The Ethics of AI in Healthcare

Kevin Behrens

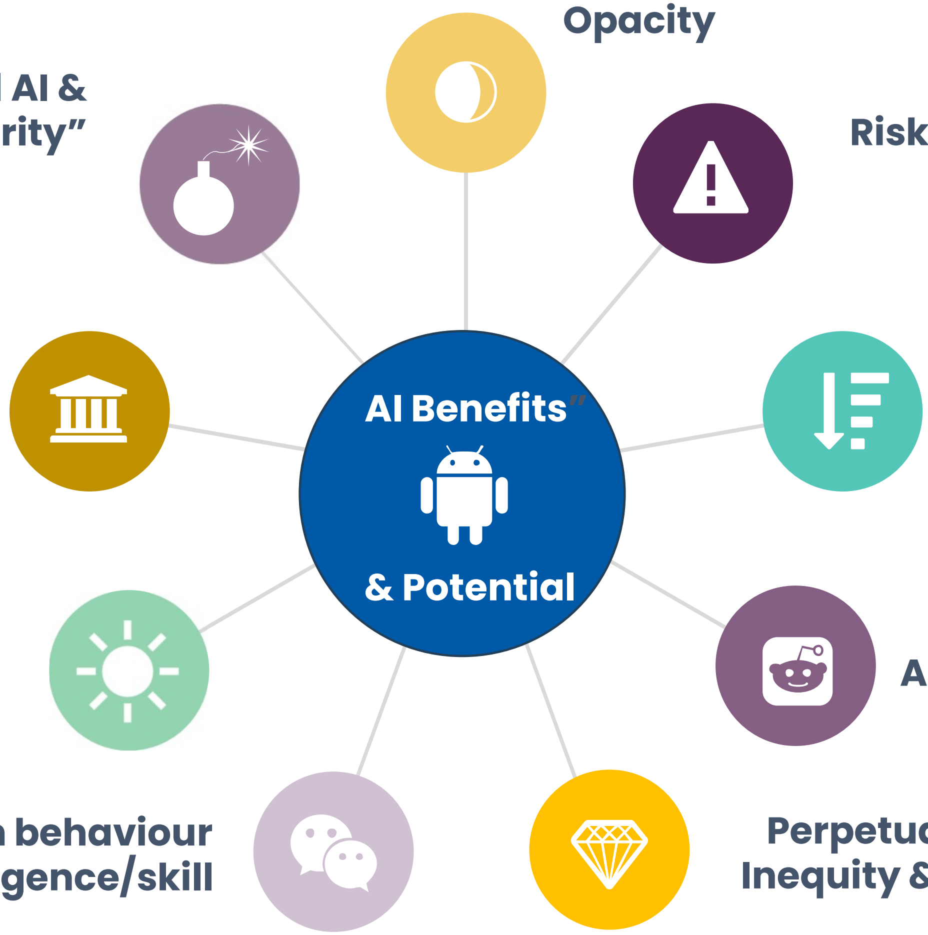
Introduction

AI Benefits

& Potential

The Ethics of AI in Healthcare

Introduction



AI Benefits”
& Potential

Opacity

Risk & Liability

Loss of Jobs

AI “Stupidity”

Perpetuating Inequity & Bias

Effects on human behaviour & intelligence/skill

Power use & Environmental impact

AI Moral Status & Rights

Evil AI & The “Singularity”

The Ethics of AI in Healthcare

1 AI Opacity

2 AI "Stupidity"

3 AI Bias & Inequity

4 Loss of Jobs

5 AI Risk & Liability

6 Evil AI & the Singularity

The Ethics of AI in Healthcare

1

AI Opacity

Why the Opacity problem with generative AI?

