# Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine

*Stefan Harrer*

Digital Health Cooperative Research Centre, Melbourne, Australia

eBioMedicine
2023;90: 104512

Published Online xxx
https://doi.org/10.1016/j.ebiom.2023.104512

## Summary

**Large Language Models (LLMs) are a key component of generative artificial intelligence (AI) applications for creating new content including text, imagery, audio, code, and videos in response to textual instructions. Without human oversight, guidance and responsible design and operation, such generative AI applications will remain a party trick with substantial potential for creating and spreading misinformation or harmful and inaccurate content at unprecedented scale. However, if positioned and developed responsibly as companions to humans augmenting but not replacing their role in decision making, knowledge retrieval and other cognitive processes, they could evolve into highly efficient, trustworthy, assistive tools for information management. This perspective describes how such tools could transform data management workflows in healthcare and medicine, explains how the underlying technology works, provides an assessment of risks and limitations, and proposes an ethical, technical, and cultural framework for responsible design, development, and deployment. It seeks to incentivise users, developers, providers, and regulators of generative AI that utilises LLMs to collectively prepare for the transformational role this technology could play in evidence-based sectors.**



Keywords: Generative artificial intelligence; Large language models; Foundation models; AI ethics; Augmented human intelligence; Information management; AI trustworthiness

## A linguistic stunt takes over the world

The feat that LLMs allow to perform goes as follows: a human user inputs a 'prompt' or an iterative series of successive prompts into an application instructing it to produce a certain piece of output. Such prompts are posed in free-flow language and can be phrases consisting of a few simple words in layman's terms up to whole paragraphs giving detailed instructions to the program in complex language–and everything in between. Prominent examples of such technology are OpenAI's chatbot ChatGPT,[1] Google's chatbots LaMDA[2] and Bard,[3] as well as Stability AI's and OpenAI's imagery generators Stable Diffusion[4] and Dall·E.[5] Topical possibilities for prompts are endless and can range from the amusing "Describe how to remove a sandwich from a toaster in the style of T.S Eliot" to consumer health related questions such as "What is the best nutrition plan for a diabetic with high blood pressure?" to prompting LLMs with entire problem subsets of the United States Medical Licensing Examination (USMLE). The LLM, in conjunction with other algorithms (a detailed description of the technology is given in the next section), will instantly and automatically generate a response to the prompt that, in case of text-to-text transformation and judging by style, grammar, presentation, and often also content, seems indistinguishable from the output a human counterpart might have produced. For example, Med-PaLM, an LLM custom developed by Google and DeepMind for the medical field recently reported USMLE accuracy results which started to approach performance levels of human clinical experts[6] and were replicated by ChatGPT.[7] BioGPT, a LLM developed by Microsoft and trained on biomedical data, has achieved human parity in certain tasks of biomedical text generation and mining.[8] Similar outcomes have been reported with other state-of-the-art LLMs for standardized tests in law and education. This capability, paired with the conversational ease of human-machine interaction and the sheer limitless scope of prompt complexity and variability which LLMs can handle is an impressive accomplishment. It lets legacy natural language processing systems such as Siri or Alexa look like stone-age technologies and has given rise to an unprecedented investment and attention frenzy in the research community, commercial industry sectors, media, and the public which has been granted access to many generative AI applications under varying degrees of access limitations.

## The technology explained

In a seminal piece of work, Google introduced a novel type of neural network algorithm, the Transformer model architecture, in 2017.[9] Transformers learn contextual information from sequential data such as, for example, time series data from wearables, videos, textual words in written or spoken language, or other audio

*E-mail address:* stefan.harrer@dhcrc.com.

signals. Transformer architectures became the basis of LLMs which can contain up to trillions of parameters.[10–12] Introduced for text-to-text conversion they were also, with minor fine-tuning and in combination with further algorithms, adapted for generating other data modalities implementing for example text-to-image, text-to-audio, and text-to-video schemes. Because they are trained on a broad set of data, and because they can perform a variety of tasks without having specifically been trained for them–for example writing a user manual, creating a painting in a particular style, or answering questions–LLMs are also called 'Foundation Models'. Their application for the creation of new content has made them a key technology in the field of generative AI.

A look inside the ChatGPT engine room illustrates the role which LLMs play as part of generative AI applications. What happens here can, in different flavours, be found in other systems as well. In layman's terms, this is how the ChatGPT algorithm was built and trained[1]: it first ingested text data scraped off the internet (OpenAI did not disclose the exact training data set) and used that data to find and learn statistical correlations between the relative positions of words. This knowledge enables the model to iteratively predict which words most likely follow when presented with an initially given word or phrase, i.e., the so-called 'prompt'. A first phase of supervised learning, during which human reviewers assessed the correctness of model responses to a select cohort of actual prompts sourced from user data of OpenAI's earlier GPT-3 model, served as the first fine-tuning round. In a second phase, more human prompts were sampled, and for each of them the model produced a selection of different responses. Human reviewers then ranked the relevance of model responses for each individual prompt, and these rankings fed a reward model that was used to further fine-tune the LLM. In a third reinforcement learning round, this process was repeated at scale with even more prompts and the LLM was fine-tuned to its present state as deployed into ChatGPT. This combined supervised, self-supervised and transfer learning approach using large amounts of unlabelled data is emblematic for Foundation models and extends their capability to learn tasks which they have never particularly been trained for.

## An illusion of intelligence: limitations and risks of large language models

As sophisticated as LLM-powered chatbot responses might look, they represent nothing more than the model's extensive statistical knowledge of which words have preceded others in text that it has previously seen. They comprehend none of the language they deal with, neither the prompts they are being fed nor their responses. It is predominantly the power of scale, i.e., the enormous size of training data and model parameters and, in case of models being custom developed for navigating specific subject matter areas such as healthcare and medicine, the targeted curation of training data, that enables LLMs to produce answers which are often correct and convincing. But make no mistake: LLMs do not understand language (or other data modalities) and have therefore been termed 'stochastic parrots' in a 2021 academic landmark paper[13] which stirred a heatedly debated controversy in the world of AI.[14]

This leads to three core limitations with respect to the content of LLM-generated data. Firstly, if models have been trained on a vast corpus of internet data with limited filtering (as for example is the case for ChatGPT or stable diffusion), they have ingested facts as much as misinformation, biased content as much as fair content, harmful materials as much as harmless ones. Without a means to assess any of these criteria before answering a prompt, LLMs are at risk—and numerous examples have demonstrated they have fallen to it—of reproducing, amplifying, and disseminating problematic content and misinformation.[15] At times a LLM might tell the truth or produce relevant, acceptable, at times surprising, creative, and appealing content. At other times it might produce or argue for the most blatant and dangerous piece of misinformation. Secondly, the model has no means to assess by itself, let alone inform the user, which one it is at any given point in time. It does not know whether the material it produces contains falsehoods, misrepresentations, or inappropriate content or whether it tells the truth. Princeton philosopher Harry Frankfurt's infamous New York Times bestseller 'On Bullshit' comes to mind, referenced, and quoted here below:

"Someone who lies and someone who tells the truth are playing on opposite sides, so to speak, in the same game. Each responds to the facts as he understands them, although the response of the one is guided by the authority of the truth, while the response of the other defies that authority and refuses to meet its demands. The bullshitter ignores these demands altogether. He does not reject the authority of the truth, as the liar does, and oppose himself to it. He pays no attention to it at all. By virtue of this, bullshit is a greater enemy of the truth than lies are."[16]

Thirdly, LLMs are probabilistic algorithms, i.e., when prompted with the same task or question multiple times, the model will return different responses which might either be different versions of previously wrong or problematic answers, or replacements of wrong answers with improved or correct ones and vice versa or constitute different versions of previous correct replies or combinations thereof. This behaviour poses a reliability and reproducibility problem requiring continuous human oversight of model operation.

