



Ethical Considerations for the Translational Application and Review of Biomedical Research Involving AI — A Briefing Document



NUS
National University
of Singapore

Centre for Biomedical Ethics
Yong Loo Lin School of Medicine



SHAPES

An NUS Centre for Biomedical Ethics initiative supported by the
Singapore Ministry of Health's National Medical Research Council

Acknowledgements

This document would not have been possible without strong collaborative efforts among the involved parties.

Alexa Nord-Bronzyk is the main author of this document, working closely with Michael Dunn as the co-author. There was strong support from the NUS Centre for Biomedical Ethics team, especially Sumytra Menon, Jerry Menikoff, and Julian Savulescu.

The foundation for the arguments presented in this document draws upon ongoing work from Science, Health and Policy-relevant Ethics in Singapore (SHAPES), an NUS Centre for Biomedical Ethics initiative supported by the Singapore Ministry of Health's National Medical Research Council. The authors involved include Kathryn Muyskens, Harisan Nasir, Angela Ballantyne, Murali and Julian Savulescu, Yonghui Ma, Jerry Menikoff, and James Hallinan.¹

The SHAPES AI working group was also instrumental in the completion of this document. Reviewers include Tsung-Ling, Ma Yonghui 马永慧, Angela Ballantyne, and Pavitra Krishnaswamy. An extended thank you to the wider working group for their ongoing support.

The SHAPES team is deeply grateful to all involved in the development of this document. Thank you all for your time, effort, and invaluable insights.

¹ See methodology for further details on the involvement of these authors.

Contents:

1. Introduction

- 1.1. What's in this document?
- 1.2. Who is it for?
- 1.3. What should you know?
- 1.4. Methodology

2. Bias— Is it ever justifiable to implement biased AI?

- 2.1. What is bias in AI?
- 2.2. Case Study: The Permissibility of Biased AI in a Biased World: An Ethical Analysis of AI for Screening and Referrals for Diabetic Retinopathy in Singapore
- 2.3. Why not just remove the bias?
- 2.4. How should utility and equity be considered in evaluating the implementation of biased AI?
- 2.5. What principles apply in deciding when implementing a biased AI is justified?
- 2.6. Strategic Measures to Mitigate Bias in Practice – What can you do?

3. Human Involvement – How and to what extent?

- 3.1. What does human involvement in AI mean?
- 3.2. Case study: Spine AI: Medical Imaging For Lumbar Spinal Stenosis
- 3.3. Leaving Humans Out of the Loop – What Are the Main Concerns?
- 3.4. What are the criteria for kicking humans out of the loop? – Benefits vs. Costs
- 3.5. What principles apply in deciding the role of human involvement?
- 3.6. Strategic Measures in Practice to Ensure Appropriate Human Involvement – What can you do?

4. The Risks of Risk Prediction – How much risk is reasonable in an already risky world?

- 4.1. What is risk prediction in AI?
- 4.2. Case Study Score for Emergency Risk Prediction (SERP) -- Machine Learning Triage Tool For Estimating Mortality After Emergency Admissions
- 4.3. SERP Risk Prediction – What is the Main Concern?
- 4.4. Assessing Risk – Are AI risks exceptional?
- 4.5. To Implement or Not- What Are the Tradeoffs?
- 4.6. Moving Forward with Implementation – How should risk mitigation strategies be evaluated?
- 4.7. What principles apply in deciding a risk threshold for implementation as well as risk mitigation strategies?
- 4.8. Strategic Measures in Practice to Evaluate Reasonable Risk Involvement – What can you do?

5. Conclusion

6. Annex

1. Introduction

1.1 What's in this document?

This document discusses key ethical considerations surrounding the pipeline of research and development activities focused on translating AI into healthcare. The focus is narrowed to three pertinent ethical issues: bias, human involvement, and risk prediction. Each theme is discussed in relation to a Singapore research case study and offers recommendations grounded in an understanding of local research practice. The guidance presented intends to offer focused insights into resolving local real-world challenges.²

1.2 Who is it for?

This document is for researchers, research institutions and IRBs in Singapore who wish to gain insights and new perspectives on how to work through key ethical issues as they navigate translational clinical research, including observational studies and RCTs. This also includes those who are in the business of translating new technology into practice and then evaluating that technology in practice once implemented, which might also include clinicians.

1.3 What should you know?

While artificial intelligence (AI) has been an academic discipline since 1956, it has received significant attention in recent years due to the advent of generative AI and the myriad resulting applications.³ The potential for applications of AI in healthcare promises great benefits not only for clinicians and patients, but also for the wider society. With those benefits also come risks and challenges. As advancements in these technologies rapidly progress, ethical issues arise as the potential for sweeping impact mounts. Below is an overview of the basic information you should know before diving into the ethical issues that follow.

Definition of AI – The present document endorses the Nuffield Foundation's definition of artificial intelligence as “any technology that performs tasks that might be considered intelligent – while recognizing that our beliefs about what counts as intelligent may change over time”.⁴ Put otherwise, when a task normally performed by humans (due to the need for human intelligence) is performed by a computer, it is thought to exhibit artificial intelligence (AI).⁵

Furthermore, this document will reference AI tools as opposed to AI in general. AI in general is too broad for our purposes by referring to any kind of AI system including artificial general

² Each section of this document draws from forthcoming papers by the CBmE SHAPES team. The ideas presented in this document reflect the conclusions from these papers and have been adapted for the purposes and audience this document is intended for. See methodology for further details.

³ James Moor, “The Dartmouth College Artificial Intelligence Conference: The next Fifty Years,” *AI Magazine* 27, no. 4 (2006): 87–87, <https://doi.org/10.1609/aimag.v27i4.1911>.

⁴ Jesse Whittlestone et al., “Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: A Road Map for Research,” <https://www.nuffieldfoundation.org/Sites/Default/Files/Files/Ethical-And-Societal-Implications-of-Data-And-AIreport-Nuffield-Foundat.pdf>, 2019.

⁵ Jesse Whittlestone et al., “Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: A Road Map for Research,” <https://www.nuffieldfoundation.org/Sites/Default/Files/Files/Ethical-And-Societal-Implications-of-Data-And-AIreport-Nuffield-Foundat.pdf>, 2019.

intelligence (AGI) – where the AI is indistinguishable from human intelligence – which does not yet exist. “AI tools” focuses on those used in practical applications and which have proven successful in practice in recent years, thus aligning with our focus on the translational application of AI.

AI Applications in Healthcare

The applications of AI now pervade nearly every aspect of medicine. A non-exhaustive list of applications include:

1. Data analysis of scientific literature
2. Mining of electronic health records (EHR)
3. Diagnostics and Screening
4. Therapeutics
5. Drug Discovery and Development
6. Clinical Care
7. Epidemiology and Prevention of Disease
8. Psychiatric Healthcare ⁶
9. Diagnosis of mood disorders such as depression
10. Health Management Systems using AI
11. Facial recognition technologies
12. Detect genetic disorders that correspond to specific facial traits ⁷
13. Patient identification
14. Eldercare (i.e. nursing carebots)

Potential Benefits

By understanding how to appropriately translate new technologies into practice, AI has the potential to transform the healthcare industry in ways that would otherwise be unattainable with human intelligence alone. Many of these possibilities are largely due to the processing power of AI using big data. The potential benefits broadly include:

1. Improved diagnosis and treatment recommendations
2. Better patient engagement and adherence, and
3. More efficient administrative activities
4. Improvements in quality and effectiveness of clinical services
5. Advancements in personalised medicine
6. More efficient randomized control trials (RCT) ⁸

Examples of how big data and AI have already begun revolutionizing the healthcare industry include the National Institutes of Health 1000 Genomes Project in the US, the partnership between DeepMind and Moorfields Eye Hospital NHS Foundation Trust performing eye scans with machine learning in the UK, and the various applications afforded by electronic health records (EHR) globally.⁹

⁶ Rebecca A. Bernert et al., “Artificial Intelligence and Suicide Prevention: A Systematic Review of Machine Learning Investigations,” *International Journal of Environmental Research and Public Health* 17, no. 16 (August 1, 2020): 5929, <https://doi.org/10.3390/ijerph17165929>.

⁷ Denys Fontaine et al., “Artificial Intelligence to Evaluate Postoperative Pain Based on Facial Expression Recognition,” *European Journal of Pain* (London, England) 26, no. 6 (July 1, 2022): 1282–91, <https://doi.org/10.1002/ejp.1948>.

⁸ Thomas Davenport and Ravi Kalakota, “The Potential for Artificial Intelligence in Healthcare,” *Future Healthcare Journal* 6, no. 2 (June 2019): 94–98, <https://doi.org/10.7861/futurehosp.6-2-94>.

⁹ “DeepMind Health Q&A | Moorfields Eye Hospital NHS Foundation Trust,” Moorfields.nhs.uk, 2018, <https://www.moorfields.nhs.uk/faq/deepmind-health-qa>.

Risks and Challenges

Alongside the promising benefits of AI come significant risks and challenges. Within AI in healthcare, the main ethical concerns broadly include,

1. Transparency/opacity
2. Risk/safety
3. Bias
4. Accountability or responsibility
5. Trustworthiness

This is complicated by the fact that the technology is relatively new and rapidly evolving, leaving gaps in the guidelines and frameworks from both the public and private sectors. The Nuffield Foundation has identified three main gaps in the current literature: “(i) No agreement around key ethical concepts and their application; (ii) lack of attention on conflicts between ideals and values; and (iii) lack of evidence of AI’s capabilities and the way it is perceived by the public”.¹⁰

The Challenge of the AI Chasm

The current document will centre around the ethical challenges surrounding what is known as the “AI chasm”. The AI chasm “describes the current gap between the development of a robust algorithm and its clinically meaningful application”.¹¹ Despite the promise of AI, only a small fraction of developed models are successfully implemented at the point of care.¹² It has been noted that circumventing the AI chasm could reduce arbitrary variation, minimize medical errors, and provide information to support clinical decision-making.¹³

Among the various ethical concerns that give rise to the AI chasm, mitigating bias and risk, and issues surrounding the human-AI relationship are particularly challenging. The norms and status quo that has developed around these issues have thus motivated the scope of this document. As the local and international community continue to develop guidelines and frameworks on how to achieve a clear and safe path forward with AI, it can be argued that these efforts have conjured fears and motivated norms erring on the side of caution when it comes to the design, development and implementation of AI.^{14,15,16}

Considering the rapid advancement of these technologies and the potential for major impact (both positive and negative), a cautious approach is required. Nonetheless, we will explore how and when certain norms can be challenged so that we may take full advantage of the potential of AI in healthcare.

¹⁰ Whittlestone, J. Nyrup, R. Alexandrova, A. et al. “A Road Map for Research”

¹¹ Melissa D McCradden et al., “A Research Ethics Framework for the Clinical Translation of Healthcare Machine Learning,” *The American Journal of Bioethics* 22, no. 5 (January 20, 2022): 1–15, <https://doi.org/10.1080/15265161.2021.2013977>.

¹² Zuzanna Angehrn et al., “Artificial Intelligence and Machine Learning Applied at the Point of Care,” *Frontiers in Pharmacology* 11, no. 759 (June 18, 2020), <https://doi.org/10.3389/fphar.2020.00759>.

¹³ Melissa D McCradden et. al. “A Research Ethics Framework”

¹⁴ Birgit et al., “Intelligent Decision Support in Medical Triage: Are People Robust to Biased Advice?,” *IEEE Intelligent Systems* 34, no. 2 (March 20, 2023), <https://doi.org/10.1093/pubmed/fdad005>.

¹⁵ Jurriaan van Diggelen et al., “Developing Effective and Resilient Human-Agent Teamwork Using Team Design Patterns,” *IEEE Intelligent Systems* 34, no. 2 (March 2019): 15–24, <https://doi.org/10.1109/mis.2018.2886671>.

¹⁶ Filippo Santoni de Sio and Jeroen van den Hoven, “Meaningful Human Control over Autonomous Systems: A Philosophical Account,” *Frontiers in Robotics and AI* 15, no. 5 (February 28, 2018), <https://doi.org/10.3389/frobt.2018.00015>

1.4 Methodology

This document takes a bottom-up approach to analysing specific ethical concerns surrounding bias, human involvement, and risk prediction in the translational application of AI in biomedical research. Each section begins with an analysis of a local Singaporean case study, revealing the commonly occurring ethical tradeoffs that need to be evaluated in making decisions about the use of AI in research. Then the principles that can be drawn upon in resolving each ethical problem are discussed and strategic guidance is offered.