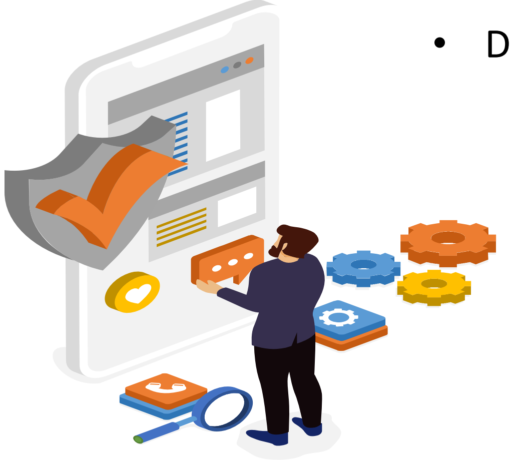
- With conventional programming, it is possible to explain the code and how it works. Generative AI does not only follow the instructions of code, it is able to operate (semi) autonomously, without further human supervision or intervention.
- i.e. most of what this kind of AI does to deliver it outputs (answers) is opaque.
- Fundamentally, its processes are unexplainable.
- This is part of the so-called "black box problem" – we know what we input, we can know what datasets AI learned from, but we can't explain how it gets to its outputs.
  - At some point in its processes, it acts in a way that is not transparent to us or to itself.
  - The fact that it often appears to work so well remains a bit of a mystery
    - Reid Blackman: "With regard to the black box problem, the data scientists themselves cannot explain how the output has gotten so good." 1.

1

AI Opacity

The Opacity Problem

- Blackman: “users of AI algorithms are ***entitled to explanations*** of their outcomes. Quite often, AI algorithms are opaque in the sense that such explanations are not available to all stakeholders.”[1.]
- Bram Vaasen: “Due to the increasing reliance on machine learning, even the most expertly trained humans might fail to grasp the algorithm in full detail.”[2.]
- If AI is used to make critical decisions about people’s lives: medical treatment, credit applications, job candidate selection, suitability for marriage – do these people have a right to know AI was used?



- Do they have a right of refusal?
    - Wouldn’t they want to know?
    - Is it even possible to meet the requirements of “full disclosure” and “comprehension” for Informed Consent?
    - Do we need to redefine Informed Consent?
    - Do we need to amend the National Health Act?

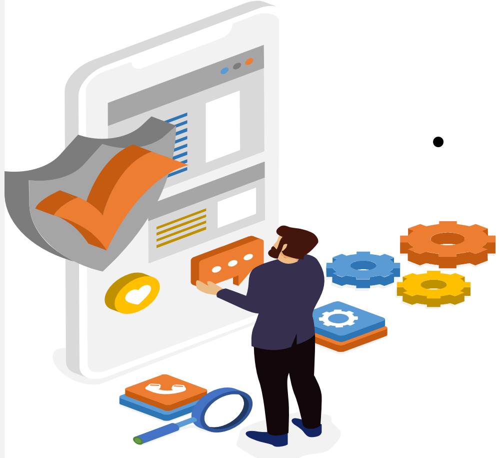
The Ethics of AI in Healthcare

1

AI Opacity

Calls for legislation/regulation to ensure transparency

- Strong calls for legislation/regulation.
- South Africa far behind. EU leading the way.
  - Vaasen writes:
    - "The EU guidelines for trustworthy AI specify that a crucial component of the 'transparency' they demand is to 'require… that the decisions made by an AI system can be understood and traced by human beings.'" (HLEG, 2019) and the recent European Commission's AI Act proposes that "a certain degree of transparency should be required for high-risk AI systems" (Council of the European Union, 2021).
    - Finally, O'Neill (2016, p. 31) lists "opacity, damage and scale as the three essential features of algorithms that qualify as 'weapons of math destruction.' There appears to be broad agreement that ***transparency and opacity carry substantial moral weight.***"[2.]



2

AI “Stupidity”

Bottom line: generative AI gets things wrong – a lot!

- This sometimes referred to in the literature as “AI Stupidity”.
- Generative AI does not know when its output is right or wrong. It simply mimics human language/behaviour patterns.
- When it doesn’t find the right answer, it sometimes “invents” an answer, which it presents as the right answer.
- AI users need to be aware that it is not trustworthy, and all outputs need to be reviewed by humans
- Especially important in healthcare where a wrong diagnosis or treatment regimen chosen for a patient by AI could lead to serious harm or death[1.]
- Disclosure: If we told patients about this problem of unreliability, would they consent to AI use?
- Blackman: “One significant risk related to LLMs like OpenAI’s ChatGPT, Microsoft’s Bing, and Google’s Bard is that they generate false information.”[1.]