## Health buyers beware: generative AI is an experimental technology not yet ready for primetime

Despite these issues, LLM-driven generative AI is not smoke in the mirror. On the contrary: the fact that some of the applications currently being in the public domain have been released prematurely overshadows but does not take away from the tremendous potential the technology shows for fundamentally transforming countless industries and professions and particularly so, information management, education, and communication in healthcare and medicine.

LLMs could assist the generation of medical reports or preauthorization letters based on keywords. Documentation constitutes a quarter to half of a doctor's time and a fifth of a nurse's time.[17,18] The use of LLMs has been demonstrated to reduce the time clinicians or other healthcare professionals spend on producing documentation content.[19,20] Thus, LLMs show promise to change clinical practice by allowing doctors to spend more time with their patients. LLMs could also help medical students to study more efficiently by generating quality practice questions alongside explanations or by breaking down complex concepts at appropriate levels of detail.[21] Clinician-patient communication often requires simplifying medical jargon, which LLMs could be a useful assistive tool for. This points to a potentially particularly impactful application of LLMs in the field of clinical trial design where their use could yield improved efficiency of clinical trial matching and clinical trial enrichment processes.[22] Electronic health records (EHR) including clinical notes are one of the biggest, fastest-growing and most information-rich data sources in existence yet resist efficient interpretation and knowledge retrieval partly due to a lack of syntactic, structural, and semantic interoperability and standardisation. LLMs could help to overcome these hurdles.[23]

Generative AI has also gained traction in the life sciences and biomedicine fields and shows substantial potential for making drug discovery and design more efficient: David Baker's work on using LLM-type models for novel biomolecule design[24] and biotech start-up Insilico Medicine's promising forays into using generative AI tools for target discovery[25] showcase how generative AI could help to massively speed up the conception and design of novel protein structures as part of an AI-native drug development cycle. Married with one of the most exciting AI breakthroughs of our times, DeepMind's AlphaFold system capable of predicting structure and thus biological function of virtually any protein in existence,[26] we are potentially looking at one of the most transformative and powerful tools humankind will ever have created to understand and navigate health and life sciences.[27]

Silicon Valley VC firm Sequoia Capital has recently enthusiastically laid out its positive view onto generative AI.[28] At the time of publication of this essay, ChatGPT had successfully passed AWS Cloud Certification[29] and Med-PaLM has started to approach human expert level performance on a subset of the USMLE exam.[6] Microsoft is reportedly in negotiations with OpenAI about a multi-billion-dollar follow-up investment in integrating ChatGPT into MS Office and has already embedded it in its search engine Bing.[30] Digital health AI guru Eric Topol has shared his encouraging views on the role LLM-powered generative AI applications could play in healthcare and medicine.[31]

It is tempting to picture the latest generation of LLMs as quasi-omniscient artificial medical doctors seamlessly tapping into the entirety of digitally documented knowledge and digging up a requested needle from that haystack at the glimpse of a prompt. However desirable, this vision suffers from misinterpretations on several levels. As Topol highlights, the role of AI systems is to augment human intelligence and to assist, not replace human decision making and knowledge retrieval. This becomes nowhere clearer than with LLM-based generative AI applications: the essence of efficient knowledge retrieval is to ask the right questions, and the art of critical thinking rests on one's ability to probe responses by assessing their validity against models of the world. LLMs can perform none of these tasks. They are powerful in-betweeners capable of efficiently and creatively narrowing down the vastness of all possible responses to a prompt to the most likely ones. But they cannot assess whether a prompt was meaningful or whether the model's response made any sense. Humans are needed in the loop to complement these shortcomings of LLMs with their own cognitive capabilities.

LLMs should therefore be treated as imperfect tools which promise to offer a tremendous efficiency gain to many workflows but need strict human supervision and action at both operational interfaces, input and output, using the proverbial garbage-in/garbage-out paradigm as yardstick to assess prompt and response quality, relevance, and appropriateness. It is no surprise that one of the fastest growing entrepreneurial markets in generative AI currently evolves around prompt generation and composition.[32] It is intellectually rewarding to see that skilfully asking questions and giving clear instructions now also turns out to be quantifiably commercially rewarding. Generative AI has helped search engines to jump the evolutionary stage to response engines–hence Microsoft's investment into integrating ChatGPT capability into its search engine Bing being a competitive code red alert for Google.

However, concerning reports on dangerous use and misguided hype of several LLM applications in the healthcare and wellness management sector have started to surface. Anecdotal accounts of excited individuals reporting on consulting ChatGPT in health-related matters and taking its feedback at face value regularly show up in social media streams. Meta's botched release of its own self-proclaimed science-specific, LLM-powered

chatbot Galactica produced disturbing pamphlets, offering–among other bogus reports skilfully presented as scientific publication including empirical quantitative research results–a detailed justification for the 'health benefits of eating crushed glass'.[33] And even Google's and DeepMind's more cautiously and responsibly built and released Med-PaLM still yields, albeit greatly diminished through careful selection of vetted training data, an extent and likelihood of potential harm which the responses of the chatbot could cause if acted upon without human expert review and guidance that is greater than human benchmarks.[31] While one might argue that obvious medical and health related misinformation should be identifiable easily by applying common sense, we should remind ourselves that, in June 2020, the Centers for Disease Control and Prevention felt compelled to issue a public advisory report detailing instructions for safe household cleaning and disinfection for COVID-19 prevention explicitly warning the public to not drink bleach.[34] The bottom line remains: uncontained creation and spread of misinformation in healthcare and medical domains can lead to serious harm.

