AIM
AI in Medicine for Kids

Ethical Considerations of Data Science and Artificial Intelligence

Melissa McCradden, PhD, MHSc

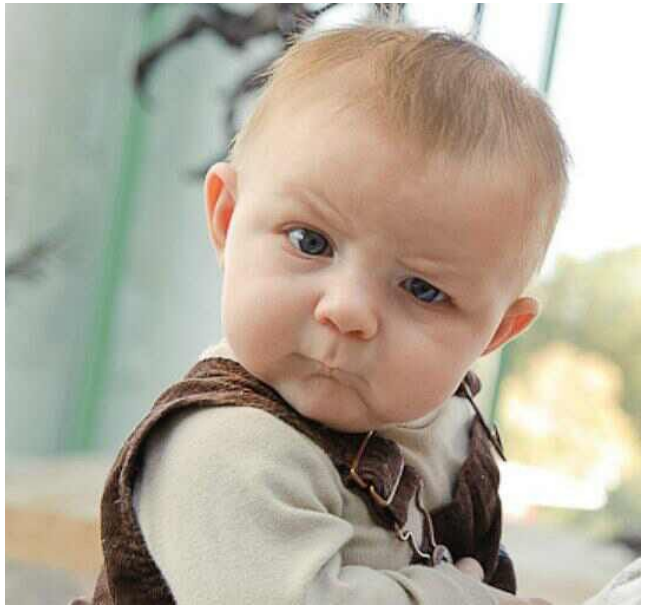
Bioethicist, The Hospital for Sick Children
Assistant Professor, University of Toronto

SickKids 2025

Agenda

Intro to Ethics & Bioethics
Ethical awareness & analysis
Issues: 1) explainability 2) bias
Critiques of AI Ethics

What ethical issues are you thinking about?



Possible signs of an ethical issue:

- Unsure of what to do or why
- Asking a ‘should I…” question
- Feeling uncomfortable


MIFF

What is ethics?

The study of moral behaviour

Morals: values concerning right and wrong; individual - can be subjective and varying

Ethics: concerning the moral parameters of particular activities; requires defensibility, evidence, publicity, (some) generalizability

What is Bioethics?

Bioethics involves critical reflection on issues relevant to health and human flourishing

- Deciding **what** we should do
- Explaining **why** we should do it
- Describing **how** we should go about doing it

Ethics is not a proxy for ‘bad things’ :)

Why?

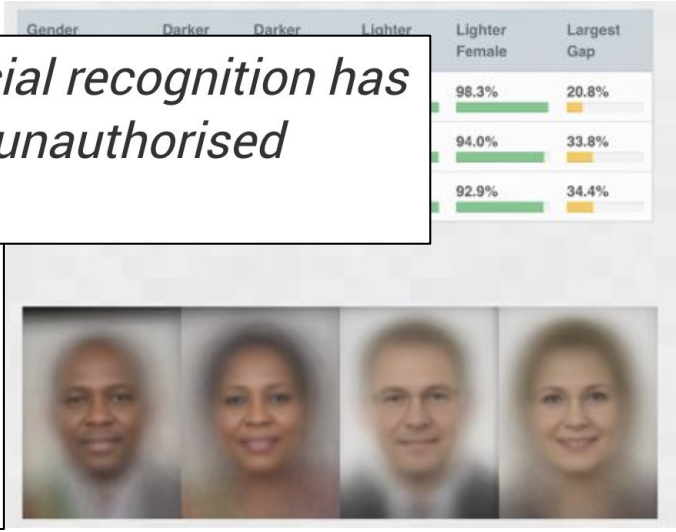
‘Bad things’ sometimes don’t start out as ‘bad’ → they are intended for security, efficiency, streamlining, or easing workflow

Our inte

ID passes are yesterday’s technology. Facial recognition has the potential to improve security, identify unauthorised access and keep track of visitors.