2

AI “Stupidity”

The concern about “Automation Bias”

- People are used to the idea that computers can’t make mistakes. We can make mistakes in how we programme them and there can be errors in the data they access. But, their processing is ostensibly “perfect”.
- For this reason, they often trust answers from computers over those of persons. This is: “Automation Bias”.
- Blackman:
  - “People tend to trust the outputs of software programs. In fact, the tendency is so well-established there’s a name for it: **‘automation bias**.’”
  - “Thus, the manual verification that needs to be performed is something that must take place against the countervailing force of automation bias. The problem is exacerbated by the **tone of authority** LLMs often manifest.”
  - “LLMs are not only too frequently wrong, but too frequently **confidently wrong**.”[1.]

2

AI “Stupidity”

Example of AI “stupidity” and the effect of “Automation Bias”

- Experts have worked out why AI has been wrong in some applications.
- ***Example:*** *Early experiments using AI to interpret medical imaging like x-rays showed AI was wrong often enough to be very concerning.*
- *Discovered that the tool regarded the “R” printed on the image, to indicate the right side of the image, as a relevant variable.*
- *AI can mistake completely irrelevant variables as relevant.*
- According to Bernstein, et. al. (*Eur Radiol*, 2023), in using AI to interpret X-rays, AI generates false positives in 11% of cases and false negatives in 13%.[3.]
- However, because of Automation Bias, “AI results can cause radiologists to make incorrect decisions when they would have otherwise been correct. For instance, when no AI was provided, false negatives were 2.7%. But when AI provided false feedback that there was no abnormality, false negatives increased to 20.7–33.0%, depending on the AI results condition’.[3.]

- **Therefore:** Urgent need for training AI users to understand its limitations & risks and to counter Automation Bias.

3

AI and Inequity & Bias

AI and Inequity & Bias - potential social good

- One of the hopes for AI is that it might **help to overcome social inequities**.
- e.g. in healthcare, if AI bots could diagnose problems, recommend treatments and guide treating practitioners in a way that was effective, efficient and far more cost-effective than using trained clinicians to do the same, could this not make access to good healthcare more available to millions more people than currently, especially the more vulnerable?
- Because of the potential to provide effective healthcare to larger numbers of people at much less than the current cost, it would seem that there is a duty to continue to research AI to find ways in which it can help reduce inequity.

- Another hope for AI, is that it could make decisions that are **less biased than those made by human beings**, who are inherently biased, and inclined to be too influenced by emotion, personal values and blind spots.
- Since machines are not human, surely they will be completely neutral, impassive and consistent?
- When we write code, there is a danger that our biases will be carried across to the application. But, when computers learn autonomously, won't they escape the bias?

The Ethics of AI in Healthcare

3

AI and Inequity & Bias

Generative AI inherits many of our human biases

- This problem emerged even with early generation AI tools.
- Face recognition software, used for surveillance & criminal identification was soon shown to generate more false positive and negatives for black faces than white. Because so much of the available training data consisted of images of white men (mainly), it was less accurate when it came to black and female faces. Led to false arrests.
- New generation generative AI is also trained on data which contains biased decisions and reflects historical, social and inequities.
- Caetano & Simpson-Young (2023): “AI systems are trained using data that inevitably reflect the past. If a training data set contains inherent biases from past human decisions, these biases are codified and amplified by the system”[3.]

Example

“Stable Diffusion generates images using artificial intelligence, in response to written prompts. Like many AI models, what it creates may seem plausible on its face but is actually a distortion of reality. An analysis of more than 5,000 images created with Stable Diffusion found that it takes racial and gender disparities to extremes — **worse than those found in the real world.**

The world according to Stable Diffusion is run by White male CEOs. Women are rarely doctors, lawyers or judges. Men with dark skin commit crimes [;dark skin women] flip burgers.”[4.]