In defence of OpenAI, it never advertised ChatGPT as trustworthy advisor but rather as a crowdsourced technology evaluation and refinement experiment. But context matters, and ethical use and development of AI technology requires to responsibly assess and mitigate risks upon releasing it. This has not happened in the case of ChatGPT,[35] and, for that matter and in different ways, also not with several comparably popular applications of generative AI bots for automatically creating imagery, audio data, and code.[15] In the wake of an unprecedented hype that likens recent progress in generative AI to the introduction of the internet itself cautious voices are getting louder. Early warners of uncontained and ethically unguided growth of LLMs[13] experienced backlash to ethical whistleblowing.[14] Other notable authoritative voices at the time of writing include DeepMind's CEO and cofounder Demis Hassabis offering a measured take on the AI revolution that is worthwhile reading,[36] AI expert Gary Marcus providing a candid perspective on the risks and limitations of generative AI,[37] Insilico's CEO Alex Zhavoronkov pointing out the accuracy limitations of generative AI in biomedicine,[38] and Kevin Roose openly exposing the naïve, bordering dangerous mindsets of some of the creators of generative AI by asking them point-blank whether they 'were worried by unleashing generative AI on the world before it was safe' and reporting that the answer he got was a concerning 'no'.[15]

## How to bring generative AI to fruition in healthcare and medicine: an ethical, technical, and cultural call to action

With the Genie out of the bottle, what needs to be done to steer LLM-powered generative AI technology towards becoming a useful and safe tool in the healthcare, medical, and clinical fields? Certainly, hectic brute-force scaling the same deep learning technology with even more training data and more complex neural network models will not get us an inch closer to this goal or to a higher level of artificial intelligence, let alone to artificial general intelligence (AGI) at that. The last thing the field of (generative) Health AI needs at this point is another haphazard release of GPT-4 or any other purely scaled up LLM. The limitations of LLMs are systemic: simply boosting training data sizes and the number of model parameters to create future versions of the same model architectures will not overcome their shortcomings but rather amplify them. Instead, the following **ethical, technical, and cultural approaches**, should be the focus of responsible (re-)design of generative AI applications using LLMs so that they may become useful and safe tools for clinician and patient users.

### AI ethics

The field of AI ethics is built on several key principles for the responsible design and application of AI technology. In a seminal 2021 report The World Health Organisation has translated these principles into guidance for the ethical use and governance of AI in health.[39] This WHO framework allows to identify the specific ethical risks that arise from incorporating LLMs into generative AI systems in healthcare and medicine. The following section discusses insights from this assessment, explains key risk factors, and suggests possible pathways for risk mitigation. Table 1 summarises the findings.

(1) **Accountability:** Frameworks for ethical release and use of generative AI applications are needed. This starts with broadly educating users on the capabilities and risks of the technology and on the responsibilities and sensitivities involved in using it and should lead to the implementation of broadly understood and accepted conduct of use codices. At this point, the field navigates through a legally blurry state where good actors try to do the right thing, but bad actors seem to get away with following an 'ask for forgiveness, not for permission' paradigm, and the nature and capabilities of LLMs are misunderstood by many of its users. Some communities, such as for example, the education and academic publishing ones, have responded by issuing bans of the technology in their sectors.[40–42] Trying to pull the emergency break is an understandable first reaction in the face of a little understood threat. But a ban cannot and should not be a solution to dealing with LLMs. Several lawsuits have been launched and are currently under way which promise to bring some legal clarity to accountability, rights and responsibilities of users and developers.[43,44] The AI

**Framework for the ethical use, design, and governance of Large Language Model (LLM)-based generative AI in health and medicine**[a]

| AI ethics principles | Human autonomy | Human well-being, safety | Transparency, explainability | Responsibility, accountability | Inclusiveness, equity | Responsiveness, sustainability |
|---|---|---|---|---|---|---|
| Definition for health and medicine applications as per WHO guidance[39] | • Humans remain in control of health-care systems and medical decisions<br>• Providers have information necessary to make safe, effective use of AI systems<br>• People understand the role AI systems play in their care<br>• Data privacy and confidentiality is protected through valid informed consent | • Designers of AI satisfy regulatory requirements for safety, accuracy and efficacy for well-defined use cases or indications<br>• Measures of quality control in practice and quality improvement in the use of AI over time are available<br>• Use of AI does not result in mental or physical harm that could be avoided by use of an alternative practice or approach | • AI is understandable to developers, medical professionals, patients, users and regulators<br>• Sufficient information is published or documented before the design or deployment of AI, and such information facilitates meaningful public consultation and debate on how the AI is designed and how it should or should not be used<br>• AI is explainable according to the capacity of those to whom it is explained | • Clear, transparent specification of the tasks that the AI can perform and the conditions under which it can achieve the desired performance are available<br>• Human stakeholders ensure that the AI can perform those tasks and that it is used under appropriate conditions and by appropriately trained people<br>• Patients and clinicians evaluate AI in development and deployment<br>• Regulatory principles are applied upstream and downstream of the algorithm by establishing points of human supervision<br>• If something goes wrong with an AI system, there is accountability<br>• Appropriate mechanisms are available for questioning and for redress for individuals and groups that are adversely affected by decisions based on AI | • AI is designed and shared to encourage the widest possible appropriate, equitable use and access, irrespective of age, sex, gender, income, race, ethnicity, sexual orientation, ability or other characteristics protected under human rights codes<br>• AI is available for use not only in contexts and for needs in high-income settings but also in lower/middle income countries<br>• AI does not encode biases to the disadvantage of identifiable groups<br>• AI minimizes inevitable disparities in power that arise between providers and patients, between policy-makers and people, and between companies and governments that create and deploy AI and those that use it<br>• AI is monitored and evaluated to identify disproportionate effects on specific groups of people | • Designers, developers and users continuously, systematically and transparently assess AI during use to determine whether it responds adequately and appropriately and according to communicated expectations and requirements<br>• AI is consistent with wider promotion of the sustainability of health systems, environments and workplaces<br>• AI is designed to minimize its environmental consequences and increase energy efficiency<br>• Governments and companies address anticipated disruptions in the workplace, including training for healthcare workers to adapt to the use of AI, and potential job losses due to use of automated systems |
| LLM-focused risk factors | • AI is used without human-in-the-loop oversight for checking correctness and appropriateness of AI-generated data<br>• Patient data and EHR data are used to develop AI without data owner consent<br>• Care models use AI without disclosing how and when | • AI is released broadly without regulatory oversight or user education<br>• AI produces misinformation or inappropriate content resulting in physical or mental harm to patients or clinicians<br>• AI is operated without offering quality or truthfulness control mechanisms in place | • AI is released without educating stakeholders about its limitations or by misrepresenting the AI's lack of language and content comprehension<br>• AI is designed and released without describing or disclosing training and testing data sources<br>• AI-generated content is not clearly recognisable and marked as such at all times | • AI is used as autonomous, 'omniscient' tool<br>• Training and validation data are not curated properly allowing the AI to ingest illegal or inappropriate content or misinformation<br>• AI is designed and deployed without sufficient clinician and patient input<br>• AI is used without regulatory oversight and without human control of quality, truthfulness or appropriateness of in-or output<br>• No accountability mechanisms are in place for holding human designers and providers of AI accountable in case of adverse outcomes resulting from AI use or release | • Datasets for training and validating AI are selected prioritising data accessibility and availability over bias mitigation, or data is procured without any or with insufficient human due diligence and oversight by subject matter experts in the field of use<br>• AI design and development are dominated by stakeholders with access to large compute-power resources limiting the diversity of developer and user community<br>• There is no assessment of potential bias in outcomes and impact of care decisions which are influenced by the use of AI | • The probabilistic nature of AI is neglected or overseen and no continuous audit of AI-generated content is performed after deployment and during use of AI<br>• Size of datasets and complexity of AI model architectures cause large carbon footprint for training at scale<br>• Human labour costs for data selection, procurement and engineering are underestimated or misrepresented leading to unfair or unhealthy working conditions |

