AI/Data Science Professional

COSC2778/COSC2792/COSC2818

Assessment Task 2: Project Proposal for Data-Driven Solution

Report

Improving Transparency and

Interoperability of Healthcare AI Systems

Thu 4:30- Group 1

Mrwan Fahad A Alhandi	**s3969393@student.rmit.edu.au**
Eric Cheung	**s3868588@student.rmit.edu.au**
Luke Dale	**s3964888@student.rmit.edu.au**
Richard Doherty	**s3863706@student.rmit.edu.au**

Introduction

Artificial Intelligence (AI) has been used in health care sector for years. Its aim is to help making better clinical decisions and user experience, in areas such as improving diagnostics through using machine learning techniques, medicine personalization, reducing errors and cost savings through automation.

In this proposal, we'll discuss the usage, current problems, pain points and challenges of AI and Machine Learning (ML) methods in the healthcare sector. Furthermore, we'll propose a data-driven solution, together with methodology and a design prototype.

Relevant stakeholders

For general AI usage in healthcare, relevant stakeholders can include a diverse range of individuals, such as:

Health professionals: Clinicians use Deep Neural Networks (DNN) to assist in tasks such as interpreting medical scans, and the results suggested that AI had a better accuracy than human practitioners. Another area which AI has been used is mental health, tools are developed include digital tracking of mental health issues via various methods, for example speech, voice and facial recognition. ML techniques have also been developed for predicting successful antidepressant medication, predicting suicide and other mental health issues. (Topol, (2019))

Technology companies and researchers: Companies which are developing the software, algorithms, hardware for AI usage, for example, companies which developed the algorithms (DNN) used in interpreting medical scans, companies that manufacture the devices to carry out the procedures of taking scans, storage of scans. Companies provide security around the process from capturing to storing scans and patients' records.

Patients and Advocacy group: They are the group who are included users being treated and benefit by the AI services & research, also the users who participate in the development of the service in providing feedback to developers. Advocacy group comes not only from patients, but also from medical practitioners such as the American Medical Association, in February 2019 edition of the AMA Journal of Ethics, it mentioned several articles about Ethics and usage of AI.

Regulators: Governments and regulatory bodies are the stakeholders. They are the bodies which set policies in areas such as privacy, security, transparency, bias, and accountability. They also regulate the use of AI. For example, in the National Statement on Ethical Conduct in Human Research (2007), it covered the various aspects of ethical conduct in human research, such as obtaining informed consent from participants, protecting the privacy and confidentiality of participants, minimising harm and maximising benefits to participants, and ensuring that the research is conducted with integrity and transparency. (National Statement on Ethical Conduct in Human Research (2007))

Insurance companies: Insurance companies have been using AI for health care in many areas such as Chatbot for customers interaction, fraud detection, personalized insurance policy. The aim is to increase accessibility for the public, reveal any early illness etc. (Role of Artificial Intelligence (AI) in Health Insurance (acko.com))

Pain points

Because the nature of AI is to use data to assist in the decision-making process, many AI applications share similar pain points, accuracy and reliability, privacy and security concerns. (Rigby 2019).

Research suggested algorithm used AI technology could not achieve a high accuracy and shown a significant difference in error rate for areas related to races and gender. (Chen et al. 2019)

In terms of privacy and security concerns, AI algorithms requires personal data to train, predict and collect to perform chatbot functions, data being collected must be protected to prevent unauthorized access. Incidents such as cyber data breach around the world highlighted the crucial role of data and cybersecurity.

Why it is important

It is important to take account of digital trust and responsible AI in developing the algorithms.

When patients require treatment from a human physician, the interactions between the patient and physician often display empathy and trust was built between them. Because patients felt the physician showing empathy, the patients trust the physician's knowledge and skills, the patients have a face-to-face interaction with the physicians which help patients feeling the treatment process is transparent, and they knew they were treated on demand, and fair.

In most cases, AI is still a black box operation, it is difficult for patients and public to fully understand the limitation. It is important to build digital trust and responsible AI, because it can foster trust and confidence in AI technology by providing ethical services. In the long term it will benefit the society while the risk and challenges are low. (Dawson et al. 2019)

Problem definition

Significant costs associated with the establishment and ongoing maintenance of cloud computing infrastructure, compounded by restrictive exclusivity agreements which encourage long-term dependencies on proprietary products and services, has seen healthcare information systems monopolised by large-scale organisations (Panch et al. 2019). Inadequate governance frameworks regulating digital privacy and information sharing practices have moreover allowed these providers to become de-facto owners of personal healthcare data (Panch et al. 2019). As observed by Lehne et al. (2019), the conditions for data driven technologies expected to drive innovation in medicine are sub-optimal.