By ELAI

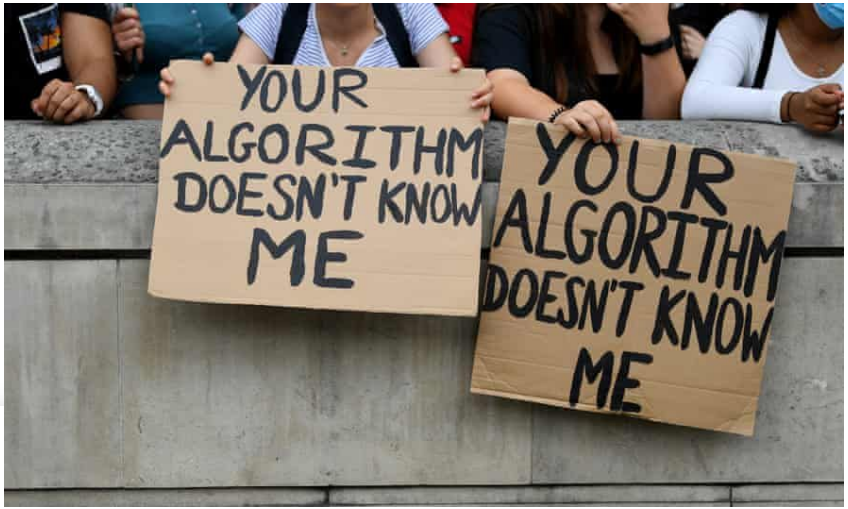
Wrongful arrest exposes racial bias in facial recognition technology



Why?

Europe

The U.K. used an algorithm to estimate exam results. The calculations favored elites.



When faced with an ethical challenge...

Expressively	Pre-reflectively	Reflectively
Through feelings, emotions, thoughts	By adherence to laws, religious tenets, codes of ethics	Through reasoned analysis, choice of ethical principles, theories, rules, values
Reactive	Compliant	BIOETHICS
“That’s totally wrong!”	“We should follow the law”	

Ethical frameworks

Most people know...

Autonomy

Beneficence

Non-maleficence

Justice

- Substantive
- Procedural
- Distributive
- Retributive
- Restorative

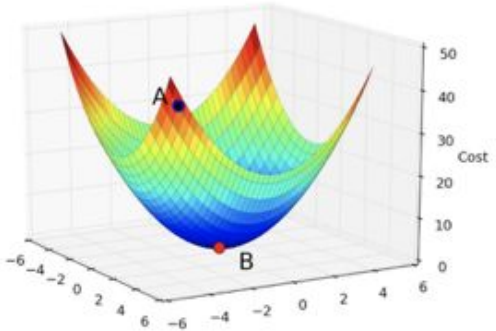
But there's also...

Power dynamics
Ethics of care
Feminist (intersectional, neoliberal, communitarian, relational, radical, empiricist)
Decolonial frameworks
More principles (equity, racial justice, solidarity, reciprocity, family-centred care, person-centred care, transparency, integrity, trust, best interests, utility, stewardship, confidentiality…

Reflective equilibrium

Deliberative, collaborative process of reflecting upon and revising our beliefs, both moral and non-moral, in a given area