Each section draws from ongoing work from the SHAPES team. Section 2 on bias draws on a paper authored by Kathryn Muyskens, Harisan Nasir, Angela Ballantyne, Murali and Julian Savulescu titled “The Permissibility of Biased AI in a Biased World: An Ethical Analysis of AI for Screening and Referrals for Diabetic Retinopathy in Singapore”. It is currently under review. Section 3 on human involvement draws from a forthcoming paper by Kathryn Muyskens, Yonghui Ma, Jerry Menikoff, James Hallinan, Julian Savulescu titled “When Can We Kick Humans ‘Out of the Loop’ - An Examination of the Use of AI in Medical Imaging for Lumbar Spinal Stenosis”. Section 4 will be developed into an academic paper as a next step with Alexa Nord-Bronzyk as the main author, Michael Dunn as the senior author, and future co-authors to be announced.

While many frameworks and guidelines offer a top-down approach to the present ethical concerns by beginning with various principles to uphold throughout ethical analysis (see Annex 1), the present document instead hopes to illuminate the complexity and nuance of our case studies as a means to pull out the salient principles at work in resolving each ethical consideration. As opposed to a principles-first approach, a bottom-up approach accounts for the unique context of each case study instead of making generalizations that might fail to accurately or adequately capture the full picture and particularities. The research community may apply this approach similarly by considering each case as unique, starting with identifying the problem to be solved and then working towards the desired outcome.

The focus on bias, human involvement, and risk prediction is motivated by the problem of the AI chasm as norms that have developed in dealing with these themes may prevent useful tools from reaching the point of care. These issues will be discussed in relation to how they present difficulties toward implementation in ways that are consistent with ethically defensible practice. In not addressing these issues, therefore, there is a risk of acting unethically or wrongly when translating AI into practice. Each section finishes by offering guidance on how to think critically and carefully about addressing relevant ethical considerations and making appropriate ethical tradeoffs.

The first section on bias explores how to ethically consider if implementing a biased AI tool is ever justifiable by evaluating the tradeoffs between utility and equity. In the subsequent section, how to appropriately understand the human-AI relationship is examined by narrowing in on the ethical challenges of “kicking humans out of the loop” (i.e. replacing a previously human task with automation). The final section discusses the risks involved with risk prediction AI tools, especially in emergency medicine, and analyses how much risk is reasonable in an already risky world. From there, the principles that apply in resolving each ethical consideration are discussed. In line with the narrow scope and focus on local context, this document endorses the definitions of principles outlined by the Bioethics Advisory Committee’s consultation paper “Ethical, Legal and Social Issues Arising from Big Data and Artificial Intelligence Use in Human Biomedical Research” where possible, and the Oxford Handbook of

Ethics of AI¹⁷ secondarily.¹⁸ Each section finishes with strategic measures guiding readers on how to handle these issues in practice.

In summary, each section highlights an ethical tradeoff to be made, how to analyse that tradeoff carefully and transparently, and how to proceed accordingly. In some cases, this will allow AI that manifests social biases, allow AI that kicks humans out of certain loops, and allow AI that is potentially beneficial but still carries unknown risks to be introduced.

Notwithstanding this tradeoff analysis, and irrespective of whether the threshold is met for further use/research, there is an ethical obligation to mitigate the different harms (or risks of harms) or wrongs that have been drawn attention to in each of the three sections. That means taking appropriate steps to i) address and minimise social biases in the implementation process, ii) continually evaluate the impact of dehumanising approaches and mitigating the potential effect of dehumanising healthcare relationships, and iii) mitigating the risks appropriately in order to ensure an optimal balance between risks and benefits is realised.

Since the focus of this document is narrow in scope, there will be certain ethical considerations that are not covered.¹⁹ The aim is to address the selected themes as they appear in the local context by concentrating on the relevant ethical issues driven by the case studies without broadening to all possible ethical considerations. Wider discussions analysing these topics in more detail can be found in various other comprehensive documents including those from the Bioethics Advisory Committee, World Health Organisation, and Ministry of Health.^{20,21,22}

¹⁷ Definitions by Alessandro Blasimme and Effy Vayena in 'The Ethics of AI in Biomedical Research, Patient Care, and Public Health

¹⁸ See Annex 1 for full definitions of principles

¹⁹ For example, issues surrounding consent in the final section on risk prediction will be left out as the ethically relevant considerations are limited by regulatory constraints in Singapore.

²⁰ Bioethics Advisory Committee, "Ethical, Legal and Social Issues Arising from Big Data and Artificial Intelligence Use in Human Biomedical Research: A Consultation Paper". May 2023. <https://www.bioethics-singapore.gov.sg/files/publications/consultation-papers/big-data-and-ai.pdf>

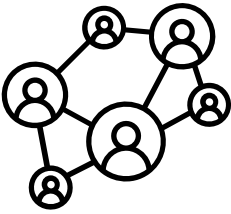
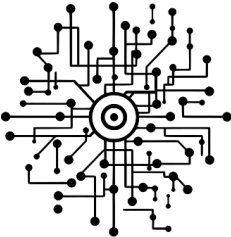
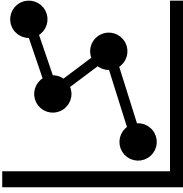
²¹ WHO tool for benchmarking ethics oversight of health-related research involving human participants. (2023) Geneva: World Health Organization. Licence: CC BY-NC-SA 3.0 IGO.

²² "MOH Artificial Intelligence in Healthcare Guidelines (AIHGle)," (October 2021), <https://www.go.gov.sg/aihgle>.

2. Bias— Is it ever justifiable to implement biased AI?²³

2.1 What is bias in AI?

Bias presents in societies as a significant ethical issue that threatens the realisation of social justice. In the translation of AI into health practice, the concern is that pre-existing social bias manifests algorithmically and statistically. Figure One below elaborates on these three forms of bias in AI.

<p style="text-align: center;">Social</p> 	<p>Differential access to the underlying social determinants of health (e.g. experiences of poverty, lack of education and living conditions); differential access to healthcare (e.g. distance to hospital or insurance status); or patterns of discrimination (such as racism and sexism). Hence, even when the data used to train the model may be representative and accurate, it may still capture and reflect objectionable aspects of the real world such as stigma, discrimination or oppression.</p>
<p style="text-align: center;">Algorithmic</p> 	<p>Prediction or outputs of a model unfairly and unjustifiably benefit or disadvantage certain individuals or groups.</p>
<p style="text-align: center;">Statistical</p> 	<p>Training data unrepresentative of the target patient population. For example, unconscious racial bias that leads to the over-policing of people of colour in the United States can generate statistical bias in arrest data if people of colour are more likely to be arrested and if arrests are considered a proxy measure of actual crime.</p>
<p>Figure One. Definitions of bias in AI.</p>	

Bias in health AI in particular gives rise to specific ethical concerns regarding justice in health care practice. The case study below describes a diagnostic tool for the screening of diabetic

²³ The case study and ideas presented in this sections 2.1-2.4 reference the following paper: Kathryn Muyskens, Harisan Nasir, Angela Ballantyne, Murali and Julian Savulescu. Under Review. “The Permissibility of Biased AI in a Biased World: An Ethical Analysis of AI for Screening and Referrals for Diabetic Retinopathy in Singapore”.

retinopathy developed by a team of researchers in Singapore that was found to be biased. The study illustrates the ethical tension between utility and equity and analyses when a certain degree of bias or injustice may be reasonable to accept in its translational application. From there, other issues arising from bias in health AI are discussed and guidance is offered for mitigating these issues.

2.2 Case Study: The Permissibility of Biased AI in a Biased World: An Ethical Analysis of AI for Screening and Referrals for Diabetic Retinopathy in Singapore

In 2017, a team from Duke-NUS led by Ting and colleagues developed a deep learning system to detect the progression of diabetic retinopathy and glaucoma (hereafter referred to as DLSDR for deep learning system for diabetic retinopathy). When a certain level of disease progression is reached, the algorithm flags the result and radiologists can decide whether to refer the patient to a specialist.

When sensitivity was set at comparable levels to trained human graders (90%), specificity was lower for AI (90%) than for trained graders (99%). When sensitivity was maximized at the cost of lower specificity, a significant difference in sensitivities between Malay (97.1%) and non-Malay patients (100% and 99.3% for Chinese and Indians respectively) was detected.

“Sensitivity” here refers to how many positive cases are detected, and “specificity” refers to how often the tool accurately identifies a patient as having a negative result. Sensitivity refers to the chance that someone with a disease tests positive when using the tool. Likewise specificity refers to the chance that someone without the disease tests negative.

In an *unbiased* system, we would expect that rates of referable diabetic retinopathy to roughly track the proportion of diabetics who are Malay (>20%) weighted by the average degree of disease control. Despite this, they are still referred to specialists less often than other groups. Only 7.3% of those referred to specialist care are Malay (Sia et al 2020)⁴. This would seem to indicate some background bias or other prior unjust disparities present in the social environment in which the DLSDR system would be deployed. We must ask, what explains this disparity? It is not within the scope of this paper to determine the answer, but possible explanations include personal (conscious or unconscious) bias on the part of the physicians or graders; disparities in patient participation in screening, distrust of the medical establishment, workplace discrimination, poverty, religious fatalism, lack of education, cultural norms or personal choice.

Implementing the DLSDR into this social context would likely result in more referrals to specialists across groups when compared with unaided clinical judgment due to the boost in efficiency that the system provides. However, it could also widen the existing disparity between Malays and non-Malays. This would be because while the sensitivity in detecting diabetic retinopathy improves for all groups, the sensitivity improves for non-Malays more than Malays. Meanwhile, most if not all of the causes of the disparity remain untouched by the implementation of the DLSDR. The question of equity versus utility arises as we are now tasked with analysing the ensuing trade-offs of implementing or not implementing the DLSDR.

2.3 Why not just remove the bias?

Due to the various ways biases are programmed into AI, such as how a problem is framed, data processing, and inappropriate deployment, it can be difficult to identify the source.²⁴ Because biases occurring in the social context cannot be removed, said biases will inevitably infiltrate any AI tool.

However, there have been efforts to improve algorithmic biases which have proven successful. For example, bias was identified in a surgical AI system (SAIS) used to assess the skill level of surgeons across various activities such as needle handling and needle driving.²⁵ The system demonstrated bias by either erroneously downgrading or erroneously upgrading skill performance. An add-on application called TWIX was added to the prediction model that mitigated the bias by improving model performance for the erroneous gradings. But this kind of strategy takes time and can risk introducing additional biases.²⁶ Even where the source(s) of bias are identifiable, the amount of time and collaboration among researchers, developers, and clinicians needed to determine the origins should be considered.²⁷

In light of these complexities, it would seem there is no straightforward option to remove biases from AI entirely. The ethical question becomes about how much bias we are willing to accept in light of the utility of a tool, especially in consideration of the degree of impact the bias has in health care practice.

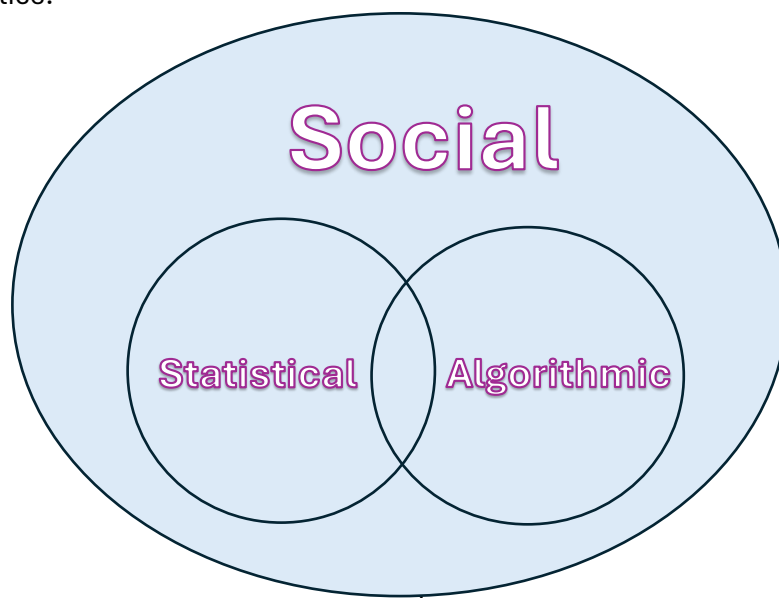


Figure Two. The diagram illustrates the inextricable ties among the various ways bias can occur in AI by highlighting how social biases will inevitably always interact and influence the manifestation of algorithmic and statistical biases. Refer to the definitions in Figure One for clarification.