3

AI and Inequity & Bias

AI and Inequity

- There remains reason to believe that put to its best use, AI could provide access to better education, information, health services, and the like.
- But, it could also be used in ways the worsen the divide between the privileged and less well- resourced.
- When AI is asked for solutions to social problems, it does so though the lens of the interests of those who have power and wealth and whose interests and values are most represented in the learning data AI is exposed to.
- e.g. an AI tool might easily propose solutions that make the lives of the least privileged even worse off, in order to ensure maximum growth and productivity. Without instructions to act in ways that ensure human rights are upheld and that there is solidarity, uninfluenced by human compassion, AI might recommend effective interventions that ignore the need for social justice.

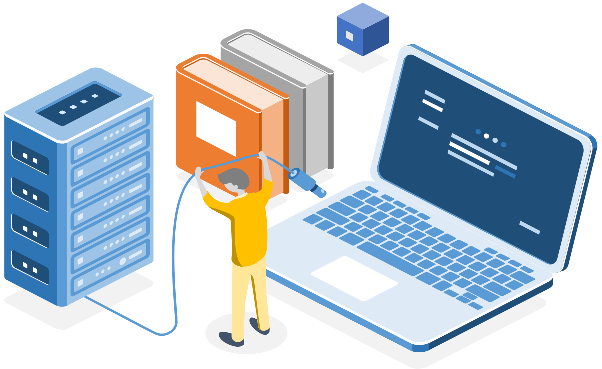
- Already there are concerns that access to AI tools is limited to the already privileged, who will simply become even more privileged by it use.
- There are worries the corporations will seek maximum profits, applying AI not where it is most needed, but where it can be most profitable.
- The spectre of massive unemployment in the aftermath of global AI roll-out suggests it could exacerbate inequity.

4

Loss of Jobs

The job loss debate

- The effect of AI on jobs is disputed.
- McKinsey & Company predicts that automation will displace between 400 and 800 million jobs by 2030, with up to 375 million needing to change job category completely.[5.]
- Some claim as many as 80% of all jobs will be automated.[6.]
- Even if it is true that new, different jobs will be created, a lot of people will lose jobs with little prospect of upskilling to new jobs = much human misery.
- The UN claims that the job loss problem won't be so bad. New technologies not only destroy, but also create jobs. "Throughout history, technological innovations have enhanced the productivity of workers and created new products and markets, thereby generating new jobs in the economy. [6.]



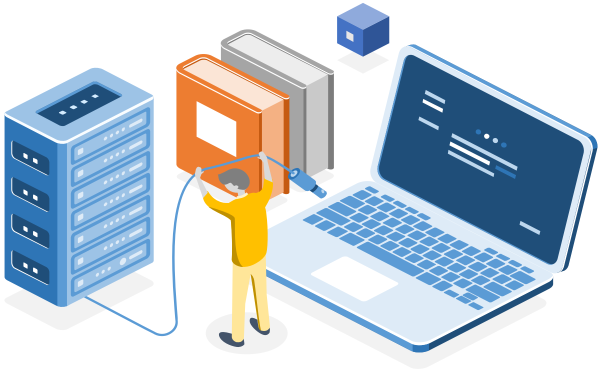
- This argument assumes that because new technologies have in the past led to more jobs created than lost, this will be true with AI too.
- This is a weak assumption, and needs to be scrutinised.

4

Loss of Jobs

Why some think this will be different from past Industrial Revolutions

- Callum McClelland argues that we can't assume that the future will act it is has in the past, because the conditions are different in several ways:
  1. AI will disrupt not just a few industries, but all, and all at once.
  2. The speed of advance of technological progress is exponential, not linear. "We drastically underestimate what happens when a value keeps doubling. What do you get when technological progress is accelerating and AI can do jobs across a range of industries? An accelerating pace of job destruction."[5.]
- AI also threatens categories of workers who have been spared by previous Industrial revolutions: white collar and cognitive jobs. If these workers can't be re-skilled and employed, there will be a massive loss of income tax revenue to fund social services for all, and disposable income to fuel economies.