Seeking coherence among principles, least tension → minimizing cost

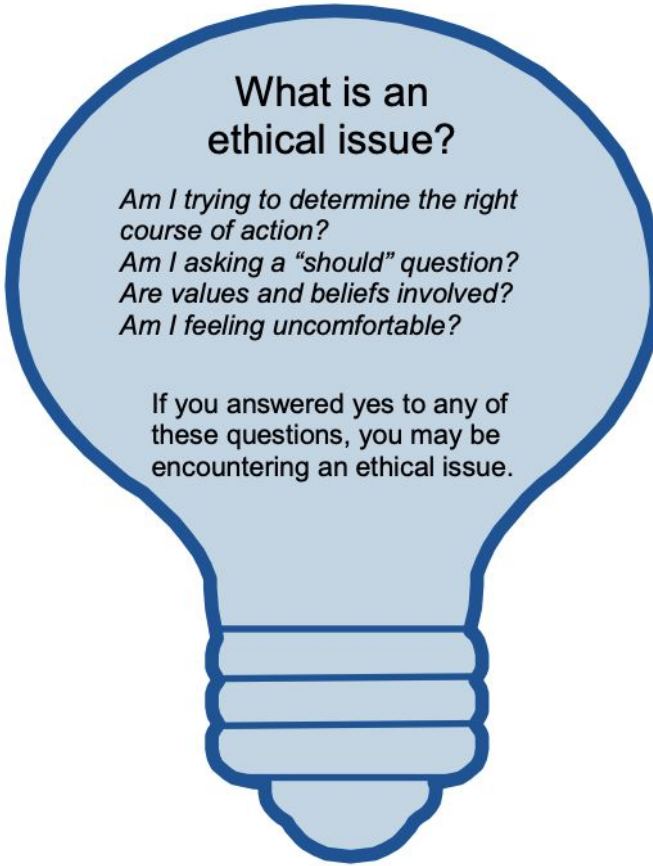


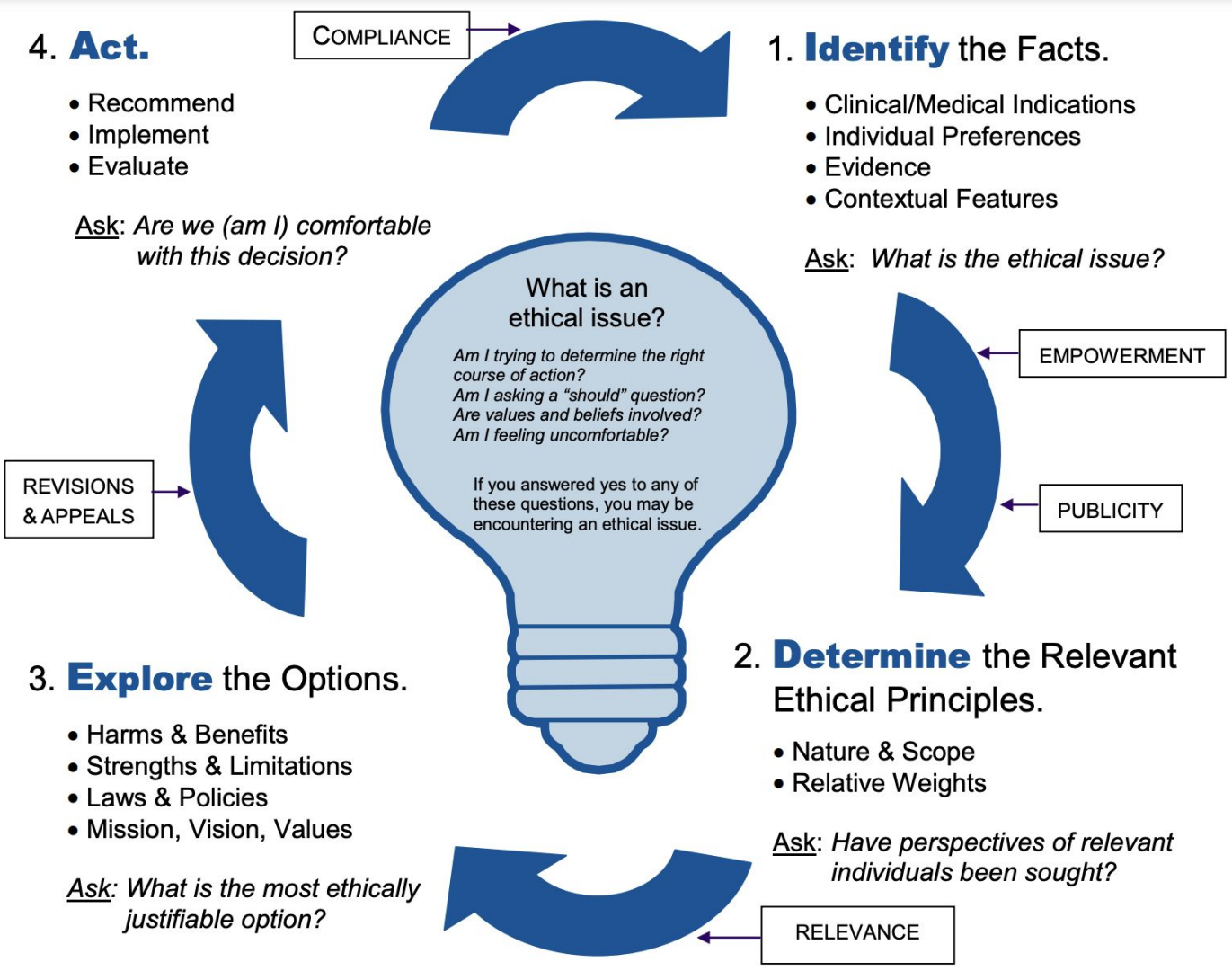
Ethical decision-making resource

Pull up this resource (link in the chat):

https://trilliumhealthpartners.ca/aboutus/Documents/IDEA-Framework-THP.pdf

Page 11-12 for ethical principles





SickKids 2025

Ethics in *Coded Bias*

“While conducting research on facial recognition technologies at the M.I.T. Media Lab, Buolamwini, a "poet of code," made the startling discovery that some algorithms could not detect dark-skinned faces or classify women with accuracy. This led to the harrowing realization that the very machine-learning algorithms intended to avoid prejudice are only as unbiased as the humans and historical data programming them.”

What ethical principles do you think are relevant to the themes in *Coded Bias*?

(in the IDEA framework or others as well!)

One more ethics message...

“Ethical AI” suggests we can achieve ethics as an end state

“Ethics of/in/for AI” suggests ethical thinking is ongoing, reflexive, and adaptive to the continued use and evolution of the technology

Why do you think this might be important?