²⁴ David Leslie et al., "Does 'AI' Stand for Augmenting Inequality in the Era of Covid-19 Healthcare?," *BMJ* 372, no. 372 (March 16, 2021): n304, <https://doi.org/10.1136/bmj.n304>.

²⁵ Dani Kiyasseh et al., "Human Visual Explanations Mitigate Bias in AI-Based Assessment of Surgeon Skills," *Npj Digital Medicine* 6, no. 1 (March 30, 2023): 1-12, <https://doi.org/10.1038/s41746-023-00766-2>.

²⁶ Mirja Mittermaier, Mariam M. Raza, and Joseph C. Kvedar, "Bias in AI-Based Models for Medical Applications: Challenges and Mitigation Strategies," *Npj Digital Medicine* 6, no. 1 (June 14, 2023): 1-3, <https://doi.org/10.1038/s41746-023-00858-z>.

²⁷ Another strategy to mitigate the influence of a biased system is to flag the bias to the clinicians so that they are aware. This would make using the tool more transparent.

Where there are biases happening in the background in various ways, the question of the ethical justification of using such a biased AI tool depends on how it interacts with the social context, particularly whether or not the tool will accentuate pre-existing concerns about injustice in health care upon its introduction. In our case, and likely almost all social contexts, we know to be characterised by bias and inequality.²⁸

2.4 How should utility and equity be considered in evaluating the implementation of biased AI?

Since at present we cannot completely remove biases from AI tools nor the social context, we must consider when using a biased AI system can be justified. More specifically, is worsening an existing disparity for the sake of utility ever defensible?

In our case, Malays will still be better off on average with the DLSDR tool than without it and gain real benefits in wellbeing, as well as wider social benefits, through the prevention of job loss and disability. By not implementing the DLSDR, the current level of inequity is maintained, and the potential utility that could have been gained across all groups is lost. In this case, placing concerns of equity higher than utility seems questionable.

Therefore, it would seem that it is justifiable to implement a biased AI tool where,

- | |
|-----------------------------------------------------------------------------|
| 1. Its introduction reduces the influence of biases, and/or |
| 2. Where the utility gained is significant enough and shared across groups. |

2.5 What principles apply in deciding when implementing a biased AI is justified?

The principle of **proportionality** can be drawn upon to help in the weighing between equity and utility in the case of implementing the DLSDR. Proportionality “requires that the methods or processes used in biomedical research are necessary and appropriate in relation to the research intent and the range of public and private interests at stake. Proportionality thus achieves a balanced relationship between the risks and benefits when incommensurable values compete. In evaluating the tradeoffs between prioritising equity over utility, it may be justifiable in certain cases justifiable to implement a biased system based on the above criteria which aims to fulfil the principle of proportionality.

The second principle at play is **justice**. Justice is primarily about fair distribution of benefits and burdens, and “justice in the context of big data and AI biomedical research requires that researchers manage and use data in a manner that does not create or reinforce bias”.²⁹ Justice is achieved when there impartial and just treatment toward all groups and prejudice and discrimination is avoided. Justice can be employed to ensure that despite widening inequities caused by the DLSDR, the overall utility would be such that justice would still be served to

²⁸ Health and Racial Discrimination: Submission by the Community Action Network (Singapore) to the UN Committee on the Elimination of Racial Discrimination on Singapore’s Compliance with the International Convention on the Elimination of All Forms of Racial Discrimination, focussing on health and racial discrimination.
https://tbinternet.ohchr.org/_layouts/15/TreatyBodyExternal/DownloadDraft.aspx?key=ICEnwWR8rbeJM801ALabP3E pMVzXUy1JQWquKqoRYVcPw16C5yhh4LWTicblXC38ZGsmA5SQWqq1qcpXX7w8w==

²⁹ Bioethics Advisory Committee, “Consultation Paper”, p.26.

Malays and other affected groups as the utility would benefit all. Other factors that may motivate the use of a biased system could be cost savings or allocation of resources.

We can imagine a scenario where the utility gained from the use of an AI tool might be as significant as in the case of the DLSDR, but the resulting inequity is such that it does not align with our threshold for proportionality. For example, if a tool promised an extreme savings in cost for certain groups by reducing the workload for clinicians, but this savings was not shared across groups, it may not be justifiable to employ the tool.

Inclusiveness and **solidarity** are also at the heart of avoiding such biases. Inclusiveness aims to respect diversity and fairly represent affected groups by stressing “the need to include all affected parties in deliberations and decision-making practices about the use of data and algorithms”.³⁰ While we are not aware of exactly how the complexities of the background bias operate, the bias in DLSDR can be guided by inclusiveness to help in justifying the criteria for using a biased system. Solidarity implies sharing both prosperity and burden to ensure nobody is left behind in the immediate and long-term.³¹ Solidarity can be realised by requiring utility to be shared across groups.

2.6 Strategic Measures to Mitigate Bias in Practice – What can you do?

While the weighing analysis above justifies the use of a system that manifests pre-existing biases, there is still an overarching obligation to mitigate bias impacting practice in the translation project. The case study presented above raises just one of many ethical concerns that biases in AI can introduce. Below are strategic measures that can help to guide the research community in mitigating bias in practical ways.

Appropriate Recruitment

Appropriate recruitment should focus on the inclusivity of the research participants as the demographic influences the potential for bias. Inclusiveness is predicated on the awareness of possible underrepresentation and overrepresentation of any certain group. Diversity in AI training datasets can help to appropriately represent affected groups. This should begin with recruitment. The body of research participants should appropriately reflect those most likely to be affected by the study.

Awareness of Audit Mechanisms

Appropriate audit mechanisms can enable researchers and IRB members to employ experts to evaluate where AI bias may exist or arise. These mechanisms can include ensuring diversity in datasets, appropriate patient selection, and auditability of systems. Where technical knowledge is required to unearth such biases, scientific review may be necessary. Nonetheless, the research community should be aware of such issues and be able to respond appropriately and offer meaningful insight where necessary.

Who is responsible for identifying biases is outside the scope of this document. These considerations will be defined by international and local guidelines as they align with other legal frameworks such as the Ministry of Health (MOH) Artificial Intelligence in Healthcare Guidelines (AIHGle), and Human Biomedical Research Act (HBRA), and Health Information Act.

³⁰ Alessandro Blasimme and Effy Vayena, 'The Ethics of AI in Biomedical Research, Patient Care, and Public Health', in Markus D. Dubber, Frank Pasquale, and Sunit Das (eds), *The Oxford Handbook of Ethics of AI* (2020; online edn, Oxford Academic, 9 July 2020), <https://doi-org.libproxy1.nus.edu.sg/10.1093/oxfordhb/9780190067397.013.45>

³¹ Miguel Luengo-Oroz, "Solidarity Should Be a Core Ethical Principle of AI," *Nature Machine Intelligence* 1, no. 11 (October 18, 2019): 494–94, <https://doi.org/10.1038/s42256-019-0115-3>.

Additionally, audit mechanisms are ever-changing. Keeping pace with the updates will help to ensure appropriate safeguards are current.

Public Engagement by Research Community

Public engagement can also be used to mitigate bias by involving the communities closest to the study by collaborating with patients and the public, sharing decision-making, raising awareness, and sharing research knowledge and findings.³² The goal of shared-decision making and public engagement is not to achieve unanimity, which would be unrealistic given the complex ethical tradeoffs involved; but rather to achieve a shared understanding of the nature of the problem, to consider diverse perspective and to be transparent about reasons and justifications.

As the Nuffield Foundation rightly points out, “negotiating tradeoffs between values can only happen when these values and the related hopes and concerns of everyone who is going to be impacted by these technologies are identified and considered”.³³ Through public deliberation, polling, dialogues, surveys, and interviews, the research community can align research goals with public concerns.³⁴ Advisory forums can also be held to discuss the conduct of research, and how the social value may legitimize the research.

In-house and project-specific designated committees can be useful in narrowing in on the specific ethical challenges of each unique study. The committee should be comprised of a diverse group including members of the public. Committee goals, members, and structure should be decided during the design stage of any study prior to IRB review.

Trial Registration

Trial registration can serve as another means to mitigate bias by promoting accountability and clarity in the study design from the outset. Studies have shown that trial registration can also be used to enhance quality and transparency, thus leading to safeguards against outcome reporting bias and spin. Nonetheless, researchers and IRB members should remain privy to the ways bias and spin may manipulate research outcomes.³⁵

³² “Patient and Public Involvement, Engagement and Participation Definitions,”

<https://www.medsci.ox.ac.uk/research/patient-and-public-involvement/section-2-what-is-patient-and-public-involvement>

³³ Whittlestone, J. Nyrup, R. Alexandrova, A. et al. “A Road Map for Research”




³⁴ Although it falls out of the scope of the current document, AI literacy would also be an essential component of public engagement.

³⁵ Jiyeon Won et al., “Trial Registration as a Safeguard against Outcome Reporting Bias and Spin? A Case Study of Randomized Controlled Trials of Acupuncture,” ed. Spyridon N. Papageorgiou, PLOS ONE 14, no. 10 (October 3, 2019): e0223305, <https://doi.org/10.1371/journal.pone.0223305>.

3. Human Involvement – How and to what extent?³⁶

3.1 What does human involvement in AI mean?

As advancements in AI continue to challenge human intelligence, concerns around how and to what extent human involvement is necessary for optimal use of these tool become problematic. Human involvement refers to the human-AI relationship and the level of interaction, agency, and/or oversight one employs in relation to how much the tool is doing autonomously. The European Commission categorizes human involvement in three main ways.³⁷

<p>Human-in-the-loop</p> 	<p>Refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable.</p>
<p>Human-out-of-the-loop</p> 	<p>Refers to the capability for human intervention during the design cycle of the system and monitoring system’s operation.</p>
<p>Human-in-command</p> 	<p>Refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation. This can include the decision not to use an AI system in a particular situation, to establish levels of human discretion during the use of the system, or to ensure the ability to override a decision made by a system.</p>
<p>Figure Three. Human-AI relationship terminology.</p>	

³⁶ The case study and ideas presented in this sections 3.1–3.3 reference the following forthcoming paper: Kathryn Muyskens, Yonghui Ma, Jerry Menikoff, James Hallinan, Julian Savulescu. Forthcoming. “When Can We Kick Humans “Out of the Loop”. Asian Bioethics Review.

³⁷ European Commission, “Directorate–General for Communications Networks, Content and Technology, Ethics guidelines for trustworthy AI”, Publications Office., <https://data.europa.eu/doi/10.2759/346720>

As tools get smarter, mounting fears spread about AI replacing humans in various ways which has contributed to the human-in-the-loop model becoming the norm in some circles. Whether as a strategy to soften fears or as legitimate guidance, many guardrails and frameworks across continents have also generalized good ethical practice as keeping humans in the loop.^{38,39} However, as AI advances, we may wish to question if there are situations when kicking humans out of the loop might better serve the desired outcome. This question is explored by taking a closer look at the application of machine learning for the detection of Lumbar Spinal Stenosis (LSS) developed by a team of researchers in Singapore.