Ever increasing volumes of data generated from electronic medical records, hospital management systems, mobile applications, medical imaging, gene sequencing and wearable biosensor monitors holds remarkable potential (Lehne et al. 2019). By enabling new machine learning techniques, the lives of millions of patients worldwide could be improved through enhanced diagnostics, personalised treatments, earlier prevention of disease, accelerated drug discovery and more efficient service delivery (Lehne et al. 2019). However, healthcare data is only useful if it can be transformed into meaningful insights (Lehne et al. 2019). Balkanisation of info systems has therefore meant model inputs are hidden in isolated incompatible silos, made difficult to exchange, process & interpret (Panch et al. 2019; Lehne et al. 2019).

Even in the case of standard computerised tomography ('CT') scans, for example — a tool regularly employed to create detailed three-dimensional images of a patient's body including bones, organs, soft tissue and tumours — outputted sensor data cannot be integrated with the commercially sourced databases used by referring practitioners or specialists (Appen 2020; healthdirect 2021). This slows the speed at which potentially critical illnesses or injuries can be identified. More importantly though, it limits the ability for "big data" analytics to detect previously unrealised correlations from large-scale observational studies and inhibits the training of complex algorithms that could automate diagnoses at rates of accuracy higher than humanly possible (Lehne et al. 2019).

For the potential of medical digitalisation to be realised, solutions are required which support greater interconnectivity between data infrastructures and interfaces, standardise semantics and syntax, and "democratise" data access (Lehne et al 2019). In short, the successful deployment of artificial intelligence in healthcare depends on interoperability (Lehne et al 2019). Broadly defined as the ability for information systems to cooperatively exchange, integrate and leverage data across organisational, regional, and national boundaries, interoperability is crucial for the integration of multi-dimensional healthcare data into unified, open repositories (Sheban-Nejad et al. 2020). Without necessary action to resolve issues of interoperability, enthusiasm regarding the potential of artificial intelligence in healthcare will need to be downgraded (Panch et al. 2019).

Significance

The lack of interoperability and transparency can lead to various issues, such as
1) Privacy and cybersecurity - According to the brief from Office of Information Security and Health Sector Cybersecurity Coordination Centre (2023), the reported health data breach in the US was increased over 35% between 2019 (21.1 million) and 2022 (28.5 million). In 2022 the incidents reported to the HHS Office of Civil Rights indicated that around 79% of the incidents were related to Hacking / IT incident. Around 57% of the incidents were Network Severs related, and 23% of the incidents were email related. It is important that privacy and cybersecurity can give patients confident to provide personal information, which is essential for AI / ML to foster. (Fig 1)

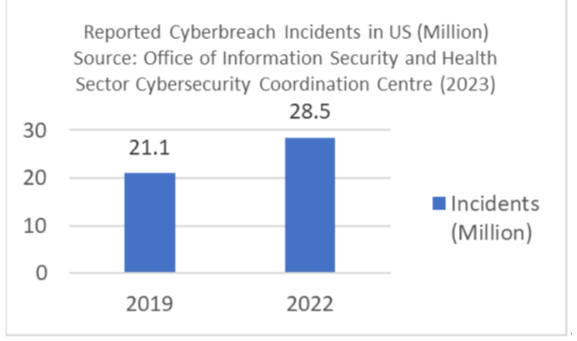


Fig 1

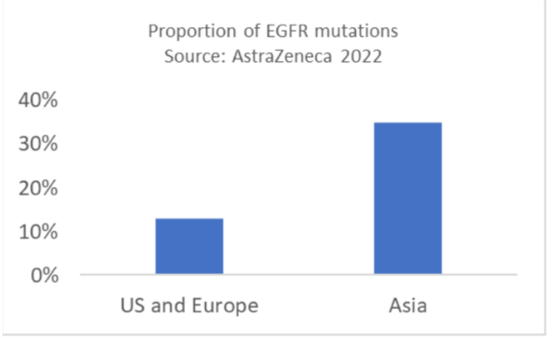


Fig 2

2) Training Data availability – According to AstraZeneca (2020), the majority of lung cancers are classified as non-small cell lung cancer (NSCLC), and the portion of EGFR (one of the biomarker for NSCLC) mutation is found to be higher in NSCLC patients in Asia (30-40%) than patients in the US and Europe, this demonstrated the challenges faced by researcher that, in order to train the cancer screening model properly, not only patient data collection from different countries is an issue, and also dealing with different regulators and jurisdictions. It is important that sufficient data to be collected to the AI /ML model to ensure fairness. (Fig 2)