Embedding ethics in AI development

nature machine intelligence

Explore Content | Journal Information | Publish With Us | Subscribe

nature > nature machine intelligence > comment > article

Comment | Published: 31 July 2020

An embedded ethics approach for AI development

Stuart McLennan, Amelia Fiske, Leo Anthony Celi, Ruth Müller, Jan Harder, Konstantin Ritt, Sami Haddadin & Alena Buyx

Nature Machine Intelligence **2**, 488–490(2020) | Cite this article

507 Accesses | **5** Citations | **80** Altmetric | Metrics

1. **Proactive** engagement
2. Regular formal and informal exchanges with ethicist
3. Increasing ethical awareness among team
4. Ethicists ideally have computational knowledge (but not necessarily expertise)

How can embedded ethics inform issues in machine learning?

1. Explainability
2. Bias

Explainable AI (XAI)

Generally, XAI focuses on helping the user ‘understand’…

- how the **model** works as a system
- how it arrived at a particular **prediction**

Two levels of explanations serve different goals

- Model-level explanations enable checks for patient safety, potential confounders, clinical face validity, fairness and bias

- Prediction-level (post-hoc) explanations are intended to inform a normative judgment: should I trust this prediction?[1]

[1]Selbst & Barocas. 2018 “The intuitive appeal of explainable machines” *Fordham Law Rev*

Why do you think explainability might be important?

Ethics arguments for XAI[1,2]

In recognition of the opacity of many high-performing AI systems, a ‘fifth principle’ for medical ethics (‘explicability’) has been proposed[1]

- To ensure accountability in decision-making
- To resolve physician-computer discrepancies
- To respect patient autonomy by ensuring informed consent

Is explainability really necessary for accountability and ensuring informed decision-making?

[1]Floridi et al., 2018 “AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations.” *Minds and Machines;* [2]Amman et al., 2021 “Explainability for artificial intelligence in healthcare: a multidisciplinary perspective” *BMC Med Inform Dec Making*;

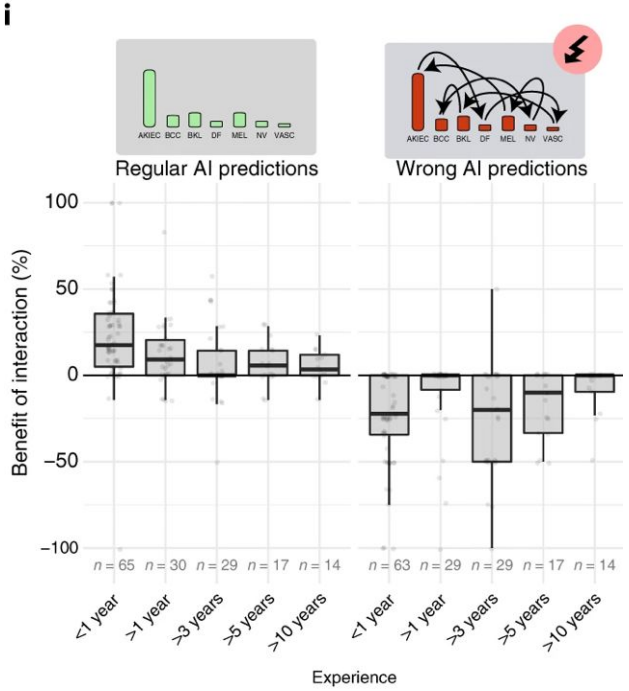
Prevalent ethics-based arguments for explainability go like this…

- Opacity is a problem - transparency is the remedy
- Transparency is achieved through explanations
- Explanations will enable clinicians to recognize incorrect predictions
- These incorrect predictions will be rejected

???

Case 1: Skin cancer recognition

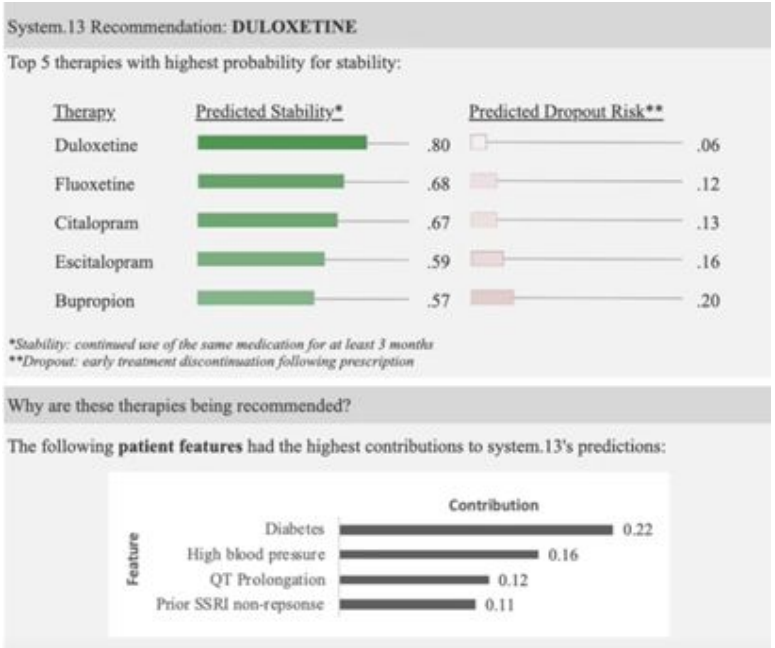
- Prediction of 7 distinct skin cancers
- When predictions were **correct**, clinician accuracy **improved**
- When predictions were **incorrect**, however, clinicians were often **misled** - no effect of years of experience
- Authors consider whether XAI/interpretable methods may prevent over-reliance