(Table 1 continues on next page)

**Framework for the ethical use, design, and governance of Large Language Model (LLM)-based generative AI in health and medicine**[a]

| AI ethics principles | Human autonomy | Human well-being, safety | Transparency, explainability | Responsibility, accountability | Inclusiveness, equity | Responsiveness, sustainability |
|---|---|---|---|---|---|---|
| (Continued from previous page) | | | | | | |
| LLM-focused risk mitigation pathways | • As a matter of regulatory approval, design and deploy AI as an assistive tool, augmenting the capabilities of human decision makers and not replacing them<br>• As a matter of regulatory approval, design AI systems to regularly produce easily accessible and quantifiable performance, usage and impact metrics explaining when and how AI is used to assist decision making and allowing to detect potential bias, either present at deployment of AI or evolving during ongoing use of AI in operational environment<br>• Study the value system of the target user group for a specific AI application, and design and develop the AI to adhere to these values accordingly<br>• As a matter of regulatory approval and machine learning best practices, declare the purpose of designing and using a specific AI system at the outset of any conceptual or development work, including the selection and curation of training or validation data and the integration of the AI system into clinical or non-clinical workflows or care models<br>• As a matter of regulatory approval and as a requirement of best-practice product release, designers, developers and providers of AI will disclose sources of all training and validation data<br>• As a matter of regulatory approval, design AI systems to clearly and transparently label any AI-generated content as such<br>• As a matter of regulatory approval and machine learning best-practices, humans ongoingly audit AI systems including the data they ingest and generate against defined quality, safety, and performance standards and ensure adherence to responsible AI conduct of use codices, data privacy and security policies including appropriate and valid consent frameworks<br>• Maintain databases for collecting, documenting and sharing the results of AI audits and insights, educate users about capabilities, limitations and risks for specific AI use cases and applications, and improve AI performance and trustworthiness through periodically retraining and redeploying updated algorithms<br>• Apply fair-work and safe-work standards when employing human developers of AI along the MLOps Cycle with a special focus on rewarding human subject matter experts and on assessing and mitigating risks for physical and mental harm that might arise for data engineers and reviewers of AI operations and performance | | | | | Subject to a combination of existing regulatory frameworks, good machine learning practice, and regulatory and technical suggestions at the time of writing |
| | • Develop and apply methods for programming values into AI systems, and integrate frameworks for monitoring adherence of AI systems to encoded values<br>• Explore technologies for 'untraining' AI algorithms in order to help models 'forget' toxic or inappropriate content or misinformation they might have been exposed to<br>• Explore technologies for automatically detecting content generated by AI<br>• Explore technologies for embedding cognitive models-of-the-world in AI systems to guide them towards language comprehension and to support models with self-guided detection of harmful or inappropriate content<br>• Explore technologies for developing truthfulness indices for AI systems<br>• Explore avenues for building use case-specific evidence-based AI applications that rely on smaller training data and less complex model architectures supporting sustainable creation of AI systems by reducing the carbon footprint of model training and testing<br>• Establish legal precedence to define the circumstances under which data may be used for training AI, and as part of this establish, communicate and enforce copyright, liability and accountability frameworks that govern the connection between data used for training and developing AI systems, the data such AI systems then generate, and the impact of decisions humans make using such data | | | | | Subject to ongoing and future exploratory technical research and legal proceedings at the time of writing |
| Governance approaches | AI ethics committees with executive decision powers integrated into AI product design, development, deployment, and auditing processes<br>targeted education and collaborative agency of all stakeholders: developers, providers, medical professionals, patients, users, and regulators<br>secure sandbox environments for safe, consumer-centric research, development, prototyping, and validation of AI applications for clinical environments | | | | | |

The proposed framework is based on the AI ethics principles and their projection onto AI systems in health as defined by The World Health Organisation 2021 Guidance on Ethics and Governance of Artificial Intelligence for Health[39] (first and second row), employs these principles to identify key risk factors specific to using Large Language Models as part of generative AI applications in healthcare and medicine (third row), and suggests possible risk mitigation pathways (fourth row). Governance approaches which could support the effective implementation of this framework are provided in the bottom row. Note that this proposed framework does not override the full scope of criteria for assessing and addressing the risks of using AI technology in health as laid out in[39] but seeks to extract and highlight those aspects that carry particular relevance and urgency with respect to the role of LLMs and generative AI in health. The framework is intended to be a basis and trigger for discussion and further refinement towards its broader uptake and integration into best-practices processes aiming to translate innovation in LLM-driven generative AI into products that ethically deliver value and minimise risks for those who use them. [a]Within this table the term 'AI' is used to describe Large Language Models (LLMs) as part of generative AI systems.

***Table 1:*** **Ethics framework for the responsible use, design, and governance of Large Language Models and generative AI applications in health and medicine.**

ethics community has begun to home in on generative AI[45] and on regulating it through a checks and balances system as a matter of urgency.[46] A rift seems to deepen between smaller start- and scaleup creators of generative AI willing to take higher risks when releasing it and BigTech corporations who–bound by high levels of public, regulatory and reputational scrutiny–double down on and promote their responsible and trustworthy AI programs in line with releasing LLM applications.[3]

(2) **Fairness:** Measures to mitigate model biases need to be put in place. In an ethically responsible AI development environment, this task is an essential part of the entire MLOps cycle. Bias can creep into AI applications in many ways and at several development stages. The most blatant neglect of addressing bias has already happened for many generative AI applications currently in public deployment by either scraping training data of the internet unseen or by curating training data using questionable methods[35] or with insufficient attention to sources for potential bias or misinformation. Med-PaLM is one noteworthy laudable exception in this respect, but ChatGPT and Stable Diffusion are not. For them fighting bias has become as much of a retrospective as a prospective issue: human ethics panels with deciding power need to review current model implementations to identify and eliminate sources of bias and misinformation. This is a hard task when dealing with models that have already been trained using biased or incorrect data. The question of how to make an AI system 'unlearn' problematic content is a complex research topic.[47] Furthermore, such ethics panels need to continuously audit the performance of already deployed models, report and weed out any problematic content they produce. This requires model developers and providers to work with human experts in the user community on identifying and documenting any errors or biased outputs which the deployed models might produce. Such information could be used for fact-checking and retraining models as well as for educating the user community on limits and risks of using the models. Note that constantly auditing LLM outputs is needed because, due to their probabilistic nature, it cannot simply be assumed that an LLM that has previously produced a correct or appropriate output to a prompt will automatically do so again when repeatedly given the same prompt. A large-scale citizen-science project recently kicked off by Gary Marcus and Ernest Davis aims to establish an error database for ChatGPT,[48] and pressure needs to be put on developers and providers of models to engage in such activities proactively and continuously.

