoutvin, J. P. G., Aldenkamp, A. P. & Nijhuis, F. J. N. Epilepsy and employment: Literature review. *Epilepsy & Behavior* **10**, 354–362 (2007).
23. Foster, E. *et al.* The costs of epilepsy in Australia: A productivity-based analysis. *Neurology* **95**, e3221–e3231 (2020).
24. Pedersen, M. *et al.* Deep learning for reliable detection of epileptogenic lesions. in *Augmenting Neurological Disorder Prediction and Rehabilitation Using Artificial Intelligence* 163–175 (Elsevier, 2022).
25. Pedersen, M. *et al.* Artificial intelligence for clinical decision support in neurology. *Brain Commun* **2**, (2020).
26. Tailby, C. *et al.* Teleneuropsychology in the time of COVID-19: The experience of The Australian Epilepsy Project. *Seizure* **83**, 89–97 (2020).
27. Kwan, P. & Brodie, M. J. Early identification of refractory epilepsy. *N. Engl. J. Med.* **342**, 314–319 (2000).
28. Wu, E. *et al.* How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat Med* **27**, 582–584 (2021).
29. Nagendran, M. *et al.* Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* **368**, m689 (2020).
30. Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* **1**, 389–399 (2019).
31. Afnan, M. A. M. *et al.* Interpretable, not black-box, artificial intelligence should be used for embryo selection. *Human Reproduction Open* **2021**, hoab040 (2021).

32. Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O. & Weisz, J. D. Expanding Explainability: Towards Social Transparency in AI systems. in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* 1–19 (Association for Computing Machinery, 2021). doi:10.1145/3411764.3445188.
33. Felzmann, H., Fosch-Villaronga, E., Lutz, C. & Tamò-Larrieux, A. Towards Transparency by Design for Artificial Intelligence. *Sci Eng Ethics* **26**, 3333–3361 (2020).
34. Larsson, S. & Heintz, F. Transparency in artificial intelligence. *Internet Policy Review* **9**, 1–16 (2020).
35. Castelvecchi, D. Can we open the black box of AI? *Nature News* **538**, 20 (2016).
36. Baud, M. O. *et al.* European trends in epilepsy surgery. *Neurology* **91**, e96–e106 (2018).
37. Sinclair, B. *et al.* Machine learning approaches for imaging-based prognostication of the outcome of surgery for mesial temporal lobe epilepsy. *Epilepsia* **63**, 1081–1092 (2022).
38. Suzuki, K. Overview of deep learning in medical imaging. *Radiol Phys Technol* **10**, 257–273 (2017).
39. Spitzer, H. *et al.* Interpretable surface-based detection of focal cortical dysplasias: a Multi-centre Epilepsy Lesion Detection study. *Brain* **145**, 3859–3871 (2022).
40. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Medicine* **13**, 1 (2015).
41. Collins, G. S. *et al.* Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* **11**, e048008 (2021).
42. Mongan, J., Moy, L. & Kahn, C. E. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiology: Artificial Intelligence* **2**, e200029 (2020).
43. Norgeot, B. *et al.* Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med* **26**, 1320–1324 (2020).
44. Vasey, B. *et al.* Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* **377**, e070904 (2022).
45. Walsh, I. *et al.* DOME: recommendations for supervised machine learning validation in biology. *Nat Methods* **18**, 1122–1127 (2021).
46. Gebru, T. *et al.* Datasheets for datasets. *Commun. ACM* **64**, 86–92 (2021).
47. Chambers, C. D. & Tzavella, L. The past, present and future of Registered Reports. *Nat Hum Behav* **6**, 29–42 (2022).

48. Allen, C. & Mehler, D. M. A. Open science challenges, benefits and tips in early career and beyond. *PLOS Biology* **17**, e3000246 (2019).
49. Levin, L. A. & Palmer, J. G. Institutional Review Boards Should Require Clinical Trial Registration. *Archives of Internal Medicine* **167**, 1576–1580 (2007).
50. Booth, A. *et al.* An international registry of systematic-review protocols. *The Lancet* **377**, 108–109 (2011).
51. Stewart, L., Moher, D. & Shekelle, P. Why prospective registration of systematic reviews makes sense. *Systematic Reviews* **1**, 7 (2012).
52. Ge, L. *et al.* Association between prospective registration and overall reporting and methodological quality of systematic reviews: a meta-epidemiological study. *Journal of Clinical Epidemiology* **93**, 45–55 (2018).
53. Cugurullo, F. & Acheampong, R. A. Fear of AI: an inquiry into the adoption of autonomous cars in spite of fear, and a theoretical framework for the study of artificial intelligence technology acceptance. *AI & Soc* (2023) doi:10.1007/s00146-022-01598-6.
54. Sanderson, K. GPT-4 is here: what scientists think. *Nature* (2023) doi:10.1038/d41586-023-00816-5.
55. Panch, T., Mattie, H. & Atun, R. Artificial intelligence and algorithmic bias: implications for health systems. *J Glob Health* **9**, 020318.
56. Chastain, D. B. *et al.* Racial Disproportionality in Covid Clinical Trials. *N Engl J Med* **383**, e59 (2020).
57. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* **51**, 584–591 (2019).
58. Alten, S. van, Domingue, B. W., Galama, T. & Marees, A. T. Reweighting the UK Biobank to reflect its underlying sampling population substantially reduces pervasive selection bias due to volunteering. 2022.05.16.22275048 Preprint at https://doi.org/10.1101/2022.05.16.22275048 (2022).
59. Wang, R., Chaudhari, P. & Davatzikos, C. Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies. *Proceedings of the National Academy of Sciences* **120**, e2211613120 (2023).
60. Ellahham, S., Ellahham, N. & Simsekler, M. C. E. Application of Artificial Intelligence in the Health Care Safety Context: Opportunities and Challenges. *Am J Med Qual* **35**, 341–348 (2020).
61. Urbina, F., Lentzos, F., Invernizzi, C. & Ekins, S. Dual use of artificial-intelligence-powered drug discovery. *Nat Mach Intell* **4**, 189–191 (2022).