Tschandl et al., "Human–computer collaboration for skin cancer recognition" 2020 *Nat Med*

Case 2: Antidepressant prescribing

- Expert-generated ranking of ADs given patient scenarios = simulated ML model
- Systematically varied scenario, prediction accuracy, and explanation
- Main effect of explanations on incorrect predictions
- Following incorrect predictions happened mostly with **feature-based explanations**



Jacobs et al., "How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection" 2021 *Translational Psychiatry*

Case 3: Sanity checks for saliency maps

- Explanations may not always be faithful to the original model

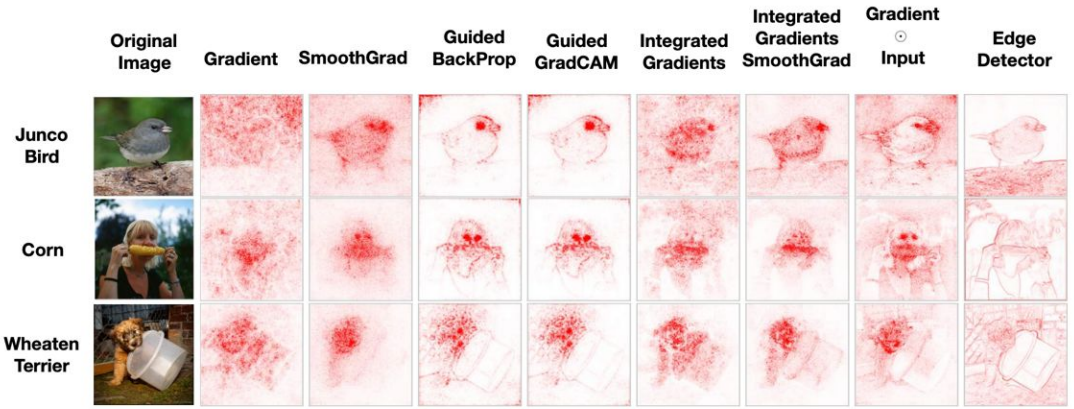


Figure 1: **Saliency maps for some common methods compared to an edge detector.** Saliency masks for 3 different inputs for an Inception v3 model trained on ImageNet. We see that an edge detector produces outputs that are strikingly similar to the outputs of some saliency methods. In fact, edge detectors can also produce masks that highlight features which coincide with what appears to be relevant to a model's class prediction. Interestingly, we find that the methods that are most similar to an edge detector, i.e., Guided Backprop and its variants, show minimal sensitivity to our randomization tests.

Adebayo et al., 2018 "Sanity checks for saliency maps" 2018 *NeurIPS*

Additional evidence

- Explainability may decrease our likelihood to detect mistakes[1]
- Explainability may confer inappropriate level of confidence in our judgments about prediction accuracy[2,3]
- Explainability does not prevent people from making worse decisions when faced with incorrect outputs compared with their own baseline judgment[4,5,6]