Case Study: SPINE AI: MEDICAL IMAGING FOR LUMBAR SPINAL STENOSIS

Developed from: Hallinan JTPD et al. (2021)⁴⁰

A team of researchers at NUH/NUS medical school in Singapore has developed an AI model based on convolutional neural networks for automated detection and classification of lumbar spinal canal, lateral recess, and neural foraminal narrowing in an MRI scan of the spine to diagnose lumbar spinal stenosis (LSS). LSS is a potentially debilitating condition affecting many adults globally, with a considerable impact on livelihood. Most patients with LSS present with lower back pain, which is also the main reason for seeking care. A large proportion of patients eventually undergo lumbar spine MRI for diagnosis and treatment planning. Lumbar spine MRI is an essential tool in the assessment of LSS for the accurate evaluation of the central canal, lateral recesses, and neural foramina. The degree of stenosis at each region plays a role in determining the appropriate treatment, but detailing such information in a report can be repetitive and time-consuming. In addition, there are multiple grading systems for LSS, with a lack of standardization.

The research team was able to show that the Spine AI model for semi-automated reporting of lumbar spine MRI scans could produce the following benefits:

More consistent and accurate grading of spinal stenosis: This can improve clinical decision-making and patient outcomes (Hallinan et al 2021), and where institutions lack radiologist expertise, the model can improve the accuracy of inexperienced readers, e.g. kappas for trainee and general radiologists increased from 0.6 to 0.9 with the model, matching the performance of a specialist spine radiologist (kappa=0.9) (Hallinan et al 2021, Lim et al 2022).

Improved scan turnaround time and radiologist productivity (which in turn reduced cost): The deep learning (DL) solution will also reduce the time taken for report generation. Based on the recent Radiology manuscript (2022) reporting time could be reduced by ~70% with, compared to without DL model assistance (e.g. 10 minutes to 3 minutes with the DL model, 7 min time saved) (Lim et al 2022). With ~67,000 MRI lumbar spines a year performed in Singapore (~10 hospitals), a saving of 7 minutes per MRI results in ~469,000 minutes (7,817 hours) saved per year in Singapore alone. The per-hour rate for radiologists is \$100 SGD, meaning there is a potential cost savings of up to \$780,000 SGD each year.

³⁸ Mittermaier, M., Raza, M.M. & Kvedar, J.C. "Bias in AI-based models for medical applications".

³⁹ Birgit van der Stigchel, Karel van den Bosch, Jurriaan van Diggelen, Pim Haselager. "Intelligent decision support in medical triage".

⁴⁰ Hallinan JTPD, Zhu L, Yang K, Makmur A, Algazwi DAR, Thian YL, Lau S, Choo YS, Eide SE, Yap QV, Chan YH, Tan JH, Kumar N, Ooi BC, Yoshioka H, Quek ST. "Deep Learning Model for Automated Detection and Classification of Central Canal, Lateral Recess, and Neural Foraminal Stenosis at Lumbar Spine MRI." *Radiology*. 2021 Jul;300(1):130-138. doi: 10.1148/radiol.2021204289. Epub 2021 May 11. PMID: 33973835.

This version of assisted AI solution for MRI spine reporting has the radiologist at the center of the process (e.g. it was not fully automated). The AI model outputs are provided as boxes overlaid on the MRI images. These can be changed as necessary by the reporting radiologist, and once they have reviewed all the outputs a text report can be automatically generated.

The researchers at NUH/NUS recommend that the AI assisted solution for MRI spine reporting have the radiologist at the center of the process, saying “a fully automated system is unlikely to be acceptable to either radiologists or clinicians” (Hallinan et al 2021). AI models of the type deployed in this instance can assist with reading images and identifying patterns in the data at a faster rate than humans alone. However, the limitations of current AI models mean that the human radiologist cannot be removed from the process without raising some serious ethical concerns, primarily, concerning issues of safety, reliability and accountability.

The authors of the study concluded, “...our deep learning (DL) model is reliable and may be used to quickly assess lumbar spinal stenosis (LSS) at MRI. In clinical practice, the diagnosis of LSS still relies on the subjective opinion of the reporting radiologist. Our DL model could provide semi-automated reporting under the supervision of a radiologist to provide more consistent and objective reporting. Further development of the DL model could involve a consensus panel of international experts to reduce any labelling errors and biases. The DL model could also be assessed for the longitudinal follow-up of LSS at MRI” (Hallinan et al 2021, 137).

3.2 Leaving Humans Out of the Loop – What are the main ethical concerns?

At first glance, Spine AI does not appear to raise many of the common ethical concerns associated with AI. First, the training data was racially heterogenous. This would suggest it is reasonable to trust the validity of its conclusions. Second, the tool’s transparency allows clinicians to easily check its conclusions by displaying 1) a reading of whether or not LSS is present and 2) highlighting the region on the scan used to make the judgment. A system that fails to be transparent or interpretable is known as a “black box”. While Spine AI’s exact grading system is indeed a “black box”, clinicians had no reservations implementing the system because they could double-check its marking. Because Spine AI also does not store data or directly interact with the patient, it is less risky than tools that do.

In terms of accountability and trustworthiness, the concerns associated with the fully automated use of AI are varied. These include 1) that the technology may not have yet proven itself sufficiently adept; 2) given these limitations, the combination of radiologists and AI could be the most cost-effective; 3) corrosive effects on trust in the clinical context; 4) the dehumanization of medicine; 5) over-reliance on AI increasing liability in medical malpractice and raising questions about accountability. These issues are addressed in the following sections.

Adept Technology

The technical performance of the AI in question should always be properly evaluated. In the case of Spine AI, the main function of pattern recognition has proven adept. Adept in the case of AI means that the tool has proven proficient and achieves the desired goal. Additionally, there are pre- and post-scan consultations with a clinician, and because eventually a surgeon would still need to make the case for moving forward with surgery, discharging radiologists may be not only advantageous but also cost-effective in this case.

The Dehumanisation of Medicine



Concerns around dehumanising medicine can carry negative connotations, but here it is important to clarify that “dehumanising” only refers to the removal of a person from the process and not removing the “humanness” from the process. In the case of Spine AI, the person removed (the radiologist) is someone the patient would ordinarily never meet. Thus, while there may be a loss of interaction between the doctor and radiologist, this does not equate with a dehumanisation in the relationship between patients and those caring for them.

Removing the human from the loop focuses on replacing a human with AI for a certain task within a process. The process will still involve humans who can monitor the task and contribute to the success of the patient pathway in other ways. This keeps concerns of trustworthiness focused on the doctor-patient relationship, and not the patient-AI relationship.⁴¹

Unlike other AI tools that may indeed reduce necessary human involvement where it is an invaluable component of the patient pathway, Spine AI is unlikely to achieve this in any meaningful way. If Spine AI were to advise on the case for surgery or influence the care pathway, then we may be more hesitant to kick the human out of the loop. The important question the research community should be concerned with is whether the dehumanisation of medicine is ethically meaningful in the case of the specific tool in question. Where kicking humans out the loop is suggested, the implications to care as well as the humans that are removed should be the focus of ethical discussion.

3.3 What are the criteria for putting humans out of the loop? - Benefits vs. Costs

Ultimately evaluating whether it is justified to put humans out of the loop in the case of Spine AI comes down to evaluating the tradeoffs. This is a concern about how the AI tool replacing the human involvement will operate, in terms of its costs and benefits as highlighted in Figure Four below.

<p>Benefits</p> 	<ul style="list-style-type: none">• Significant savings in cost and time• Less manpower required• Improved accuracy of inexperienced trainees• Decreased turnaround time for image reading
<p>Costs</p> 	<ul style="list-style-type: none">• Overall risk to patient is low as it is an image-reading tool and LSS is not life threatening• Radiologists “kicked out of the loop”
<p>Figure Four. The costs and benefits associated with implementing Spine AI.</p>	

⁴¹ This also raises ethical considerations surrounding the patient perspective. Should the patient have the right to know if his or her report was read by Spine AI or a radiologist? Should a patient have the right to refuse Spine AI replacing radiologist in their medical care? These are important considerations when reviewing the implementation of any new tool, but due to the limitations in the scope of this document we will not analyze them here.

So when is it ok to kick humans out of the loop?

1. The technology is as effective (or better) than a human at the given task (e.g. error rates are equal to or lower than human experts).
2. The risk to patients (or any humans involved) is low in the event of an error.
3. The wellbeing that is gained by the speed, accuracy, and cost-efficiency of automation is high.

According to this criteria, Spine AI is a good example of when kick humans out of the loop would be ethically justifiable: the putative benefits outweigh the expected risks.⁴²

3.4 What principles apply in deciding the role of human involvement?

The principle of **proportionality** can be employed in evaluating the use of Spine AI by weighing the tradeoffs between benefits and costs of implementation. The significant benefits of using Spine AI serves in justifying the tradeoff of kicking radiologists out of the loop in this particular task and in line with the above criteria.

Interpretability and **transparency** help in defending a “human out of the loop” model where the more interpretable an AI tool is, the ethical considerations that follow can be illuminated more clearly than in a “black box” model. The interpretability of Spine AI meant there was little pushback from users regarding its implementation as the ethical issues could be more easily understood. When thinking about transparency, aim to calibrate how transparent a system is to the context and impact it may have where other considerations are at play such as data protection, safety, and security.

Closely related to transparency is the principle of **accountability**. Accountability can be drawn upon in evaluating at what stage Spine AI would contribute to the diagnosis and patient care pathway. As it is just an image reading tool, it would be the surgeon at the end who is accountable for making a case for a patient to undergo surgery, thus lowering the overall risk of the tool.

3.5 Strategic Measures in Practice to Determine Appropriate Human Involvement – What can you do?

While the weighing analysis above justifies putting humans out of the loop in this instance, there is still an overarching obligation to ensure appropriate human involvement, especially in consideration of any potential ripple effects. Below are strategic measures that can help to guide the research community in analysing the appropriate human-AI relationship in practical ways.

Measure Baseline Performance

When evaluating the utility of any AI adoption, baseline performance should be understood to measure if the AI is indeed improving overall performance.

It is not always the case that the addition of AI will streamline performance or improve workflows. For example, a study involving radiologists varied the availability of AI assistance

⁴² This is especially the case in low resource settings where medical professionals may be in short supply. In extreme situations, this could mean that some scans may never get read at all, thus making the case for Spine AI even stronger.

and contextual information to study the effectiveness of human-AI collaboration and reported that “humans—when supplemented by AI—did not perform any better than their counterparts”.⁴³

Measuring baseline performance helps to give perspective to the appropriate human-AI involvement by understanding where and how AI is contributing to a task or process and how to properly manage expectations.

Is the AI tool a decision-making tool?

An AI system that makes decisions tends to pose an overall greater level of risk as their predictions can influence the patient care pathway meaning their conclusions hold more weight.⁴⁴ Image-reading tools such as Spine AI serve to supplement the important decisions that contribute to diagnosis and treatment, but not make any ultimate decision where harm to the patient is a concern.

AI’s lack of emotional capacity should also be considered if the system is involved in decisions where human factors such as empathy may be especially important.

Have a Plan to Ensure Accountability

Regardless of how intelligent an AI system is, all tools are prone to fallibility. Should the AI arrive at incorrect conclusions or make wrong decisions, including a plan addressing accountability can help to ensure the appropriate actor is held responsible. This can be especially complicated due to the fluctuations of involvement among various actors involved in a research study making attribution of responsibility uncertain, otherwise known as “diffusion of responsibility”.⁴⁵

Cross-functional communication among all involved in a research study including but not limited to AI researchers, biomedical researchers, developers, and clinicians should be maintained so that appropriate accountability is upheld. For example, if the source of error was due to erroneous code, then it would most likely be the responsibility of the AI algorithm researchers. A plan scoping potential errors and the associated responsible actor should be agreed upon during the design stage of the study.

Be Realistic about the Prospect of Automation

Due to the nature of the work done by radiologists, it is plausible to imagine many of their everyday tasks becoming automated. These tasks may include (1) automated image segmentation, lesion detection, measurement, labelling, and comparisons with historical images; (2) generating radiology reports, particularly with the application of natural language processing and natural language generation; (3) semantic error detection reports; (4) data mining research; and (5) improved business intelligence systems that allow real-time dashboarding and alert systems, workflow analysis and improvement, outcomes measures and performance assessment.⁴⁶

⁴³ Christos Makridis et al., “Informing the Ethical Review of Human Subjects Research Utilizing Artificial Intelligence,” *Frontiers in Computer Science* 5, no. 5 (September 14, 2023), <https://doi.org/10.3389/fcomp.2023.1235226>.