62. Mello, M. M., Lieou, V. & Goodman, S. N. Clinical Trial Participants' Views of the Risks and Benefits of Data Sharing. *New England Journal of Medicine* **378**, 2202–2211 (2018).
63. Erlich, Y., Shor, T., Pe'er, I. & Carmi, S. Identity inference of genomic data using long-range familial searches. *Science* **362**, 690–694 (2018).
64. Schwarz, C. G. *et al.* Changing the face of neuroimaging research: Comparing a new MRI de-facing technique with popular alternatives. *NeuroImage* **231**, 117845 (2021).
65. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. Preprint at https://doi.org/10.48550/arXiv.2112.10752 (2022).
66. Goodfellow, I. J. *et al.* Generative Adversarial Networks. Preprint at https://doi.org/10.48550/arXiv.1406.2661 (2014).
67. Pinaya, W. H. L. *et al.* Brain Imaging Generation with Latent Diffusion Models. *arXiv.org* https://arxiv.org/abs/2209.07162v1 (2022) doi:10.48550/arXiv.2209.07162.
68. Marquand, A. F., Rezek, I., Buitelaar, J. & Beckmann, C. F. Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. *Biological Psychiatry* **80**, 552–561 (2016).
69. Coeckelbergh, M. Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Sci Eng Ethics* **26**, 2051–2068 (2020).
70. Constantinescu, M., Voinea, C., Uszkai, R. & Vică, C. Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context. *Ethics Inf Technol* **23**, 803–814 (2021).
71. Tigard, D. W. Responsible AI and moral responsibility: a common appreciation. *AI Ethics* **1**, 113–117 (2021).
72. Kral, T. *et al.* Outcome of epilepsy surgery in focal cortical dysplasia. *J Neurol Neurosurg Psychiatry* **74**, 183–188 (2003).
73. Deitte, L. A. *et al.* Entrustable Professional Activities: Ten Things Radiologists Do. *Acad Radiol* **23**, 374–381 (2016).
74. Gennatas, E. D. *et al.* Expert-augmented machine learning. *PNAS* **117**, 4571–4577 (2020).
75. Zheng, N. *et al.* Hybrid-augmented intelligence: collaboration and cognition. *Frontiers Inf Technol Electronic Eng* **18**, 153–179 (2017).
76. Bazoukis, G. *et al.* The inclusion of augmented intelligence in medicine: A framework for successful implementation. *Cell Reports Medicine* **3**, 100485 (2022).

77. Lederman, A., Lederman, R. & Verspoor, K. Tasks as needs: reframing the paradigm of clinical natural language processing research for real-world decision support. *Journal of the American Medical Informatics Association* **29**, 1810–1817 (2022).
78. Cowls, J., Tsamados, A., Taddeo, M. & Floridi, L. The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. *AI & Soc* **38**, 283–307 (2023).
79. Strubell, E., Ganesh, A. & McCallum, A. Energy and Policy Considerations for Deep Learning in NLP. in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* 3645–3650 (Association for Computational Linguistics, 2019). doi:10.18653/v1/P19-1355.
80. Dhar, P. The carbon impact of artificial intelligence. *Nature Machine Intelligence* **2**, 423–425 (2020).
81. Schwartz, R., Dodge, J., Smith, N. A. & Etzioni, O. Green AI. *Commun. ACM* **63**, 54–63 (2020).
82. Patterson, D. *et al.* The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. *Computer* **55**, 18–28 (2022).
83. Wu, C.-J. *et al.* Sustainable AI: Environmental Implications, Challenges and Opportunities. (2021) doi:10.48550/arXiv.2111.00364.
84. Anatomy of an AI System. *Anatomy of an AI System* http://www.anatomyof.ai.
85. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–623 (Association for Computing Machinery, 2021). doi:10.1145/3442188.3445922.
86. Henderson, P. *et al.* Towards the systematic reporting of the energy and carbon footprints of machine learning. *J. Mach. Learn. Res.* **21**, 248:10039-248:10081 (2020).
87. Lottick, K., Susai, S., Friedler, S. A. & Wilson, J. P. Energy Usage Reports: Environmental awareness as part of algorithmic accountability. Preprint at https://doi.org/10.48550/arXiv.1911.08354 (2019).
88. Lacoste, A., Luccioni, A., Schmidt, V. & Dandres, T. Quantifying the Carbon Emissions of Machine Learning. Preprint at https://doi.org/10.48550/arXiv.1910.09700 (2019).
89. Ritchie, F. *Five Safes: designing data access for research*. (2016). doi:10.13140/RG.2.1.3661.1604.

90. Mitchell, M. *et al.* Model Cards for Model Reporting. in *Proceedingsof the Conference on Fairness, Accountability, and Transparency* 220–229 (Association for Computing Machinery, 2019). doi:10.1145/3287560.3287596.
91. Liang, P. *et al.* Holistic Evaluation of Language Models. Preprint at http://arxiv.org/abs/2211.09110 (2022).