3) Transparency and Accountability– In the article from STAT (2018), it documented that some training patient data which IBM Watson used in the cancer treatment system was synthetic, not from real patient. In addition, the production information from the IBM website in 2017 stated the model was trained with thousands of real patients' data, but according to the July 2017 presentation, it stated the range of 635 case of lung cancer to 106 ovarian cancers, which was not thousands as the IBM website suggested. Therefore, it raised transparency concern about the source and recommendation by IBM Watson.

Proposed data-driven solution

Stakeholders can look at improving the data collection strategy, currently, AI/ML in health care only use data from a limited source and scale, we propose researchers should begin incorporating inclusivity and fairness in AI model building process. We suggest researchers to consolidate and coordinate in data collection in global scale (such as organise through World Health Organisation) so that sufficient training data to cover the whole population and minimise bias.

Since the proposed data collection strategy can end up enormous amount of data collected, we proposed using models such as Large Language Model (LLM) or Reinforcement Learning with Human Feedback (RLHF) to train and perform tasks, for example stakeholders can utilise LLM to support appropriate treatment plan and diagnosis. Incorporate humanised face to provide human like user experience.

For data security and digital trust arise from the immersive amount of data and stakeholders involve, digital trust will need to be addressed. We propose a global recognised body such as WHO, to establish a new regulatory standard, which covers and combine the strictest standards from all current privacy legislation (e.g., California Consumer Privacy Act in the US and General Data Protection Regulation

(GDPR) in Europe), such that the stakeholders will have confidence in providing and using the data for AI/ML.

Our proposals will have potential ethical concern, firstly because it is a global scale operation, countries which do not have resources to implement the strategies will pose a risk, secondly, stakeholders such as health professionals might not be willing to take part of the initiatives due to job security concern. To handle these issues, we propose WHO to coordinate funding across countries and educate health professionals the benefits of AI/ML.

Methodology

To mitigate the emergence of black boxes in AI systems, address issues concerning digital trust, overcome interoperability challenges, it is imperative to adhere to the principles of Explainable AI (XAI) and FATE (Fairness, Accountability, Transparency, and Ethics). A framework of the eight core principles for AI is given by CSIRO in 2018. Accordingly, a comprehensive report should be generated during each step in the data science lifecycle to cover as much as possible of these eight principles.