⁴⁴ Michael R. MacIntyre et al., “Ethical Considerations for the Use of Artificial Intelligence in Medical Decision-Making Capacity Assessments,” *Psychiatry Research* 328 (September 7, 2023): 115466, <https://doi.org/10.1016/j.psychres.2023.115466>.

⁴⁵ Hannah Bleher and Matthias Braun, “Diffused Responsibility: Attributions of Responsibility in the Use of AI-Driven Clinical Decision Support Systems,” *AI and Ethics* 2, no. 4 (January 24, 2022), <https://doi.org/10.1007/s43681-022-00135-x>.

⁴⁶ Muyskens, “When Can We Kick Humans “Out of the Loop?”, in reference to Ho et al (2019),

Radiologists are not alone in this prospect, but that does not necessarily mean AI needs to be seen as a threat. While kicking radiologists out of the loop may indeed reduce the need for their aid in initial imaging analysis, it can be argued that the tool will only free up time for more meaningful work and unlock the opportunity to redefine their role in novel ways.

Automation will indeed represent challenges for specific roles and present risks in how healthcare professionals may or may not accurately use new technology. Nonetheless, automation may open doors for various other roles to be redefined in meaningful ways as well by allowing individuals to focus less on mundane tasks and more on interesting and particularly human ones. When the prospect of automation is a realistic and desirable probability, the ethical concern should turn to where the individuals being affected can direct their attention in novel and fulfilling ways.

Still, there are many tasks that simply cannot be replaced by AI and so evaluating automation on a case-by-case basis is necessary. Consideration of the social consequences of automation should also be exercised to ensure the wider impact of the study is justifiable. The research community involved in a study should be aware of the social ripple effects an AI system may induce such as job loss and difficulties with allocation of resources.

4. The Risks of Risk Prediction – How much risk is reasonable in an already risky world? ⁴⁷

4.1 What is risk prediction in AI?

Among the various risks associated with AI, risk prediction tools raise important ethical concerns due to their use in high stakes environments and their relationship with clinical judgment. According to the European Parliament,

Risk prediction focuses on assessing the likelihood of individuals experiencing a specific health condition or outcomes. It typically generates probabilities for a wide array of outcomes ranging from death to adverse disease events (e.g. stroke, myocardial infarction, bone fracture). The process involves the identification of individuals with certain diseases or conditions and their classification according to stage, severity, and other characteristics. These individuals may subsequently be targeted to receive specific medical interventions (Miotto et al., 2016; Steele et al., 2018; Fihn et al., 2019). ⁴⁸

Although risk prediction models have long been available, there have been questions about their value in the clinical setting because of worries about their limited predictive accuracy. The use of AI poses new opportunities to improve accuracy. In Singapore, there have been recent examples demonstrating these improvements. This includes our case study below of a machine learning triage tool called Score for Emergency Risk Prediction (SERP) used for estimating mortality after emergency admissions. RapidAI is another example of a successful risk prediction tool whereby stroke patients are identified in less than a minute, thus shaving off precious minutes of response time. ⁴⁹

Concerns about levels of risk and risk assessment are heightened within the context of emergency medicine where healthcare professionals must navigate critical medical urgencies in a high-pressure environment. As a result, issues with overcrowding and excessive delays have raised concerns about the quality of care in emergency departments globally and in Singapore. ^{50,51}

⁴⁷ Please note that this is an ongoing case study and the ideas resented here may be updated as new information arises.

⁴⁸ Eleanor Bird, Jasmin Fox-Skelly, Nicola Jenner, Ruth Larbey, Emma Weitkamp, Alan Winfield, Panel for the Future of Science and Technology, European Parliamentary Research Service, and Scientific Foresight Unit (STOA). "The Ethics of Artificial Intelligence: Issues and Initiatives." Report. STOA | Panel for the Future of Science and Technology. <https://doi.org/10.2861/6644>.

⁴⁹ National University Hospital. "AI-POWERED TRIAGE TOOL HELPS DOCTORS TREAT STROKES FASTER." (May 18, 2023). Press release.

<https://www.nuhs.edu.sg/sites/nuhs/NUHS%20Assets/News%20Documents/NUHS%20Corp/Media%20Releases/2023/Media-release-AI-powered-triage-tool-helps-doctors-treat-strokes-faster.pdf>.

⁵⁰ Ru Ying Fong et al., "Comparison of the Emergency Severity Index versus the Patient Acuity Category Scale in an Emergency Setting," *International Emergency Nursing* 41 (November 2018): 13-18, <https://doi.org/10.1016/j.ienj.2018.05.001>.

⁵¹ Naser B Elkum, CarolAnne Barrett, and Hisham Al-Omran, "Canadian Emergency Department Triage and Acuity Scale: Implementation in a Tertiary Care Center in Saudi Arabia," *BMC Emergency Medicine* 11, no. 1 (February 10, 2011), <https://doi.org/10.1186/1471-227x-11-3>.

The focus here is on the risks associated with AI triage tools, and how these risks should be evaluated appropriately, particularly in light of concerns about trust and well-recognized technological biases.

Technology Bias and Automation Bias

Technology bias refers to resistance toward new technologies either due to fear that it is too new to trust, to conform to the status quo, or both. The recent GE HealthCare Reimagining Better Health Study revealed that there are major issues surrounding trust in medical AI.

Clinicians are highly skeptical of the quality of data used to train AI algorithms, with only 33% of experienced providers saying they have trust in the data. A further 44 % believe that AI is vulnerable to built-in biases that could have a negative impact on patient care and outcomes [...] Globally, 55% of clinicians don't feel that AI is ready for medical use.⁵²

It has also recently been noted that despite equal assessments of reliability, people trust autonomous medical AI decisions less than decisions by human physicians, and interestingly, AI outperforming a doctor does not increase trust.⁵³

Like AI in healthcare, self-driving cars are also prone to this technology bias. Despite proving less-accident prone than human drivers, self-driving cars are often considered more dangerous.³⁹ That is not to say that we shouldn't be cautious in adopting these new technologies, but that we should be aware of the underlying reasons for caution and ensure they fairly reflect the level of potential risk.

In contrast, automation bias refers to the overreliance on AI. Overreliance can occur as “humans tend to trust algorithms once they have proven their efficiency and lose critical consideration for what they do”⁵⁴, making novice clinicians especially vulnerable. This can lead to healthcare professionals incorrectly using new tools by blindly trusting the AI conclusions and forgetting to practice clinical judgment. Automation bias thus can pose equally significant risks as technology bias and has proven a concern in several studies.^{55,56}

Technology bias and automation bias threaten the opportunity to effectively analyse the performance of AI meaning we may lose out on the potential major benefits. Moreover, this could result in a threat to the health of the patient when the tool is used incorrectly. In light of these concerns, the case study that follows will lead us to explore how to orient AI-related risks within the wider context of risks associated with risk prediction and triaging tools so that we can ethically discern if and how to implement such tools into clinical practice where a clinical trial may not be possible. As the evaluation of any new intervention (involving AI or not) involves some degree of risk, and since we may not always be able to remove risk completely, we shall

⁵² GE HealthCare. “Reimagining Better Health.” Report. June 2023. <https://www.gehealthcare.com/-/jssmedia/gehc/us/images/insights/reimagining-better-health/ge-healthcarereimagining-better-healthstudyjune222023jb25147xx.pdf?rev=-1>

⁵³ Georgiana Juravle et al., “Trust in Artificial Intelligence for Medical Diagnoses,” *Progress in Brain Research* 253 (2020): 263–82, <https://doi.org/10.1016/bs.pbr.2020.06.006>.

⁵⁴ GlobalData Thematic Intelligence, “The Ethics of AI-Powered Medical Triage,” *Medical Device Network*, August 9, 2023, <https://www.medicaldevice-network.com/sectors/healthcare/ai-medical-triage-ethics/#:-:text=However%2C%20studies%20have%20suggested%20that>.

⁵⁵ Raja Parasuraman and Dietrich H. Manzey, “Complacency and Bias in Human Use of Automation: An Attentional Integration,” *Human Factors: The Journal of the Human Factors and Ergonomics Society* 52, no. 3 (June 2010): 381–410, <https://doi.org/10.1177/0018720810376055>.

⁵⁶ Marina Chugunova and Daniela Sele, “We and It: An Interdisciplinary Review of the Experimental Evidence on How Humans Interact with Machines,” *Journal of Behavioral and Experimental Economics* 99 (August 2022): 101897, <https://doi.org/10.1016/j.socec.2022.101897>.

come to find that ethical concerns should focus on minimizing risk and assessing the threshold of acceptable risk.

4.2 Case Study: Score for Emergency Risk Prediction (SERP) – Machine Learning Triage Tool for Estimating Mortality After Emergency Admissions⁵⁷

Pre-admission triage includes an assessment of vital signs (body, temperature, pulse rate, respiratory rate and systolic and diastolic blood pressure) as well as oxygen saturation and mental status. Assessment is also reliant on the subjective judgement of healthcare professionals who should be attentive to ‘red flags’ such as patient disorientation or confusion, lethargy, severe pain, or distress.

In Singapore, all EDs currently use the national triage system known as the Patient Acuity Category Scale (PACS). PACS uses a symptom-based differential diagnosis approach based on patients’ presenting complaints and objective assessments such as vital signs and the Glasgow Coma Scale.

On this scale there are four categories:

P1: requires immediate attention

P2: in severe distress and requires critical care

P3: ambulant and presents with mild to moderate symptoms

P4: nonemergency cases more appropriately managed in the primary care setting

Duke-NUS recently developed a new interpretable AI model to profile the 30-day mortality risk at admission. This tool adopts a machine learning model called Autoscore to produce a new triage outcome score known as the Score for Emergency Risk Prediction (SERP). SERP is an additive, points-based scoring tool, which makes it quick to calculate, easier to explain and easier to interpret. Note that, like PACS, SERP is a ‘severity classification index’. It does not instigate or determine a care pathway.

A retrospective cohort study examined ED admissions at Singapore General Hospital between January 2009 and December 2016 using the hospital’s Electronic Health Records. This study examined SERP scores for 224,666 patients in the model training cohort and 42,676 patients in the testing cohort.

The analysis showed that SERP had better prediction scores for mortality risk at 30 days than existing, commonly-applied clinical triage scores, including PACS. However, it is currently unknown whether SERP can improve outcomes in actual clinical practice as further evidence is needed to validate its real-world predictive capabilities.

Due to certain interpretations of the limitations under Singapore’s Human Biomedical Research Act (HBRA) 2015, an evaluation of SERP would not be able to move forward as a clinical trial. Part 3 section 6b of the HBRA states that a waiver of requirement for appropriate consent for emergency research is only approvable where, “there is no professionally accepted standard of treatment or the available treatments are unproven or

⁵⁷ Feng Xie et al., “Development and Assessment of an Interpretable Machine Learning Triage Tool for Estimating Mortality after Emergency Admissions,” *JAMA Network Open* 4, no. 8 (August 27, 2021): e2118467, <https://doi.org/10.1001/jamanetworkopen.2021.18467>.

are unsatisfactory”.⁵⁸ Since PACS is the professionally accepted standard and SERP is unproven, the law suggests a clinical trial evaluating SERP would require a waiver of consent.⁵⁹ However, it is virtually impossible to gain consent from those in the most serious risk categories (or from a proxy, given the time pressures), meaning the ethical consideration turns to whether or not to implement SERP in clinical practice considering the risks and benefits with what data is available.

Asymmetric approaches to risk mitigation were also proposed with the aim of lowering risk. This will be further detailed will below.

4.3 SERP Risk Prediction – What is the Main Concern?

As discussed in the previous section, whether an AI tool makes decisions or not plays an important role in evaluating the level of risk it poses. As SERP is only a ‘severity classification index’, it is less risky than a decision-making tool because it will provide healthcare professionals with information that they will use to inform their ultimate triage judgment. Their interpretation is what will ultimately influence the patient care pathway. Therefore, the main risk associated with SERP are the real-world implications of false negatives and false positives as outlined below.