(3) **Data privacy and selection:** Legal and ethical frameworks for selecting and managing training data are needed. Heated debates and, as of a few weeks ago also the first high-profile lawsuit[49] have ensued in the software engineering, creative, publishing, and arts sectors around copyright issues related to works being created by LLMs which had been trained on human artists' work without obtaining their explicit consent to approve that use of their data. In the healthcare and medical sector, debates on training data and AI generated materials will not be confined to data ownership and copyright issues but touch on liability and consent issues as well. Electronic health records range among the most highly restricted and sensitive data sources and require to be treated as such, and creating potentially harmful health advice has further reaching implications than infringing an artist's copyrights. Furthermore, as explained earlier, in evidence-driven domains such as healthcare and medicine the act of choosing and accessing suitable training data to develop generative AI applications has not only legal and ethical but also model performance related implications. At the time of writing, the legal lay of the land for managing data ingested into and produced by LLMs is complex and awaits the creation of clarifying legal precedence. James Vincent has published a comprehensive overview of the issues at hand and the powers at work.[50]

(4) **Transparency:** LLMs bridge prompts with responses without being inherently capable of showing the logic behind their work. This task is left to the human operator who will only be able to do a reliable job if the model provides insights into its data sources and labels AI-generated content as such. Some of the core reasons for Galactica's swift downfall was the fact that it made up scientific references and cited those in its responses.[51] A non-excusable cardinal sin in the world of academia and research. One of the greatest threats caused by undetectable generative AI content is that if malicious, it could contaminate knowledgebases at scale. Turbocharged by the power of automated chatbots, camouflaged generative AI content could become a dangerous source of misinformation permanently leaving its toxic mark on knowledgebases.[52] One can only imagine how misinformation broadly disseminated and hiding under the cloak of scientific truth could influence debates on such hotly contented topics as for example vaccination rules. OpenAI is reportedly planning to introduce a 'Watermark' feature labelling content created by ChatGPT as such. The efficiency of this effort is in doubt.[53] DetectGPT is an experimental AI tool currently under development at Stanford University that allows to automatically detect whether a text sample has been created by a human or by a chatbot using

an LLM. The program was reportedly able to correctly determine authorship for 95% of test cases across five popular LLMs, but more work is needed to bolster DetectGPT's resistance to various evasion strategies.[54] Other, rather exotic but nevertheless potentially valid approaches to marking data with so-called 'radioactive' samples which would allow to retrospectively detect its traces in LLM outputs that have used such contaminated data as training data are being explored by reputable developers.[55] Hugging Face and project BLOOM follow a radically transparent approach to developing generative AI technology.[56] Integrating humans in the generative AI MLOps cycle brings its own ethical challenges: the detection of harmful content can expose human data annotators to mental health risks, and the immense amount of data that has to be reviewed at scale could lead to exploitative working conditions akin to those seen at assembly lines in the early stages of the industrial revolution.[35] In evidence-based sectors such as healthcare and medicine data need to be curated by highly skilled clinical domain experts which results in often-underestimated invisible clinical labour cost. As generative AI systems will become more prolific, the workload and accountability aspect of data curation by humans will have to be addressed to avoid bottle necks and to keep human experts incentivised to partake in data selection and review.[57]

(5) **Explainability:** Closely tied to transparency and a complex area of research within the field of AI,[58] a maximum level of explainability should be a key design feature of generative AI systems as it provides an important datapoint to the human user and operator whose role it is to validate the correctness and soundness of the AI's work. Transparent and explainable generative AI systems could be assigned a truthfulness index which would serve as an important measure for assessing the trustworthiness of an LLM when assisting clinicians and patients. The awareness for the importance of such indices however has just emerged in the field of AI in 2022, and at the time of writing only one index called TruthfulQA exists.[59] Research into truthfulness indices for generative AI solutions should be at the core of every LLM development program in healthcare and medicine. Having said this, numerous programs to build, test, and apply frameworks and tools for developing trustworthy AI systems are established in the field. For example, IBM's 'AI 360 Toolkit' offers industry-grade tools to assess and improve fairness, explainability and robustness of AI systems.[60] Google has introduced similar tools as part of its AI Principles program: the Model Remediation Library, Learning Interpretability Tool and Know Your Data Catalogue allow to identify bias, to improve fairness of models, to backtrack errors from model outputs to training data, and to scrutinise data with respect to its origin content and labels.[61] Microsoft's One Engineering System develops tools with similar capabilities.[62] Such frameworks are applicable to LLM technology, and it will be imperative to integrate them in the MLOps cycle for developing trustworthy generative AI applications in healthcare and medicine.

Explainability is not the sole yardstick to determine the clinical usefulness of an insight which an AI system has produced: while a high level of explainability is desirable, it is possible to use agnostic learning on high-dimensional data to produce clinically useful conclusions without being able to fully explain how the underlying AI system reached them.[63] Such cases illustrate the paramount need for human expert validation and review of AI generated insights before they are translated into actions. But they also highlight the fact that AI technology can assist humans with selecting endpoints in AI-driven clinical interventions which they might not have considered without enlisting the help of AI.

(6) **Value and purpose alignment:** Generative AI shines a special light on AI's so-called Alignment Problem[64]: it describes the ethical and existential risks that emerge when machines do not follow or violate the values and purpose of their human creators and users. The constellation from which an alignment problem can arise requires a human value system to be in place on one hand, and a clearly defined purpose of what an AI system is supposed to be doing and why on the other hand. To then solve an alignment problem requires the ability to program values into an AI system and to control its adherence to them. All these aspects need to be at the forefront of developing LLM-driven generative AI applications in healthcare and medicine. Releasing them into the wild without declaring the values they are intended to follow, without defining a purpose as to what they should accomplish, and without explaining how these are implemented in the AI system and how model users and creators can monitor and enforce that the system adheres to them, AI technology is naïve at best and dangerous at worst. Value and purpose frameworks will depend on application sectors, users, and developer communities but they should always be declared and implemented in any AI system as a first step of responsible AI development. The medical community has a strong set of values as set out in the Hippocratic Oath, and the 'do no harm' imperative should be at the center of all exploration and application of generative AI for assisting in the management of health and wellness. OpenAI, in an effort to counteract toxic use of GPT-3, has made

first steps to address the alignment issue by introducing a custom-designed LLM called InstructGPT which is reportedly better in capturing the intentions of human users and taking them into account when producing an output.[65]

## AI technology

Besides these ethical design features, there are also conceptional opportunities for technically altering the architecture of generative AI systems using LLMs to put them on the path to language comprehension and suitability for deployment in healthcare and medicine applications. The most fundamental and challenging one is to develop and integrate so-called 'models of the world'. A contested and debated area of research at the intersection of AI and cognitive neuroscience,[66,67] there are several approaches to how such models of the world could be designed and operated. Ultimately, humans develop, refine, and continuously apply them to understand and manipulate the world: they enable us to plan, reason, learn and have common sense. In the world of AI research, these cognitive abilities are currently–and have been for a while–overshadowed by the dominance of one of them: (deep) learning. To develop artificial, truly intelligent systems the field needs to not exclude but go beyond learning and take a more nuanced approach to researching and developing all aspects of intelligence. Generative AI technology and LLMs constitute both, a showcase of how turbocharging one narrow aspect of AI can yield an extremely promising and powerful tool but ultimately dumb system, and a call to action and unique opportunity for mitigating this outcome by diversifying the field of AI research. Analytical methods from the field of cognitive psychology have been used to shed light on the working principles of LLMs, pointing out their cognitive limitations but possibly also opening avenues for linking cognitive models of the world into them.[68] Meta has recently taken a first exciting step into this direction in form of their CICERO chatbot.[69] Capable of mastering the complex strategy game Diplomacy, the underlying LLM architecture veers away from the pure scaling-up trajectory and instead reflects attempts to incorporate cognitive capabilities beyond learning, notably bringing reasoning and planning into the game (figuratively and literally). A promising differentiation from the pure scaling strategy[70] which still seems to be the mainstream trajectory of LLM research and evolution judging by the buzz pre-empting an expected impending release of OpenAI's next-generation, even higher-parameter LLM GPT-4.