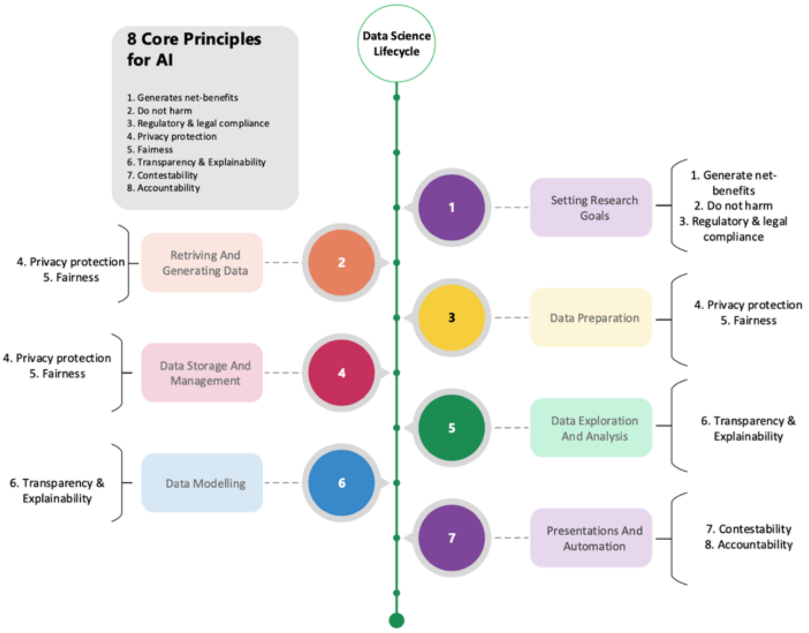


Figure: Data Science Lifecycle

1. **Setting the research goals:** During the stage of defining research goals, the primary focus lies in identifying the specific task and desired outputs that are anticipated from the model. These outputs could encompass personalized patient care plans, improved medical imaging, or administrative support, for instance. Organizations are required to articulate their intended objectives with the model and commit to employing it solely for those purposes, thereby ensuring informed consent and alignment with the desired goals. In this phase, the responsibilities and privileges of both organizations and users need to be distinctly defined, ensuring that either party can be held responsible in the event of a breach of these rights.

2. **Retrieving and generating data:** Based on the task identified in the previous step, the information/data required in this step must be clear. The organisations must clarify the type of data that will be collected from individuals and consent that no additional data will be collected. Also, the organisations must clarify the methods of collecting the data. Organisations must avoid any type of bias when collecting data such as historical, population, sampling, or social biases. When collecting data from users, their consent must be taken as it is part to ensure privacy. One of interoperability challenges is the Inconsistent data and lack of standardized data structure. The principal strategy for attaining interoperability rests on the utilization of open Application Programming Interfaces, commonly known as APIs. APIs facilitate communication between divergent applications and systems, thereby promoting the exchange of the data and protected health information (PHI) across Electronic Health Records (EHRs) and various health information technology platforms (Ali, 2022).

3. **Data preparation:** Whether organisations are cleaning the data from errors and outliers, transforming the data, or combining data, they must explain the reason behind changing users' data in both statistical language to professional and to use a language that is clear, concise, and accessible to public. The development or use of the AI system must not result in unfair discrimination against individuals, communities, or groups. This require particular attention to ensure the training data is free from bias or characteristics which may cause the algorithm to behave unfairly.

4. **Data Storage and Management:** In this step, despite what type of storage the organisations are using, the two focuses here is to ensure data privacy and security. There are several things to consider ensuring data privacy such as regularity compliance, data minimization and data breach notifications. Regularity compliance is to ensure various laws and regulations govern how personal data should be handled. Data minimization is that organisations should only collect the minimum amount of personal data necessary for their purposes and should not keep it for longer than needed. And finally, when a data breach occur, organisations are required to notify affected individuals. For data security, concepts of cybersecurity such as access control, data encryption, data backup, firewalls, and network security.

5. **Data Exploration and Analysis:** The type of statistical analysis that are made by the organisation must be clarified. For example, WHO states in privacy policy that they perform statistical analysis on personal data. Feature engineering and machine learning techniques can unravel interesting information about individuals. Hence, the user must have a clear understanding of the kind of statistical analysis being used, its results and its intended purpose. On the other hand, Intentional or not, data visualisations can obfuscate and deceive audience (Bresciani and Eppler 2008). Organisations must adhere principles to deliver visualisations that are clear and meaningful. IBM have identified five key principles for an effective data visualisation. The data visualisation must be understandable, essential, impactful, consistent, and contextual.

6. **Data Modelling:** In real life, we are tempted to trust persons if they can explain to us why and how they do what they do. Having a general understanding of how the AI system is built is crucial for the user to build that trust and transparency. A good example of this is chat gpt. In the main page of OpenAI, samples, methods limitations and iterative deployment are explained. Familiarity with the system's limitations can significantly enhance the user's comfort and confidence in using it. The before to use page of chat gpt explains the capabilities and the limitations of it. For example, one of the capabilities is that it remembers what user said earlier in conversation. One of the limitations is that it may occasionally generate incorrect information (OpenAI, 2022).

7. **Presentations and automation:** When the AI system is operational, it is crucial for the institution to continuously revisit and apply the eight foundational guidelines for AI put forth by CSIRO and Data61. Nonetheless, priority should be assigned to the principles of contestability and accountability. The principle of contestability implies that if a person's situation is affected by an algorithm, a proficient mechanism must exist that enables the person to challenge the usage or results of that algorithm. In situations where a user experiences any anticipated or unexpected effects, the organization involved bears the responsibility and is to be held accountable (CSIRO, 2022).