False negatives – SERP incorrectly underestimates mortality risks, informing clinical decisions that fail to attenuate risk of death or serious harm correctly, leading to worse patient outcomes

False positives – SERP incorrectly overestimates mortality risks, leading to over-treatment and the unnecessary use of limited medical resources

4.4 Assessing Risk – Are AI risks exceptional?

This section aims to explore how to ethically assess risk in the case of SERP considering the risks of false negatives and false positives. The assessment will help in determining an ethical approach to deploying SERP in the clinic for the first time in terms of assessing and managing risk.

Influence of Emotional Responses

Alongside PACS are various other frameworks used for assessing risk such as the Canadian Emergency Department Triage and Acuity Scale (CTAS), Emergency Severity Index (ESI), Manchester Triage System (MTS), Australasian Triage Scale, and Korean Triage and Acuity Scale (KTAS).^{60,61}

Each scale uses similar metrics (vital signs, healthcare professional judgment, etc.) and evaluates risk on a severity scale of four or five. However, different scales may focus on

⁵⁸ Human Biomedical Research Act

⁵⁹ It could be argued that SERP has indeed proved it is not “unsatisfactory” in comparison to PACS with the data collected in the retrospective study, and so a clinical trial should be permitted. However, analysing that argument is beyond the scope of this paper.

⁶⁰ Michael Christ et al., “Modern Triage in the Emergency Department,” *Deutsches Aerzteblatt Online* 107, no. 50 (December 17, 2010), <https://doi.org/10.3238/arztebl.2010.0892>.

⁶¹ Jae Yong Yu et al., “An External Validation Study of the Score for Emergency Risk Prediction (SERP), an Interpretable Machine Learning-Based Triage Score for the Emergency Department,” *Scientific Reports* 12, no. 1 (October 19, 2022), <https://doi.org/10.1038/s41598-022-22233-w>.

different areas. For example, the main differences between ESI and PACS are that ESI incorporates resource needs in the triage ratings, whereas PACS triage is based solely on presenting symptoms and objective clinical data.⁶² The Australasian Triage scale focuses on the time a patient can safely wait.⁶³

Each of these scales, including PACS, has the potential to undertriage (failure to identify acutely severe illness) or overtriage (overestimation of patient acuity) which may have dire consequences depending on the medical severity of the situation.⁶⁴ This is why healthcare professionals are given the jurisdiction to override PACS if their reassessment suggests it is necessary, and if there is doubt, they are encouraged to uptriage.⁶⁵

In the instance of overcrowding, for example, patients may be assigned a P1 score instead of P2 due to subconscious pressure put on the healthcare professional.⁶⁶ While these decisions are often well-intended, they may lead to inefficiencies in the workflow in the best case, and otherwise preventable deaths in the worst.

Like the possible false negatives and false positives that could occur due to a SERP failure, the current scoring systems as well as healthcare professionals are prone to similar mistakes. These kinds of mistakes may often be due to the high-pressure nature of the role, susceptibility to emotional responses, and environmental distractors such as noise and task interruptions; obstacles AI needn't overcome.^{67,68}

On the one hand, emotions can be incredibly helpful to clinicians in making ethical decisions by invoking empathy, care, and compassion. On the other hand, healthcare professionals may need to decompartmentalize their emotions in order to avoid rash decision-making that may deviate from standards set by institutional and regulatory frameworks.⁶⁹ In emergency medicine, it is particularly difficult to keep judgments impartial in the face of suffering patients and urgent requirements. While AI's lack of emotional responses may help to keep decision-making clearer, it can be argued that this will not always produce the best outcome either as both hold comparable levels of risk.

Interpretability

Interpretability in AI seeks to develop tools that transparently allow humans to understand the results and output created by the algorithms. This understanding needn't necessarily require any technical knowledge, but rather a high-level understanding of how the tool works so that the user can make sense of how the algorithm comes to its conclusions. This is unlike black box models which create predictions that are too complicated for human understanding.

⁶² Ru Ying Fong et al. "Comparison of the Emergency Severity Index"

⁶³ Charles C. Yancey and Maria C. O'Rourke, "[Figure, Australasian Triage Scale Figure. Contributed by Charles Yancey]," [www.ncbi.nlm.nih.gov](https://www.ncbi.nlm.nih.gov/books/NBK557583/figure/article-93329.image.f1/), July 30, 2021, <https://www.ncbi.nlm.nih.gov/books/NBK557583/figure/article-93329.image.f1/>.

⁶⁴ Florian F. Grossmann et al., "At Risk of Undertriage? Testing the Performance and Accuracy of the Emergency Severity Index in Older Emergency Department Patients," *Annals of Emergency Medicine* 60, no. 3 (September 2012): 317-325.e3, <https://doi.org/10.1016/j.annemergmed.2011.12.013>.

⁶⁵ Yoges

⁶⁶ CTAS National Working Group and Canadian Association of Emergency Physicians. "The Canadian Triage and Acuity Scale: Education Manual," 2012. https://caep.ca/wp-content/uploads/2017/06/module_1_slides_v2.5_2012.pdf.

⁶⁷ Hugh Gorick, "Factors That Affect Nurses' Triage Decisions in the Emergency Department: A Literature Review," *Emergency Nurse* 30, no. 3 (May 1, 2022): 14-19, <https://doi.org/10.7748/en.2022.e2123>.

⁶⁸ Philippe Delmas et al., "Effects of Environmental Distractors on Nurse Emergency Triage Accuracy: A Pilot Study Protocol," *Pilot and Feasibility Studies* 6, no. 1 (November 7, 2020), <https://doi.org/10.1186/s40814-020-00717-8>.

⁶⁹ Marisa Almeida et al., "Emotional Management Strategies in Prehospital Nurses: A Scoping Review," *Nursing Reports* 13, no. 4 (December 1, 2023): 1524-38, <https://doi.org/10.3390/nursrep13040128>.

While interpretability cannot be strictly defined, “an interpretable machine learning model is constrained in model form so that it is either useful to someone, or obeys structural knowledge of the domain, such as monotonicity, causality, structural (generative) constraints, additivity, or physical constraints that come from domain knowledge.⁷⁰ In other words, the tool provides its own domain-specific explanation such that the end-user can grasp the processing of inputs and resulting outputs.⁷¹

The potential risks of non-interpretable black box models are vast and can have severe consequences. Due to a lack of transparency and accountability, such tools have led to the release of dangerous criminals on bail⁷² and poor use of limited valuable resources.⁷³ However, an explainable and interpretable model poses far less risk by giving users the opportunity to use their own judgment and expertise in evaluating the quality of the tool’s output.

Thus, interpretability results in higher assurance in the validity and safety of the tool insofar as it may help to reduce underlying concerns the user may have, such as technology or automation bias and issues of trust. That is, interpretability itself may not necessarily describe anything about the riskiness of a tool, but the more explainable a tool is, the more comfortable the user will probably be trusting it and thus be more likely to use it correctly. Interpretability helps to ensure the user will appropriately employ the tool and not impose his or her own biases toward technology, whether positive or negative.⁷⁴ For these reasons among others, it has been argued that models that are explainable should be required in the clinical setting.⁷⁵ Further, arguments also support that the way forward in designing models is to create interpretable models in the first place rather than try to explain black box models.⁷⁶

SERP is an interpretable model. The SERP scoring models are derived from AutoScore, a machine learning points-based clinical score generation algorithm using just five variables: demographic characteristics, administrative variables, medical history in the preceding year, vital signs, and comorbidities. This makes it clear to the care team why some patients are given higher scores than others. In comparison to other complex models, point-based scores also prove more explainable by enabling users to easily build interpretable clinical scores. These scores can then be implemented and validated in clinical practice.⁷⁷

⁷⁰ Cynthia Rudin, “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead,” *Nature Machine Intelligence* 1, no. 5 (May 2019): 206–15, <https://doi.org/10.1038/s42256-019-0048-x>.

⁷¹ Yohei Okada, Yilin Ning, and Marcus Eng, “Explainable AI in Emergency Medicine: An Overview,” *Clinical and Experimental Emergency Medicine* 10, no. 4 (November 28, 2023): 354–62, <https://doi.org/10.15441/ceem.23.145>.

⁷² Rebecca Wexler, “When a Computer Program Keeps You in Jail,” *New York Times*, January 1, 2017.

⁷³ Kush R. Varshney and Homa Alemzadeh, “On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products,” *Big Data* 5, no. 3 (September 2017): 246–55, <https://doi.org/10.1089/big.2016.0051>.

⁷⁴ Saif Khairat et al., “Reasons for Physicians Not Adopting Clinical Decision Support Systems: Critical Analysis,” *JMIR Medical Informatics* 6, no. 2 (April 18, 2018): e24, <https://doi.org/10.2196/medinform.8912>.

⁷⁵ Yohei Okada, Yilin Ning, and Marcus Eng, “Explainable AI in Emergency Medicine”

⁷⁶ Cynthia Rudin, “Stop Explaining Black Box”.

⁷⁷ Feng Xie et al., “AutoScore: A Machine Learning–Based Automatic Clinical Score Generator and Its Application to Mortality Prediction Using Electronic Health Records,” *JMIR Medical Informatics* 8, no. 10 (October 21, 2020): e21798, <https://doi.org/10.2196/21798>.

Sufficient Evidence for Adept Technology

The conclusions from the retrospective cohort study in Singapore demonstrate SERP has better performance than existing triage scores. Additional evidence from a retrospective study in Korea also supports SERPs effectiveness. The Korean study aimed to externally validate SERP against other conventional scores, including the Korean Triage Acuity Scale (KTAS). The study found the performance of SERP to be superior to other scores for in-hospital and 30-day mortality prediction.⁷⁸ Nonetheless, accurate prediction retrospectively does not imply necessary efficaciousness in the real world. There are other concerns at play such as that the testing data ends up being not representative of future patients, but this may be considered reasonable risk if it is proportionate to the benefits.

Thus, the main issue is that SERP cannot be evaluated under a clinical trial so researchers must decide if implementing SERP is justifiable with what data is available. AI risk should be ethically evaluated similarly to other medical interventions and processes appropriate to the level of risk they pose. The assessment of risk needs to be calibrated to levels of risk that are acceptable for translation of new tech or other interventions into clinical contexts, and not necessarily treated as an exceptional type of risk. The riskiness of SERP not incorporating emotions into the calculation is comparable to the riskiness of the emotional responses of healthcare professionals in affecting false negatives and positives. The foundational evidence from the retrospective studies in Singapore and Korea, coupled with its interpretability, also suggests an appropriate risk threshold has been reached in order to move forward with implementing SERP. An analysis of the tradeoffs will be further explored in the next section.

4.5 To Implement or Not- What Are the Tradeoffs?

Assuming SERP cannot be evaluated under a clinical trial, we must assess whether or not moving forward with implementing SERP into clinical practice ethically justifiable. This can be achieved through a benefit and risk analysis.

The main question is whether we have enough sufficient retrospective evidence to justify translating SERP into practice such that we are satisfied that the potential benefits are sufficient, or rather, that SERP would indeed perform better than PACS at predicting mortality risk. This benefit would have to be proportionate to the potential risks of implementing SERP where a reasonable level of risk may be allowed considering the risks already present in the real world (i.e. healthcare professionals making decisions in a high-pressure environment). Reasonable risk refers to the proportionality of the risk compared to the benefits where the risk correlates to the potential harm induced and the benefits to the utility of the outcomes.

The potential benefit of SERP, evidenced in the results from the retrospective studies whereby SERP significantly improves prediction scores for mortality, is proportionate to the risk of SERP which lies mainly in that we do not have prospective data to prove it will indeed improve outcomes in actual clinical practice. The fact that SERP is not a decision-making tool suggests harm to the patient is lower than if it were to dictate a patient care pathway, but we cannot fully understand the risks of SERP until we have prospective data. As discussed above, the risks surrounding technology and automation bias are of important concern, but do not pose an exceptional risk to the already accepted risks in the emergency department. Therefore, it would seem that the overall benefits outweigh the risks, and implementing SERP is ethically justifiable.

⁷⁸ Jae Yong Yu et al., "An External Validation Study".

Benefits: Better performance in scores for in-hospital and 30-day mortality prediction evidenced in two retrospective studies

Risks: We do not yet fully understand the risks because we lack prospective data, but we believe risk to be low



Figure Five. The benefit to risk ratio considered in the ethical implementation of SERP.