Another technical approach to changing the nature of LLM development is to explore ways for improving model usability and performance for evidence-based knowledge domains such as healthcare and medicine by designing smaller models and training them on smaller data rather than throwing ever more data at ever more complex models. Google's Med-PaLM in contrast to OpenAI's ChatGPT is a good example taking this alternative pathway. Other examples include Microsoft's BioBERT,[71] PubMedBERT,[72] and BioGPT.[8] We have seen this development in the field of deep learning before where a drive away from the infamous 'big data' paradigm and the introduction of transfer learning and data engineering has produced a generation of smaller yet impactful AI models.[73] ChatGPT uses a downsized version of the earlier GPT-3 LLM, but this is more of a proxy effect as a larger initial neural network trained on big data was needed before smaller offspring neural networks could be created from it. Especially for applications of generative AI tools in healthcare and medicine, developers will run into the 'big data' problem, i.e., will have to face the fact that while theoretically electronic health records are a multimodal bottomless ocean of big data poised to be analysed using deep learning models,[74] when it comes to getting access to such data, they very much are not. Whereas it was at least irresponsible in the case of non-medical LLMs to scrape training data off the internet sight unseen, it might be illegal or outright impossible to do this for medical LLMs in the healthcare domain.

Healthcare does move slowly but inevitably from a provider-centric monopolised to a patient-centric democratized state where patients gain more and more power over what happens with their data.[75,76] While this works in favour of developing LLMs in healthcare and medicine, health data accessibility, quality and maturity on the ground differs vastly between different geographies, jurisdictions and subdomains which will require a strategy for generative health AI developers to work with lesser data. Another welcome side effect of switching to smaller models and datasets would be an increased diversity of the LLM developer community: at the moment, the scale-up strategy for crunching out ever more complex LLMs requires massive compute power at a level that only big tech enterprises or start-ups backed by them can sustainably afford. Moving to less compute power hungry model architectures and training regimes would not only reduce the carbon footprint of training such models,[39] but make it sustainable for smaller entities, and desirably so also for medical or clinical institutions to participate meaningfully in the responsible and evidence-based creation of LLMs. This would boost the diversity of the generative health AI developer community, thus improving its ability to counteract bias and misinformation[13,31] and helping to balance an increasingly commercialisation-heavy approach to bringing LLM-driven services to market with ethically and ecologically oriented ones.[77]

## A culture of collaboration should drive future strategy

Besides ethical and technical approaches to transforming LLM powered chatbots into useful and

trustworthy tools for human users in healthcare and medicine, there is also a cultural aspect to be considered. The often-heralded Silicon Valley paradigm of "move fast and break things" does not apply to healthcare and medicine. This makes Health AI a fault line of innovation where different cultures of conceiving, developing, and deploying new technologies tend to clash. Successfully overcoming trust and acceptance hurdles for innovation transfer requires close collaboration and communication between all involved stakeholder groups: clinician and patient users, technology developers and regulators. To make generative AI systems and LLMs useful decision support and knowledge retrieval services for clinicians and patients means taking the time to collectively assess the opportunity and risk landscape and to develop research, trialling, implementation, and auditing standards following the technical and ethical themes laid out above (Table 1). In the past and more than once, I have had a front row seat at the intersection of industrial AI research and clinical healthcare sectors that allowed me to witness the rejection spiral which high-potential AI technology can be drawn into when not developed and deployed in lockstep with target clinical and patient adopter communities. There often are no second opportunities to get things right after releasing AI technology prematurely or hastily in the healthtech sector: user and regulator trust are easy to lose and very hard to regain.

Translating this knowledge into a sustainable and realistic strategy and timeline for bringing LLM-based generative AI technology to fruition in healthcare points to a two-stage agenda: in a first phase, covering the immediate and mid-term future, focusing on the ethical approaches laid out above will allow to establish and improve the safety, trustworthiness, and usefulness of already released models for selected Health AI applications and for those currently under development. We will see more nuanced and risk-conscious experimentation with research-grade generative AI systems accompanied by increased scrutiny of regulatory bodies, and first commercial product offerings that are targeted and regulated for very specific niche applications in health data management such as summarising or creating reports. In a second phase, reaching into the long-term future, exploring technical model-of-the-world as well as scale-down instead of scale-up approaches might create generative AI models which understand the data they handle, be it text, audio, or imagery. It is at this point that LLMs might become intelligent and trustworthy artificial companions to clinicians and patients opening a wide spectrum of possible applications in managing health and wellness from Software as a Medical Device systems in the clinical diagnostics and prognostics space to digital personal health coaches in consumer health and non-clinical sectors. I am extremely excited about this moon-shot goal and inspired by thinking about the transformative role generative AI and LLMs could one day play in healthcare and medicine. But I am also acutely aware that we are by no means there yet and that, despite the prevailing hype, LLM-powered generative AI may only gain the trust and endorsement of clinicians and patients if the research and development community aims for equal levels of ethical and technical integrity as it progresses this transformative technology to market maturity.

#### Contributors

SH is responsible for all parts of the presented study including conceptualisation, investigation, methodology, project administration, resources, validation, visualisation, editing, writing, reviewing, and revising the original draft as well as the published manuscript.

#### Declaration of interests

SH is an inventor on granted US Patent 11,250,219 B2 'Cognitive Natural Language Generation with Style Model'.

#### Acknowledgements

SH conducted this study as an employee of the Digital Health Cooperative Research Centre (DHCRC) and accepts responsibility to submit for publication.

#### Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2023.104512.