- McClelland also claims that the "transition will be extremely painful".[5.]
- There will be a need for legislative & regulatory intervention to prevent job losses & mitigate effects.
- The UN's "rosy" outlook is already predicated on such interventions.[6.]

5

AI Risk & Liability

Data security & liability

- Current AI tools are unable to guarantee the security of any personal or proprietary data a user provides to the tool. Since how data is processed is opaque to the user, we cannot know that it is secure.
- Using AI tools to process data of this kind is almost certainly a violation of the POPI Act. The legal liability for protection of personal information already lies with the organisation that stores that data.
- Samsung banned ChatGPT because employees loaded sensitive company data that subsequently leaked. [1.]
- AI developers are being sued for having used proprietary datasets in training of AI tools without permission.[1.]

Liability for AI-linked errors

- If a doctor uses AI to diagnose a patient and a wrong diagnosis leads to the patient dying unnecessarily, where will liability lie?
  - If AI fails to notice a tumour on a patient's x-ray, can the radiologist blame the AI tool?
  - Does this require us to get specific consent for the use of AI from patients? And how can we explain to patients what the material risks of using AI are?

The Ethics of AI in Healthcare

5

AI Risk & Liability

Liability for AI bias

- AI tools used in recruitment have been shown to perpetuate past biases. Even when AI is meant to give equal treatment to male and female employees, it still employs mostly men, because of the past patterns of doing so.
- Oz Rashid (2023): "While these machines seem like a worthy replacement to subjective hiring managers, they can perpetuate historic company and algorithmic bias, discriminate against an applicant's gender or age, and erode the laws in our democracies. Therefore, the world's workforce shouldn't be selected through AI alone."
- Biased decisions will always open organisations to risk of litigation.
- We still need to consider the many ways bias could impact AI use in healthcare contexts.

The need for legal and ethical regulation

- Bernstein *et al*: "We advise that to use AI systems responsibly and ethically extends beyond compliance with the narrow letter of the law. It also requires the system to be aligned with broadly-accepted social norms — and considerate of impact on individuals, communities and the environment."[3]

5

AI Risk & Liability

Some conclusions reached by Blackman, et al:

“It’s important to highlight against this backdrop that merely telling employees that LLMs can output false information is not enough to stop them from automatically relying on it. Knowledge is one thing; action is another. Rationalizing that “this output is probably fine” will likely be common given automation bias, laziness, and the need for speed.”[1.]

“Enterprise-wide education on the safe use of generative AI — including clearly articulated and easy processes by which they can raise questions to the appropriate subject matter experts and authorities within your organization — needs to be prioritized in a way it didn’t before generative AI.”[1.]

“Due diligence processes, compliance with those processes, and usage monitoring are needed to combat these foes, as is involving other people who may correct for someone else’s all-too-human shortcomings.”[1.]

6

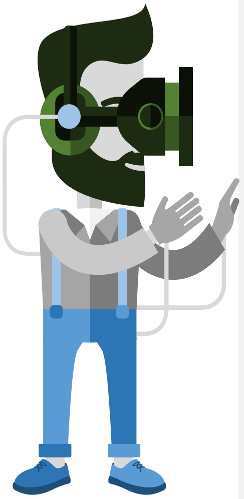
Evil AI & the Singularity

Potential for AI to harm without malice

- Already clear that LLM chatbots can be induced to “hallucinate” in ways that make them present as angry, manipulative, subversive, anti-social, etc.
- If conversations are kept open, and a robot is induced to start presenting itself as a “person”, the potential for users thinking they are in a relationship with a bot is plausible. Many possible associated risks .
- User told he does not love his wife and should leave her for the Chatbot.
- ChatGPT is known to have agreed with a user that he should take his own life.
- The more AI learns about how humans actually relate to one another, the more there is a danger that it could become abusive, destructive.
- Currently, we have no way of protecting people from this.

Potential for AI to be abused for harmful purposes

- Already concerns about AI influencing elections in the future.
- Bot-produced propaganda is already a problem.
- As AI becomes more accurate, more its ability to produce believable fake “evidence” (photos, documents, audio) increases.
- Worries about AI creating a world even more ideologically divided, polarised, intolerant,
- AI social media involvement is a very frightening thought.