What if there were not enough evidence to justify implementation?

If the retrospective data from the two SERP studies was not sufficient to justify implementation, an observational study could be conducted so long as no research interventions or tests are conducted. This would mean that any interventions or tests would be done in the absence of the observational study and are not provided for research purposes. This type of research would fall under a separate consent waiver provision involving data and tissue research which does not have the same requirements preventing a clinical trial of SERP described above.

Implementing SERP – A Regulatory Issue

Local Regulatory Issues Impacting Implementation

If at some point a hospital intended to implement the use of SERP for triaging patients, it may well have to clear a regulatory hurdle in Singapore. In particular, it may need to be reviewed and perhaps approved by the Singapore Health Sciences Authority (HSA). That step would raise, from a regulatory viewpoint, questions relating to the relationship between the risks and benefits from using this system that thus far have been discussed in this document in the context of an ethical (and not a regulatory) analysis.

In particular, HSA regulates and has to approve a variety of health-related products before they can be marketed and used in Singapore. This includes drugs and devices. Most commonly, devices are actual physical objects, such as a pacemaker. And some devices require the use of software to operate. But it can also be the case that software that is not connected to any physical device may nonetheless meet the definition of a medical device that requires regulatory review. This is commonly referred to as “Software as a Medical Device,” or SaMD. It could end up being the case that SERP, given that the scores it generates are intended to affect the triaging of patients, would need such a review. (See, e.g., HSA, Regulatory Guidelines for Software Medical Devices – A Life Cycle Approach, Revision 2.0, April 2022; HSA, Guidelines on Risk Classification of Standalone Medical Mobile Applications (SaMD) and Qualification of Clinical Decision Support Software (CDSS), July 2021.)

4.6 Steps toward full implementation – How should risk mitigation strategies be evaluated?

Technology bias, automation bias, and issues with human trust in AI has the potential to interfere with the benefits of SERP should these biases cause users to inappropriately apply its scores. If a healthcare professional approaches SERP with resistance toward its output, then SERP isn't given the chance to prove if it can indeed produce desirable outcomes in actual clinical practice. In the following sections, we will evaluate the unique ethical challenges of moving forward with the implementation of SERP by analysing proposed risk mitigation strategies.

Risk mitigation – An Asymmetric Approach?

While the risk of implementing SERP appears to be low, human biases, false positives and negatives, and a lack of prospective data all do indeed pose important risks and therefore suggest risk mitigation strategies should be considered. Risk mitigation strategies aim to minimize potential risks, but are separate from research and do not constitute research. Instead, risk mitigation strategies intend to improve care practice by evaluating newly implemented tools.

One risk mitigation strategy involves an asymmetric approach whereby the clinician receives both the SERP and PACS scores and can shift how the patient is managed in line with the SERP score, but in only one direction.

If the SERP score was lower than the PACS score (i.e. the AI model identified the patient as being more critically unwell), the patient could be managed according to the SERP score.

However, if the SERP score was higher than the PACS score (i.e. the AI model identified the patient as being less critically unwell) the patient's SERP score should be discounted in triage decision-making.

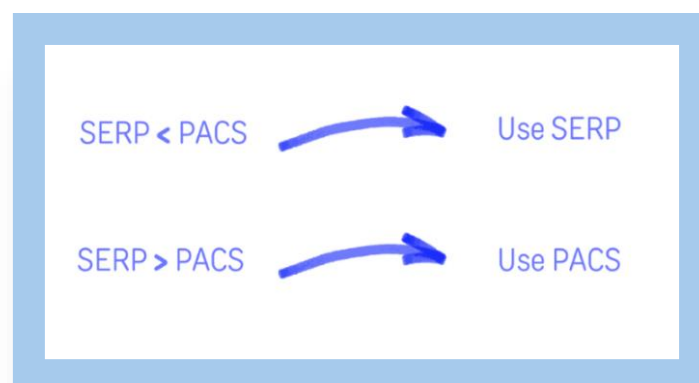


Figure Six. An asymmetric approach to risk mitigation with SERP and PACS scores.



Data could continue to be collected in under a quality improvement paradigm which provides a systematic approach to healthcare by “making processes safe, efficient, patient-centred,

timely, effective and equitable”.⁷⁹ The ongoing evaluation of SERP aims to ensure risks and benefits are being appropriately managed in the practice of triaging. Again, this is separate from research, especially as the goal is not to produce generalisable knowledge.

However, this approach does not come without tradeoffs. In the next section we will evaluate the tradeoffs of the asymmetric approach and discuss reasonable risk in risk mitigation strategies.

Evaluating the Tradeoffs of the Asymmetric Approach- What is Reasonable Risk Mitigation?

The subsequent ethical question revolves around the risks and benefits of the asymmetric approach as outlined below.

<p style="text-align: center;">Benefits</p> 	<ul style="list-style-type: none"> • Will likely reduce the risk of the real-world impact of false negatives\ndertriage (but not at all clear that this will reduce patient harm overall, given what is known about comparative false negative rates) • Will likely prevent risks associated with automation bias
<p style="text-align: center;">Risks</p> 	<ul style="list-style-type: none"> • Violates ethical principle of respect for clinical judgment • Undermines the quality of the data: data on the real-world effectiveness of SERP in critically ill patients will be incomplete • Potential to reinforce technology bias by suggesting necessary human manipulation of scores
<p>Figure Seven. The benefits and risks associated with the asymmetric approach to risk mitigation with SERP and PACS scores.</p>	

The main ethical concern with risk mitigations strategies like the asymmetrical approach is constraining the individual management of a case. We risk influencing clinical judgment and treating patients as though they are just a manifestation of their score, and not as actual persons. The risk increases as the complexities of a patient's case and need for care increases.

⁷⁹ SingHealth Duke-NUS Institute for Patient Safety & Quality (IPSQ), “Error,” www.singhealth.com.sg. January 2020, <https://www.singhealth.com.sg/Documents/IPSQ%20-%20Toolkit/Quality%20Improvement%20Toolkit%20-%20Version%2001a.pdf>.

Therefore, it can be argued that respect for clinical judgment is violated by directing the healthcare professional on which score to use rather than allowing for a higher exercising of professional expertise. However, this is only the case insofar as the score influences the healthcare professional's final triaging score. Respect for clinical judgment will still be upheld in the deciding of the actual patient care pathway.

The more practical concern is that this type of asymmetric approach would manipulate the "natural" data making it difficult to effectively evaluate SERP's actual performance. Instead, using SERP and PACS together and following SERP would give us the cleanest data possible to accurately evaluate SERP and thus improve care practice. As the healthcare professional will have the final say by exercising clinical judgment in the patient care pathway, the risk to the patient remains low.

The model for implementing SERP under a quality improvement framework would involve a titrating approach whereby PACS and SERP are both used alongside one another until/if there is enough robust data evidencing that SERP indeed improves patient outcomes in clinical practice, at which point clinical care can move to SERP only. Healthcare professionals can closely monitor the scoring process without intervening so that accurate data is collected. It would only be appropriate to intervene should changes be necessary. This maintains the likelihood that there will be less of an impact due to risk false negatives while maintaining respect for clinical judgment and hopefully also keeping the data of evaluating the performance of SERP as "natural" as possible.

From there, efforts to minimise risk should be maintained. Asymmetric approaches raise ethical concerns regarding respect for clinical judgment in the face of technology and automation biases. Although these biases are well-supported in the literature, it does not mean that we can assume all clinicians will be swayed by them. Information provision and a monitoring approach will be important in the aim of minimizing risk.

4.7 What principles apply in deciding a risk threshold for implementation as well as risk mitigation strategies?

There have been two overarching ethical deliberations thus far: 1) concerns surrounding the riskiness of the implementation of SERP, and 2) the concerns involved in risk mitigation strategies. The principles that apply are outlined in accordance below.

The principle of **proportionality** plays into in the weighing of the risks and benefits of how to implement SERP. The risk posed by SERP is not exceptional and the evidence from the retrospective studies make it reasonable to move forward with further collection of data as SERP is implemented as a quality improvement mechanism. Proportionality was also employed in evaluating how to move forward with implementation where reasonable risk coupled with the potential for impactful benefits justifies implementing SERP into clinical care.

Respect for persons can be drawn upon in evaluating risk mitigation strategies as the asymmetric approach potentially fails to treat patients as persons. "The principle for respect for persons requires that welfare and interests of data subjects and research participants should be duly protected and their right to make their own decisions without being coerced, misled or kept in ignorance should not be ignored".⁸⁰ Evaluating risk mitigation strategies considers the principle of **respect for clinical judgment**, although a certain level of risk in is justifiable as it may not be possible to obtain natural data without doing so. Nonetheless,

⁸⁰ Bioethics Advisory Committee. "A Consultation Paper". p. 34

clinical judgment would always be respected in how the patient pathway would move forward, even if it means overriding the SERP score.

Monitoring and **adaptivity** are at play in the suggesting healthcare professionals perform close surveillance of over the scores and ensure alignment with their professional judgment. Adaptivity is also important to the titrating model whereby the confidence of healthcare professional is aimed to reflect the performance of SERP and is used as a measure of understanding when more trust can be put in the model.

Reflexivity “prescribes careful scrutiny and assessment of emerging risks in the short run as well as in the long run in terms of downstream effects”.⁸¹ While we cannot fully understand the risks that could manifest from implementing SERP, the close monitoring approach employs reflexivity to ensure risks and benefits are appropriate managed.

4.8 Strategic Measures in Practice – What can you do to evaluate and manage reasonable risk?

This section aims to offer guidance on what you can do evaluate and manage reasonable risk. Once a risk threshold is achieved, minimizing risk becomes crucial.

Minimizing Risk- What are the questions to ask when evaluating reasonable risk?

The 2021 EU proposal on AI legislation classifies the riskiness of AI tools according to the severity of harm they may induce. Although the document does not specifically address AI in healthcare, it is likely that medical AI devices will be classified as high risk due to the safety and privacy concerns of AI in healthcare.⁸²

This is not to say that all medical AI tools would be high risk. This is especially true in the case of Spine AI and other certain image-reading tools that require less transparency as the clinical assessment of the results and do not ultimately inform the patient care pathway.

Risk prediction tools, especially in emergency medicine, pose higher risks due to the issue of urgency and potential for harm. However, just as clinicians rely on their colleagues, standard medical processes, and other technologies, they too must rely on new AI models appropriate to the level of risk the model poses. Where removing risk is not possible, the focus should be on minimizing it to the extent that implementing the new AI still provides significant utility.

Evaluating the threshold of reasonable risk of an AI risk prediction model should be akin to the evaluation for assessing the threshold of acceptable risk occurring from non-AI related processes and interventions. Should the implications of each be comparable, the project can move forward, and further implications of risk evaluated.

Below are proposed considerations when assessing the risks of an AI risk-prediction model.

1. Is the level of risk posed by the new AI tool comparable to the risks involved in the current process you are seeking to automate?
2. Are the risks reasonable? I.e. Is there proportionality in a risk to benefit ratio where the risk correlates to the potential harm induced and the benefits to the utility of the outcomes?

⁸¹ Alessandro Blasimme and Effy Vayena, 'The Ethics of AI', p 715.

⁸² European Commission, Directorate-General for Communications Networks, Content and Technology. CNECT. Proposal for a regulation. April 21, 2021. <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex:52021PC0206>

3. Is there sufficient evidence of an appropriate balance between risks and benefits? (i.e. evidential foundations for adept technology)
4. Have we taken appropriate steps to minimize risk? (i.e. monitoring, governing-human oversight, paired scoring to look for major deviations)
5. Is there another way evaluate the performance of the tool that would involve less risk while maintaining the purported benefits?

Conduct Careful Monitoring

Where possible, it is important to collect the most natural data possible in order to accurately evaluate the performance of any AI system. Therefore, implementing SERP under a quality improvement paradigm minimizes risk by monitoring the system's performance. Intervening with the system should only occur when there is sufficient natural data and unless necessary changes must be made. It is important to collect robust data before any human intervention so that accurate data can be collected interventions are not made solely, for example, on the basis of an individual clinician's judgment. This could be carried out by running both SERP and PACS side by side with close monitoring for evaluation of performance. Then, if emerging evaluations are satisfactory, the hospital can move to just SERP.