#### References

1 OpenAI. Introducing ChatGPT. sourced from: https://openai.com/blog/chatgpt; 2023. https://dpo.org/10.48550/arXiv.2203.02155.
2 Thoppilan R, De Freitas D, Hall J, et al. Lamda: language models for dialog applications. *arXiv*. 2022. https://doi.org/10.48550/arXiv.2201.08239. preprint arXiv:2201.08239.
3 Pichai S. *An important next step on our AI journey*. Google Blog, The Keyword; 2023. sourced from: https://blog.google/technology/ai/bard-google-ai-search-updates/.
4 stability.ai. Stable diffusion v2.1 and DreamStudio updates. sourced from: https://stability.ai/blog/stablediffusion2-1-release7-dec-2022; 2022.
5 OpenAI. Dall·E 2. sourced from: https://openai.com/dall-e-2/; 2023.
6 Singhal K, Azizi S, Tu T, et al. Large Language models encode clinical knowledge. *arXiv*. 2022. https://doi.org/10.48550/arXiv.2212.13138. preprint arXiv:2212.13138.
7 Kung TH, Cheatham M, Medinilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using Large Language Models. *medRxiv*. 2022. https://doi.org/10.1371/journal.pdig.0000198.
8 Luo R, Sun L, Xia Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform*. 2022;23(6):bbac409.
9 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30:1–11.
10 Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv*. 2018. https://doi.org/10.48550/arXiv.1810.04805. preprint arXiv:1810.04805.
11 Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877–1901.
12 Scao TL, Fan A, Akiki C, et al. Bloom: a 176b-parameter open-access multilingual language model. *arXiv*. 2022. https://doi.org/10.48550/arXiv.2211.05100. preprint arXiv:2211.05100.
13 Bender EM, 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*. 2021:610–623.

14 Lyons K. *Timnit Gebru's actual paper may explain why Google ejected her. The Verge*; 2020. sourced from: https://www.theverge.com/2020/12/5/22155985/paper-timnit-gebru-fired-google-large-language-models-search-ai.
15 Roose K. *A coming-out party for generative AI, Silicon Valley's new craze. The New York Times*; 2022. sourced from: https://www.nytimes.com/2022/10/21/technology/generative-ai.html.
16 *"On bullshit". The importance of what we care about: philosophical essays.* Cambridge: Cambridge University Press; 1988:117–133. https://doi.org/10.1017/CBO9780511818172.011.
17 Clynch N, Kellett J. Medical documentation: part of the solution, or part of the problem? A narrative review of the literature on the time spent on and value of medical documentation. *Int J Med Inform.* 2015;84(4):221–228.
18 Henry T. *Do you spend more time on administrative tasks than your peers.* Chicago, IL: American Medical Association; 2018.
19 Shen Y, Heacock L, Elias J, et al. ChatGPT and other Large Language Models are double-edged swords. *Radiology.* 2023:230163.
20 Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *arXiv.* 2022. https://doi.org/10.48550/arXiv.2212.14882. preprint arXiv:2212.14882.
21 Rushabh HD, Simar SB. Promises – and pitfalls – of ChatGPT-assisted medicine. sourced from: https://www.statnews.com/2023/02/01/promises-pitfalls-chatgpt-assisted-medicine/; 2023.
22 Harrer S, Shah P, Antony B, Hu J. Artificial intelligence for clinical trial design. *Trends Pharmacol Sci.* 2019;40(8):577–591.
23 Gordon R. *Large language models help decipher clinical notes.* MIT News, Massachusetts Institute of Technology; 2022. sourced from: https://news.mit.edu/2022/large-language-models-help-decipher-clinical-notes-1201.
24 Callaway E. Scientists are using AI to dream up revolutionary new proteins. *Nature.* 2022;609:661. sourced from: https://www.nature.com/articles/d41586-022-02947-7.
25 Philippidis A. *Insilico gains FDA's first orphan drug designation for AI candidate.* Genetic Engineering & Biotechnology News; 2023. sourced from: https://www.genengnews.com/artificial-intelligence/insilico-gains-fdas-first-orphan-drug-designation-for-ai-candidate/.
26 Callaway E. 'The entire protein universe': AI predicts shape of nearly every known protein. *Nature.* 2022;608(7921):15. sourced from: https://www.nature.com/articles/d41586-022-02083-2.
27 Crawford E. *AlphaFold works with other AI tools to go from target to hit moleculer in 30 days.* ChemistryWorld; 2023. sourced from: https://www.chemistryworld.com/news/alphafold-works-with-other-ai-tools-to-go-from-target-to-hit-molecule-in-30-days/4016935.article.
28 Huang S, Grady P, GPT-3. *Generative AI: a creative new world.* Sequoia Capital; 2022. sourced from: Generative AI: A Creative New World | Sequoia Capital US/Europe.
29 Kovanovic V. *The dawn of AI has come, and its implications for education couldn't be more significant.* Freethink; 2023. sourced from: https://www.freethink.com/society/chatgpt-education.
30 Lardinois F. *Microsoft launches the new BING, with ChatGPT built in.* TechCrunch; 2023. sourced from: https://techcrunch.com/2023/02/07/microsoft-launches-the-new-bing-with-chatgpt-built-in/.
31 Topol E. *When M.D. Is a machine doctor.* GroundTruths; 2023. sourced from: https://erictopol.substack.com/p/when-md-is-a-machine-doctor.
32 Broderick R. *The wild world of PromptBase, the ebay for generative AI prompts.* FastCompany; 2022. sourced from: https://www.fastcompany.com/90825418/promptbase-generative-ai-prompt-marketplace.
33 Greene T. *Meta takes new AI system offline because Twitter users are mean.* TheNextWeb; 2022. sourced from: https://thenextweb.com/news/meta-takes-new-ai-system-offline-because-twitter-users-mean.
34 Gharpure R, Hunter CM, Schnall AH, et al. Knowledge and practices regarding safe household cleaning and disinfection for COVID-19 prevention – United States, May 2020. *MMWR Morb Mortal Wkly Rep.* 2020;69(23):705–709. US Department of Health and Human Cervices / Centers for Disease Control and Prevention.
35 Perrigo B. *OpenAI used Kenyan workers on less than $2 per hour.* Time Magazine; 2023. sourced from: https://time.com/6247678/openai-chatgpt-kenya-workers/.
36 Perrigo B. *DeepMind's CEO helped take AI mainstream. Now he's urging caution.* Time Magazine; 2023. sourced from: https://time.com/6246119/demis-hassabis-deepmind-interview/.
37 The Ezra Klein Show. *A sceptical take on the AI revolution.* The New York Times; 2023. sourced from: https://www.nytimes.com/2023/01/06/opinion/ezra-klein-podcast-gary-marcus.html.
38 Zhavoronkov A. Caution with AI-generated content in biomedicine. *Nature*; 2023. https://doi.org/10.1038/d41591-023-00014-w. sourced from: https://www.nature.com/articles/d41591-023-00014-w.
39 *Ethics and governance of artificial intelligence for health: WHO guidance.* Geneva: World Health Organization; 2021.
40 Lipman J, Distler R. *Schools shouldn't ban access to ChatGPT.* Time Magazine; 2023. sourced from: https://time.com/6246574/schools-shouldnt-ban-access-to-chatgpt/.
41 Vincent J. *Top AI conference bans use of ChatGPT and AI language tools to write academic papers.* The Verge; 2023. sourced from: https://www.theverge.com/2023/1/5/23540291/chatgpt-ai-writing-tool-banned-writing-academic-icml-paper.
42 Vishwam S. *Scientific journals ban ChatGPT use by researchers to author studies.* The Independent; 2023. sourced from: https://www.independent.co.uk/tech/chatgpt-ai-journals-ban-author-b2270334.html?utm_campaign=fullarticle&utm_medium=referral&utm_source=inshorts.
43 Vincent J. *AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit.* The Verge; 2023. sourced from: https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart.
44 Metz C. *Lawsuit takes aim at the way AI is built.* The New York Times; 2022. sourced from: https://www.nytimes.com/2022/11/23/technology/copilot-microsoft-ai-lawsuit.html.
45 Coalition for Health AI. *Blueprint for trustworthy AI implementation guidance and assurance for healthcare.* CHAI Consultation Paper; 2022. sourced from: https://coalitionforhealthai.org/papers/Blueprint%20for%20Trustworthy%20AI%20IG%20&%20Assurance%20for%20Health.pdf.
46 Larsen B, Narayan J. *Generative AI, a game-changer that society and industry need to be ready for. The World Economic Forum*; 2023. sourced from: https://www.weforum.org/agenda/2023/01/davos23-generative-ai-a-game-changer-industries-and-society-code-developers/.
47 *We forgot to give neural networks the ability to forget.* Forbes; 2023. sourced from: https://www.forbes.com/sites/ashoka/2023/01/25/we-forgot-to-give-neural-networks-the-ability-to-forget/?sh=900b47f6853f.
48 Marcus G, Davis E. *Large language models like ChatGPT say the darnedest things – ChatGPT/LLM errors tracker.* The Road to AI We Can Trust; 2023. sourced from: https://garymarcus.substack.com/p/large-language-models-like-chatgpt.
49 Goldman S. *Stable diffusion AI art lawsuit, plus caution from OpenAI, DeepMind | the AI beat.* VentureBeat; 2023. sourced from: https://venturebeat.com/ai/stable-diffusion-lawsuit-plus-words-of-caution-from-openai-deepmind-the-ai-beat/.
50 Vincent J. *The scary truth about AI copyright is nobody knows what will happen next.* The Verge; 2022. sourced from: https://www.theverge.com/23444685/generative-ai-copyright-infringement-legal-fair-use-training-data.
51 Snoswell AJ, Burgess J. *A galaxy of deep science fakes: the problems with Galactica AI.* siliconrepublic; 2022. sourced from: https://www.siliconrepublic.com/machines/galactica-ai-meta-fake-science-misinformation.
52 Heikkilä M. *How AI-generated text is poisoning the internet.* MIT Technology Review; 2022. sourced from: https://www.technologyreview.com/2022/12/20/1065667/how-ai-generated-text-is-poisoning-the-internet/.
53 Wiggers K. *OpenAI's attempts to watermark AI text hit limits.* TechCrunch; 2022. sourced from: https://techcrunch.com/2022/12/10/openais-attempts-to-watermark-ai-text-hit-limits/.
54 Miller K. *Human writer or AI? Scholars build a detection tool.* Stanford University Human-Centered *AI (HAI)*; 2023. sourced from: https://hai.stanford.edu/news/human-writer-or-ai-scholars-build-detection-tool?utm_source=linkedin&utm_medium=social&utm_content=UComm_linkedin_Stanford-University_202302151523_sf175174497&utm_campaign=&sf175174497=1.
55 Newton C. *Can 'radioactive data' save the internet from AI's influence?* Platformer; 2023. sourced from: https://www.platformer.news/p/can-radioactive-data-save-the-internet.
56 Heikkilä M. *Inside a radical new project to democratize AI.* MIT Technology Review; 2022. sourced from: https://www.technologyreview.com/2022/07/12/1055817/inside-a-radical-new-project-to-democratize-ai/?utm_campaign=site_visitor.unpaid.engagement&utm_source=LinkedIn&utm_medium=tr_social.
57 Ulloa M, Rothrock B, Ahmad FS, Jacobs M. Invisible clinical labor driving the successful integration of AI in healthcare. *Front Comput Sci.* 2022;4:157.