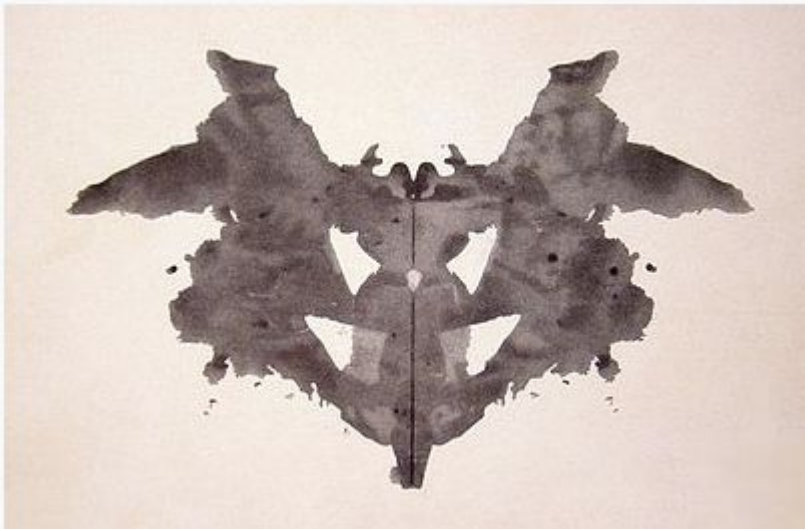
[1]Poursabzi-Sangdeh 2019 "Manipulating and measuring model interpretability" arXiv preprint; [2]Ghassemi et al., 2018 “ClinicalVis: Supporting clinical task=focused design evaluation.” arXiv preprint; [3]Eiband et al., 2019 “The impact of placebic explanations on trust in intelligent systems.” CHI; [4]Bansal et al., 2021 “Does the whole exceed its parts? The effect of AI explanations on complementary team performance.” CHI; [5]Buçina et al., 2020 “Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems.” ACM; [6]Zhang et al., 2020 “Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision-making.” FAT*20

Explanations: the new Rorschach test?

We may see what we're expecting to see or want to see[1] → confirmatory bias?

Introduces a second source of potential error

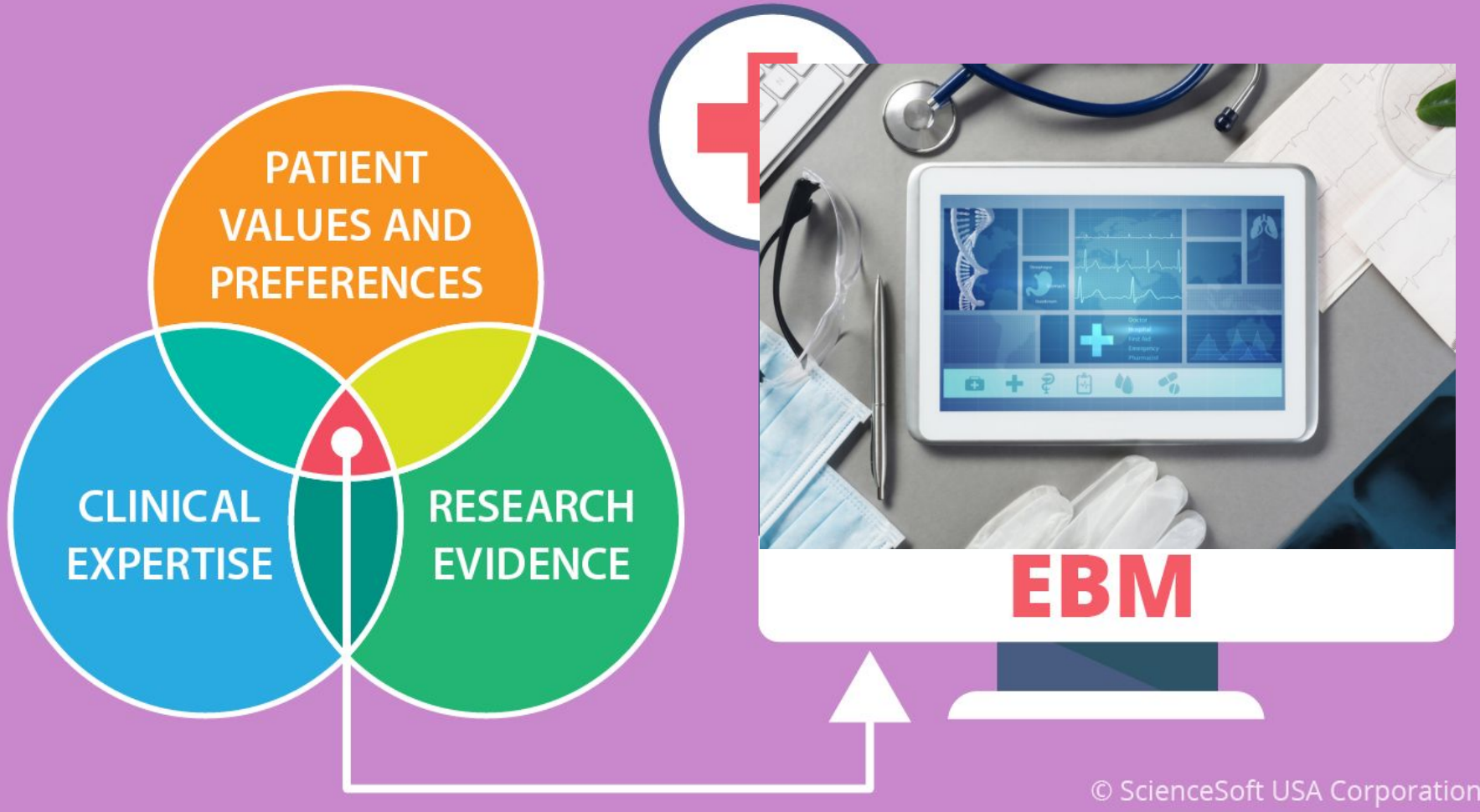
Displaces the importance and reliance on real-world evidence to ground judgments



[1]Rudin, 2019 "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nat Med Int*

Does explainability achieve its ethics-related goals?