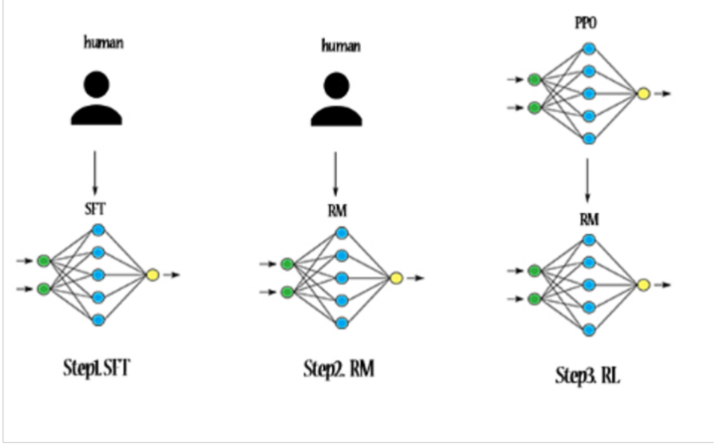
Design Prototype

LLMs – Large Language Models, have now been applied to a wide range of tasks that are applicable to Healthcare AI systems. Language translation, text responses and generative AI text conversations can now be enabled by LLMs. The ability of OpenAI's ChatGPT to engage in human-like conversations is now widely known and accepted. One example of where LLMs can now be used effectively is with healthcare system AI chatbots such as a next generation of the WHO's Florence 2.0 deployed during the Covid pandemic.

"LLMs have shown remarkable capabilities in a wide range of NLP tasks. However, these models may sometimes exhibit unintended behaviors, e.g., fabricating false information, pursuing inaccurate objectives, and producing harmful, misleading, and biased expressions. To avert these unexpected behaviors, human alignment has been proposed to make LLMs act with human expectations (La Vivien, 2023)."

To address these key trust-related concerns (especially in the context of Healthcare AI systems), we propose that the technique of Reinforcement Learning with Human Feedback (**RLHF**) be embedded in the training methods of Healthcare AI systems. RLHF incorporates humans in the training loop and uses human preferences as a "reward signal" to fine-tune LLMs. (La Vivien, 2023).

RLHF - Reinforcement Learning with Human Feedback steps (La Vivien, 2023)



1. ***Supervised fine-tuning (SFT)***

"Collect demo data & train a supervised policy. Labellers provide desired behaviour demos on input prompt distribution. Fine-tune pretrained model using supervised learning."

2. ***Rewording model training (RM)***

"Collect comparison data & train a reward model. Team collects a dataset of comparisons between model outputs labellers indicate which output they prefer for a given input. Then trains a reward model to predict the human-preferred output."

3. ***Reinforcement Learning fine-tuning (RL)***

*"Optimize a policy against the reward model using **PPO**. The team uses the output of the RM as a scalar reward. Then fine-tunes the supervised policy to optimize this reward using the Proximal Policy Optimization PPO algorithm.*

Conclusion

Incorporation of our proposed use of LLMs and most importantly, RLHF, to be further defined and stated within the guidelines specified by governing bodies such as WHO referenced in our problem definition above. Progression through a full Data Science Lifecycle with emphasis on the ***Eight Core Principles of AI*** discussed in the Methodology section above should be included as part of any Heathcare AI system development. Existing Global IT Risk & Audit firms should all be well positioned to help make sure the guidelines become best practice and measure up to generally accepted international standards for future AI audit and compliance.

Bibliography

Ali, N. (2022) EHR interoperability challenges and solutions. Available at: https://www.ehrinpractice.com/ehr-interoperability-challenges-solutions.html#:~:text=The%20most%20common%20approach%20for,and%20health%20information%20technology%20systems. (Accessed: 30 May 2023).

Bresciani, S., and M. J. Eppler. 2008. “The risks of visualization: A classification of disadvantages graphic representations of information.” Institute for Corp Commo https://pdfs.semanticscholar.org/23d2/3f5152c9b8b34f104b43d1c862ee62d2edac.pdf.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.
CSIRO (2018) Australia Ai Ethics Framework, CSIRO. Available at: https://www.csiro.au/en/research/technology-space/ai/ai-ethics-framework (Accessed: 31 May 2023).

OpenAI (2022) ChatGPT, Introducing ChatGPT. Available at: https://openai.com/blog/chatgpt (Accessed: 01 June 2023).