Automation and Tech Bias Information Provision

Include plans for proper information provision for clinicians to be aware of the effects of automation and tech bias that can potentially interfere with properly using an AI tool. This should include the risks to natural data, potential extension of study and added costs. Often these biases are unconscious and giving users of new technologies the information they need to appropriately employ the tool can help to mitigate some of the risk.

5. Conclusion

This document provides an analysis of three relevant ethical issues arising from the use of AI in biomedical research in Singapore: bias, human involvement, and risk.

In working through three case studies and analysing the risk-benefit ratio, the threshold for further use/research reached has demonstrated that certain harms or risks of harms may be justified when the benefits are sufficient. In some cases, this will allow AI that manifests social biases, allow AI that kicks humans out of certain loops, and allow AI that is potentially beneficial but still carries unknown risks to be introduced.

Nonetheless, there is an ethical obligation to mitigate these harms and risks through rigorous and ongoing risk mitigation and minimization strategies. Strategic measures to ensure ethical research are put forth for the research community to implement.

Principles, especially proportionality, are drawn upon to understand how these issues can be ethically deliberated. The principles are found through a bottom-up approach whereby the analysis of the case studies informs the salient principles.

Each analysis addresses real-world ethical issues that apply to the local Singaporean context and offer guidance on how to manage difficult ethical tradeoffs. The research community will be able to apply these strategies in ongoing and future work.

As the nature of AI is to progress at rapid speeds, this document will serve as a working document that will be updated accordingly as necessary. The last update of this document was on May 8th,

6. Annex 1

Variations of principles suggested to guide the ethical use of AI suggested by public and private sector parties.

Bioethics Advisory Committee⁸³

1. **Respect for persons:** Respect for persons includes respecting their right to make decisions without being coerced, misled, or kept in ignorance. The BAC refers to this as autonomy which can be broadly defined as the right of individuals to decide and act on their own volition and according to their own assessment of their interests. The welfare and interests of individuals are to be protected, especially when their autonomy is impaired or lacking. This principle underlies the importance that is often given to informed or appropriate consent to participate in research, protection of privacy, safeguarding confidentiality, and avoiding or minimising harm to research participants. The principle of respect for autonomy also includes proper regard for religious and cultural diversity in understanding of what constitutes the good or good life.
 - i. The principle of respect for persons or autonomy in big data and AI use in biomedical research can be demonstrated in the moral stance or attitude towards individuals (or groups). One of the ways this principle can be conveyed is through adequate communication. [...]
2. **Solidarity:** The BAC asserts that as some degree of mutual obligation exists between the individual and society, common interests of society may constrain individuals' autonomy and interests in specified circumstances. The principle of solidarity reflects the willingness and moral obligations of individuals to share the costs associated with research participation, such as potential risks, in return for the common good. Solidarity thus reflects the importance of altruism and other prosocial motivations and justifications as a basis for participation in biomedical research. There is a need to balance the interests of the public or society with the rights and interests of individual participants. Conflicting and irreconcilable ethical perspectives should be resolved by balancing public and individual interests. Based on the principle of solidarity, the BAC acknowledges that public interest may override individual rights and interests in certain circumstances, such as in public health and epidemiological research; and where appropriate safeguards are in place and the research poses minimal risk, requirements for obtaining informed consent or appropriate consent may be subordinated to those of public interests.
 - i. In the context of big data and AI use in biomedical research, data protection has been a key tenet of the governance model focused on privacy and individual rights. Such a governance model has been criticised for its focus on individual rights and interests, at the cost of collective and group interests.⁵ A solidarity-based data governance model may need to be considered to address this issue to promote sound biomedical research and to foster equitable and collective sharing in the benefits and costs of digital practices, while also appropriately respecting individual autonomy.

⁸³ Bioethics Advisory Committee. "A Consultation Paper". 24-28.

3. Justice: The principle of justice in the context of biomedical research encompasses the general principles of fairness and equity, which imply that access to the benefits of research, and the burden of supporting it, should be equitably and fairly shared in society. In the event that research yields an immediate benefit that could apply to participants in the research, reciprocity as a sub-set/element of the principle of justice would dictate that the benefits be offered to them. The principle of justice also implies that researchers and their institutions shoulder some responsibility for the welfare of participants in the event of adverse outcomes arising directly from their participation in the research.
 - i. Justice in the context of big data and AI biomedical research requires that researchers manage and use data in a manner that does not create or reinforce bias. Algorithms that have been trained using data obtained from biased systems (e.g., data predominantly obtained from a single group based on race, ethnicity, country of origin, or socioeconomic class) are likely to produce biased results, leading to decisional bias or skewed conclusions. [...]
4. Proportionality: The principle of proportionality requires that the methods or processes used in biomedical research are necessary and appropriate in relation to the research intent and the range of public and private interests at stake.³ Regulation of biomedical research should be proportional to the degree of possible threats to individual freedom, welfare, or the public good. As such, interference with individuals' autonomy, including their decisions, actions, or rights in carrying out or participating in research, should not exceed what is needed to achieve regulatory aims of mitigating anticipated threats and risks, and in promoting public interest. The risks in biomedical research and stringency of its regulation are acceptable if they are proportionate to potential benefits to the participant or others (e.g., future patients).
 - i. When assessing the processing of personal data for big data and AI use in biomedical research, proportionality requires that only personal data which is adequate for data robustness and quality and is relevant for the purposes of data processing is collected and used. Equally, the right to protection of personal data, while important, may not be the singular or primary objective in all situations and must be considered in relation to the common good, and be balanced against other fundamental rights, and executed in accordance with the principle of proportionality.⁸ Thus, for adequately anonymised or securely de-identified data, a 'light touch' or moderate regulation may be most appropriate in balancing individual rights with public interests. This entails that there should be safeguards in place to mitigate the risk of re-identification while allowing uses of the data for sound scientific research.
5. Sustainability: The principle of sustainability can be understood broadly to support arguments for the fair and just conservation of nature and minimisation of resource depletion for the good of the planet. Thus, research processes and outcomes should not unfairly jeopardise or prejudice the welfare of future generations.
 - i. The advent of big data and AI technologies can either benefit sustainability objectives or hinder their realisation, depending on their applications. Researchers have a complementary responsibility to reduce the environmental impact of big data and AI systems, including but not limited to their carbon footprint and energy consumption, to minimise climate change and environmental risk factors, and avoid the

unsustainable exploitation, use and transformation of natural resources contributing to the deterioration of the environment and the degradation of ecosystems.

- ii. Given their synergistic relationship, big data and AI, when used in tandem, can be harnessed to provide effective solutions to address environmental challenges and issues, and achieve sustainable development. Big data techniques allow for effective handling, processing and analysis of environmental data that may be complex in terms of volume, heterogeneity, and velocity. Integrating machine learning with big data can deepen the understanding of patterns from environmental data and allow meaningful insights to be drawn from the data
6. Integrity, transparency, and accountability: Researchers and their institutions should uphold the highest possible standards of professional and moral conduct during the conduct of biomedical research (principle of integrity), and should open their decision-making considerations, processes, and actions to public scrutiny (principle of transparency). The level of transparency should always be calibrated to the context and impact, as there may be a need to balance transparency with other principles such as data protection, safety, and security. For example, there may be circumstances where individuals are not aware of how their data is being accessed or used. Nonetheless, they should be fully apprised when a decision is informed by or made based on AI algorithms, especially when it affects their safety, interests or rights, and they should be able to access the reasons, including ethical reasons, for such decisions.⁷ Transparency relates closely to the principle of responsibility and accountability. Ethical responsibility and liability for the decisions and actions arising directly from research studies should be attributed to researchers and their institutions.
 7. Consistency: The principle of consistency dictates that the same ethical standards should be applied across similar situations to ensure fairness and trustworthiness. In this regard, IRBs and equivalent bodies should adhere to a practice of consistency. This includes using the same or similar required standards to evaluate research applications and protocols for research studies involving the use of big data and AI to protect the welfare, rights, and privacy of human subjects participating in these studies. IRBs should adhere to standards set out in advisories or guidelines issued by national advisory bodies, i.e., BAC's 2021 Ethics Guidelines.
 8. Stakeholder engagement: Stakeholder engagement extends beyond dissemination of information and further requires that decision-makers consider the views of all stakeholders, and take these into account where possible. Researchers and institutions should first define the stakeholders to be engaged and the processes for such engagements, particularly if they are considering access to significant data resources. Researchers and institutions who intend to use big data in biomedical research should consult relevant stakeholders such as research participants to explain the purpose of data usage and the parties who would be accessing their data. Similarly, for the design and development of AI algorithms and models, researchers and institutions should engage key stakeholders such as users, developers, and the public to understand the views, feedback, and concerns of the various groups. Meaningful stakeholder engagement happens when there is an opportunity to influence what happens in the future. In the biomedical research context, this might be input to research design, ethical oversight or overall governance of the research and the research findings.

Oxford Handbook of Ethics of AI “The Ethics of AI in Biomedical Research, Patient Care, and Public Health”⁸⁴

1. *Adaptivity* refers to the capacity of governance bodies and mechanisms to guarantee appropriate forms of oversight for new data sources and new data analytics that get incorporated in research, patient care, or public health activities.
2. *Flexibility* is the capacity to treat different data types based on both their source *and* on their actual use, it is premised on the consideration that data acquire specific ethical meaning in different contexts of use.
3. *Inclusiveness* stresses the need to include all affected parties in deliberations and decision-making practices about the use of data and algorithms in specific ambits. This component refers in particular to communities and actors that are historically marginalized, vulnerable, or otherwise excluded from the circuits of power, such as minorities and patient constituencies.
4. *Reflexivity* prescribes careful scrutiny and assessment of emerging risks in the short run as well as in the long run in terms of the downstream effects of big data and AI on interests, rights, and values, for example in terms of fair access to healthcare services, discrimination, stigmatization, medicalization, overdiagnosis, and so on.
5. *Responsiveness* refers therefor to the need for adequate mechanisms to mitigate effects of unauthorized access to personal health-related information.
6. *Monitoring* expresses the need to predispose regular scrutiny of data-related activities and their effects on health-related practices in order to anticipate the emergence on new vulnerabilities and undesirable outcome.

FUTURE -AI (EU)⁸⁵

Fairness
Universality
Traceability
Usability
Robustness
Explainability
Patient- Centricity

AI In Healthcare Guidelines AIHGle – MOH⁸⁶

Fairness
Responsibility
Transparency
Explainability

World Health Organization⁸⁷

1. Avoid harming others (sometimes called “Do no harm” or nonmaleficence).
2. Promote the well-being of others when possible (sometimes called “beneficence”). Risks of harm should be minimized, while maximizing benefits.

⁸⁴Alessandro Blasimme and Effy Vayena, ‘The Ethics of AI’, p 715.

⁸⁵ FUTURE-AI. Best practices for trustworthy AI in medicine. <https://future-ai.eu/>

⁸⁶ MOH Artificial Intelligence in Healthcare Guidelines (AIHGle).

⁸⁷ Ethics and governance of artificial intelligence for health: WHO guidance. (2021) Geneva: World Health Organization;. Licence: CC BY-NC-SA 3.0 IGO.

3. Expected risks should be balanced against expected benefits. Ensure that all persons are treated fairly, which includes the requirement to ensure that no person or group is subject to discrimination, neglect, manipulation, domination or abuse (sometimes called “justice” or “fairness”).
4. Deal with persons in ways that respect their interests in making decisions about their lives and their person, including health-care decisions, according to informed understanding of the nature of the choice to be made, its significance, the person’s interests and the likely consequences of the alternatives (sometimes called “respect for persons” or “autonomy”).

Indian Council of Medical Research- Ethical Guidelines for Application of Artificial Intelligence in Biomedical Research and Healthcare ⁸⁸



⁸⁸ ICMR, Ethical guidelines for application of Artificial Intelligence in Biomedical Research and Healthcare, (2023), 978-93-5811-343-3.
https://main.icmr.nic.in/sites/default/files/upload_documents/Ethical_Guidelines_AI_Healthcare_2023.pdf