No.



Work in collaboration with Shalmali Joshi and Alex John London

Accountability in medical decision-making

Medical recommendations are made by *reasoned judgments* in the face of incomplete causal knowledge[1,2]

Medical decisions are made by capable patients or surrogates on the basis of knowing (*HCCA,* 1996):

- The anticipated risks and benefits of the proposed plan
- The likely consequences of possible actions

XAI prioritizes knowledge of the model over knowledge of the patient and their context[3]

[1]London, 2019 “Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability” Hastings Center Report; [2]Ferry-Danini “What is the problem with the opacity of artificial intelligence in medicine?” *C4E Talks*, Jan 2021; [3]Work in progress, Shalmali Joshi, James Anderson, & Alex John London

A different ethical framework for accountable decision-making

Centring the moral commitments of clinicians to their patients tells us:

1. Implement systems that offer a net benefit over status quo (don't trade one bias for another)
2. Evidence-based, performance evaluated in the real world
3. Informed decision-making is achieved by explaining the reasons for a recommendation and likely consequences of different courses of action
4. Opacity is not the problem, automation bias is - we must get better at interacting with technologies to prevent harms and mistakes

Gender bias concerns raised over GP app

Written by **Sam Trendall** on 13 September 2019 in **Features**

Onlookers are asking why the chatbot created by Babylon Health – which provides the GP at Hand service – is offering such different guidance to men and women. But the company tells *PublicTechnology* its service is working as intended.

SCIENTIFIC REPORTS

Article | OPEN | Published: 15 April 2019

Genetic risk factors identified in populations of European descent do not improve the prediction of osteoporotic fracture and bone mineral density in Chinese populations

Yu-Mei Li, Cheng Peng, Ji-Gang Zhang, Wei Zhu, Chao Xu, Yong Lin, Xiao-Ying Fu, Qing Tian, Lei Zhang, Yang Xiang, Victor Sheng & Hong-Wen Deng

Scientific Reports **9**, Article number: 6086 (2019) | Download Citation

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

HEALTH

AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind

Machine learning has the potential to save thousands of people from skin cancer each year—while putting others at greater risk.

ANGELA LASHBROOK AUG 16, 2018

Ethical Machine Learning in Health Care

Irene Y. Chen,[1] Emma Pierson,[2] Sherri Rose,[3] Shalmali Joshi,[4] Kadija Ferryman,[5] and Marzyeh Ghassemi[4,6]

[1]Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA; email: iychen@mit.edu
[2]Microsoft Research, Cambridge, MA, 02143, USA
[3]Center for Health Policy and Center for Primary Care and Outcomes Research, Stanford University, Stanford, CA, 94305, USA
[4]Vector Institute, Toronto, ON, Canada
[5]Department of Technology, Culture, and Society, Tandon School of Engineering, New York University, Brooklyn, NY, 11201, USA
[6]Department of Computer Science, University of Toronto, Toronto, ON, Canada