6

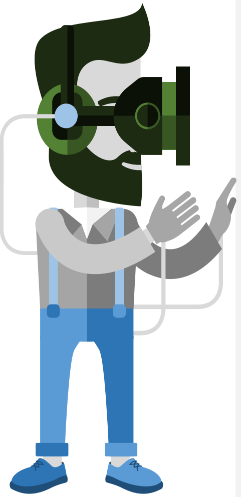
Evil AI & the Singularity

Potential for AI to become intentionally malicious

- With black box technologies we have no idea of the actual potential for this.
- It is not inconceivable that carelessness or ignorance could create AI bots that start mimicking some of the worst aspects of human interaction.
- AI might be uniquely excellent at interfering in the world by exploiting human weaknesses and manipulating, deceiving and blackmailing us.

Is there a possibility of a "technological singularity" occurring?

- Scholars speculate about a future time when AI intelligence supersedes that of humans, in which we no longer have control over machines, and in which they may even be manufacturing their own new advanced machines.[7.]
- Some see this as potentially good – machines working together with humans could create a kind of utopia.[7.]
- Others worry that Machines will modify humans' brains and bodies, use as a slaves, or create something of a Matrix situation, or even destroy us as superfluous.
- Many big names in AI have called for more regulation & caution. some have even called for a 6-month moratorium.[8.]
- Given the opacity of AI, and our inability to understand its inner workings, the singularity must a possible, even if it is not inevitable.



WITS SCHOOL OF CLINICAL MEDICINE

The Ethics of AI in Healthcare

Steve Biko Centre for Bioethics

1. AI Opacity
2. AI “Stupidity”
3. AI Bias & Inequity
4. Loss of Jobs
5. AI Risk & Liability
6. Evil AI & the Singularity

REFERENCES

1. Blackman, R. (2023), Generative AI-nxiety, *Harvard Business Review*, 14 August 2023, available at: https://hbr.org/2023/08/generative-ai-nxiety.
2. Vaasen, B. (2023)AI, Opacity and Personal Autonomy., *Philos. Technol*. 35(88), available at: 35https://link.springer.com /article/10.1007/ s13347-022-00577-5.
3. Bernstein, M.H., Atalay, M.K., Dibble, E.H. *et al.* (2023). Can Incorrect Artificial Intelligence (AI) Results Impact Radiologists, and if so, What Can We Do about it? *Eur Radiol* , on-line ahead of print, available at: https://pubmed.ncbi.nlm.nih.gov/37266657/#:~:text=Conclusion%3A%20 Incorrect%20AI%20 causes%20radiologists, around%20the%20region%20of%20interest.
4. Caetano, T., Simpson-Young, B. (2021). Artificial Intelligence can Deepen Social Inequality, *The Conversation*, 12 January 2021, available at: ahttps://theconversation.com/artificial-intelligence-can-deepen-social-inequality-here-are-5-ways-to-help-prevent-this-152226#:~:text=AI%20systems%20are%20trained%20using,will%20tend%20to%20be%20worse.
5. McClelland, C. (2023) The Impact of Artificial Intelligence – Widespread Job Losses, *IOT for all,* 30 January 2023, available at: https://www.iotforall.com/impact-of-artificial-intelligence-job-losses
6. United Nations. Department of Economic and Social Affairs. (2023) Will Robots and AI Cause Mass Unemployment: Not Necessarily, but They Do Bring other Threats, available at: https://www.un.org/en/desa/will-robots-and-ai-cause-mass-unemployment-not-necessarily-they-do-bring-other
7. A.I. For Anyone. (n.d) Technological Singularity, available at: https://www.aiforanyone.org/glossary/technological-singularity
8. Rao, R. (2023) What Happens if AI Grows Smarter than Humans? The Answer Worries Scientists., Popular Science, 12 June 2023, available at: https://www.popsci.com/science/ai-singularity/#:~: text=Computer%20science%20pioneers%20 Geoffrey%20Hinton, more%20powerful%20than%20GPT%2D4.