Rigby ,Michael J. (2019) Ethical Dimensions of Using Artificial Intelligence in Health Care ,Journal of Ethics | AMA (ama-assn.org) (Accessed 1 May 2023) https://journalofethics.ama-assn.org/article/ethical-dimensions-using-artificial-intelligence-health-care/2019-02

National Health and Medical Research Council, (2018), National Statement on Ethical Conduct in Human Research (2007) - Updated 2018. Accessed 1 May 2023. https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2007-updated-2018

Role of Artificial Intelligence (AI) in Health Insurance (acko.com)
Team Acko 2023, Understanding the Role of AI in Health Insurance, Team Acko, accessed 1 May 2023. https://www.acko.com/health-insurance/ai-artificial-intelligence-in-health-insurance/

Chen I, Szolovits P, and Ghassemi M (2019) Can AI Help Reduce Disparities in General Medical and Mental Health Care Journal of Ethics | AMA (ama-assn.org) https://journalofethics.ama-assn.org/article/can-ai-help-reduce-disparities-general-medical-and-mental-health-care/2019-02 (Assessed 1 May 2023)

Morgan M (2023) Why Artificial Intelligence Is Becoming A Cybersecurity Imperative And How To Implement It (forbes.com), accessed 1 May 2023 https://www.forbes.com/sites/forbestechcouncil/2023/03/15/why-artificial-intelligence-is-becoming-a-cybersecurity-imperative-and-how-to-implement-it/?sh=5172d1d9610d

Dawson D and Schleiger E* , Horton J, McLaughlin J, Robinson C∞, Quezada G, Scowcroft J, and Hajkowicz S† (2019) Artificial Intelligence: Australia’s Ethics Framework. Data61 CSIRO, Australia. Accessed 1 May 2023 https://www.csiro.au/-/media/D61/Reports/Artificial-Intelligence-ethics-framework.pdf

Chen I Y., Szolovits P,, and Ghassemi M, 2019 Can AI Help Reduce Disparities in General Medical and Mental Health Care? | Journal of Ethics | American Medical Association (ama-assn.org) accessed 1 May 2023. https://journalofethics.ama-assn.org/article/can-ai-help-reduce-disparities-general-medical-and-mental-health-care/2019-02

Office of Information Security & Health Sector Cybersecurity Coordination Center, (2023), Data Exfiltration Trends in Healthcare. accessed 1 May 2023. https://www.hhs.gov/sites/default/files/data-exfiltration-in-healthcare-tlpclear.pdf

Ross C and Swetlitz I (2018), IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show, STAT, accessed 1 May 2023. https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/

AstraZeneca (2020) Lung Cancer in Asia , accessed 1 May 2023 (https://www.astrazeneca.com/content/dam/az/our-focus-areas/Oncology/2020/lungcancer/Lung%20Cancer%20in%20Asia%20Backgrounder_APPROVED_2020.pdf)

Appen (2020) *What does interoperability mean for the future of machine learning?*, Appen website, accessed 15 May 2023. https://appen.com/blog/what-does-interoperability-mean-for-the-future-of-machine-learning/?amp.

Harrer S (2023) 'Attention is not all you need: The complicated case of ethical using large language models in healthcare and medicine', *eBioMedicine*, 90. https://doi.org/10.1016/j.ebiom.2023.104512.

healthdirect (2021) *CT scan*, healthdirect website, accessed 15 May 2023. https://www.healthdirect.gov.au/ct-scan.

Lehne M, Sass J, Essenwanger A, Schepers J and Thun S (2019) 'Why digital medicine depends on interoperability', *npc Digital Medicine*, 2(79). https://doi.org/10.1038/s41746-019-0158-1.

Panch T, Mattie H and Celi L (2019) 'The "inconvenient truth" about AI in healthcare', *npj Digital Medicine*, 2(77). https://doi.org/10.1038/s41746-019-0155-4.

Sheban-Nejad A, Michalowski M and Buckeridge D (2020) 'Explainability and interpretability: Keys to deep medicine', in Sheban-Nejad A, Michalowski M and Buckeridge D (Eds) *Explainable AI in Healthcare and Medicine*, Springer Cham. https://doi.org/10.1007/978-3-030-53352-6_1.

La Vivien Post (2023), 'How ChatGPT works – Architecture illustrated (viewed 1 June 2023). https://www.lavivienpost.com/how-chatgpt-works-architecture-illustrated/