58 Blackman R, Ammanath B. *When – and why – you should explain how your AI works*. Harvard Business Review; 2022. sourced from: https://hbr.org/2022/08/when-and-why-you-should-explain-how-your-ai-works?ab=hero-main-text.
59 Lin S, Hilton J, Evans O. TruthfulQA: measuring how models mimic human falsehoods. *arXiv*. 2021. https://doi.org/10.48550/arXiv.2109.07958. preprint arXiv:2109.07958.
60 Hosurmath M, Bhojwani P, Kumar S, et al. *The AI 360 toolkit: AI models explained*. IBM Developer Blog; 2021. sourced from: https://developer.ibm.com/articles/the-ai-360-toolkit-ai-models-explained/.
61 Croak M, Gennai J. *Responsible AI: looking back at 2022, and to the future*. Google Blog The Keyword; 2023. sourced from: https://blog.google/technology/ai/responsible-ai-looking-back-at-2022-and-to-the-future/.
62 Microsoft AI. Responsible AI resources. sourced from: https://www.microsoft.com/en-us/ai/responsible-ai-resources; 2023.
63 Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med*. 2018;24(11):1716–1720.
64 Christian B. *The Alignment Problem: machine learning and human values*. W. W. Norton & Company; 2020. ISBN 0393635821.
65 Dominguez D. *OpenAI introduces InstructGPT language model to follow human instructions*. InfoQ; 2022. sourced from: https://www.infoq.com/news/2022/02/openai-instructgpt/.
66 Lohr S. *One man's dream of fusing AI with common sense*. The New York Times; 2022. sourced from: One Man's Dream of Fusing A.I. With Common Sense - The New York Times (nytimes.com).
67 Tiernan R. *Meta's AI guru LeCun: most of today's AI approaches will never lead to true intelligence*. ZDNet; 2022. sourced from: Meta's AI guru LeCun: Most of today's AI approaches will never lead to true intelligence | ZDNET.
68 Binz M, Schulz E. Using cognitive psychology to understand GPT-3. *Proc Natl Acad Sci U S A*. 2023;120(6):e2218523120.
69 Verma P. *Meta's new AI is skilled at a ruthless, power-seeking game*. The Washington Post; 2022. sourced from: https://www.washingtonpost.com/technology/2022/12/01/meta-diplomacy-ai-cicero/.
70 Sevilla J, Heim L, Ho A, Besiroglu T, Hobbhahn M, Villalobos P. Compute trends across three eras of machine learning. *arXiv*. 2022. preprint arXiv:2202.05924. In 2022 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.
71 Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234–1240.
72 Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthc*. 2021;3(1):1–23.
73 Strickland E. *Andrew Ng: unbiggen AI*. *IEEE Spectrum*; 2022. sourced from: https://spectrum.ieee.org/andrew-ng-data-centric-ai.
74 Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. *Nat Med*. 2022;28(9):1773–1784.
75 Topol E. *The patient will see you now*. Little Brown Publishing; 2016. ISBN 9780465040025.
76 Harrer S. *Commercialising digital health: trading on a dynamic data marketplace*. *Forbes*; 2021. sourced from: https://www.forbes.com/sites/forbestechcouncil/2021/06/01/commercializing-digital-health-trading-on-a-dynamic-data-marketplace/?sh=a938ee76f5cb.
77 Tiku N, De Vynck G, Oremus W. *Big Tech was moving cautiously on AI. Then came ChatGPT*. The Washington Post; 2023. sourced from: https://www.washingtonpost.com/technology/2023/01/27/chatgpt-google-meta/.