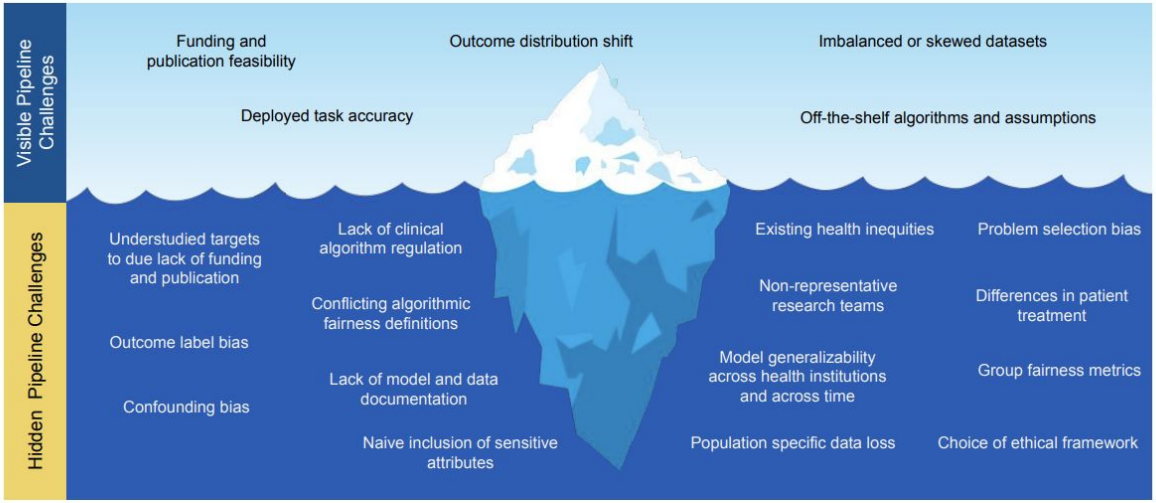


Figure 2

The model development pipeline contains many challenges for ethical machine learning for health care. We highlight both visible and hidden challenges.

BEHAVIOR & SOCIETY | OPINION

Yes, Science Is Political

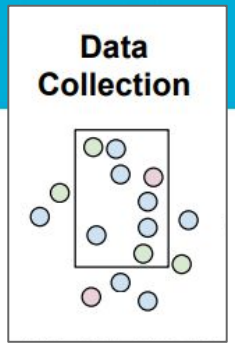
Scientists need to acknowledge that fact—and to act on it in these most dire of times

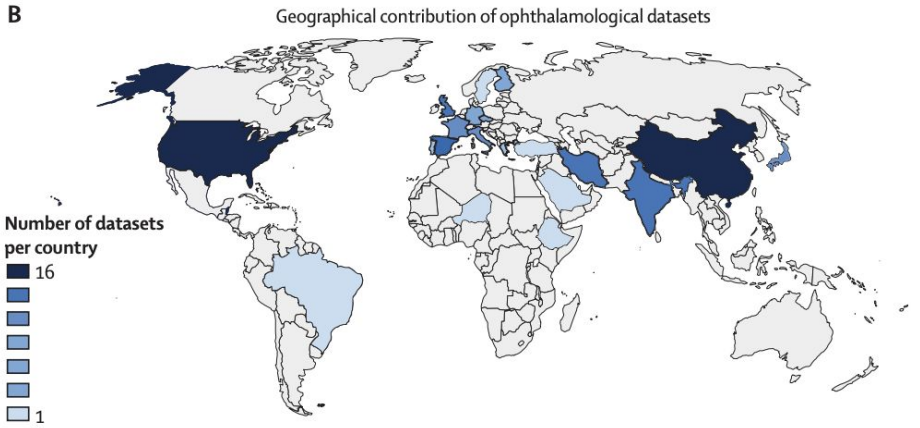
By Alyssa Shearer, Ingrid Joylyn Paredes, Tiara Ahmad, Christopher Jackson on October 8, 2020

SCIENTIFIC AMERICAN®

- Worldwide disparities in the study and funding health research
  - Gap for: health among citizens of the Global South, poverty-related diseases, racialized groups, women's health, mental health
- Health data poverty[2]
- Some problems are better studied and understood than others

[1]Chen et al., "Ethical machine learning in health care" 2020 *arXiv preprint*; [2]Ibrahim et al., "Health data poverty: an assailable barrier to equitable digital health care" *Lancet Digital Health* 2021





- Representation[2]
- Data quality
  - Interventional trial data, social media data, health records
- Diversity in scientific workforce

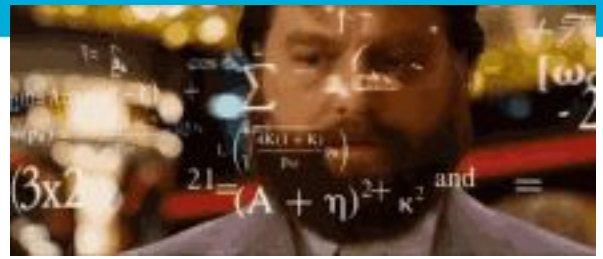
[1]Chen et al., ”Ethical machine learning in health care” 2020 *arXiv preprint*; [2]Khan et al., “A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalizability” *Lancet Digital Health* 2021

Outcome Definition

- Labels reflect the current state of knowledge on a topic and are affected by societal attitudes, policy, law
- Label noise[1]: access, incentives, inconsistencies, structural biases
- Both under- and over-diagnosis[2]
- Risk of preserving our axioms of the past

[1]Chen et al., "Ethical machine learning in health care" 2020 *arXiv preprint*; [2]Mullainathan & Obermeyer, "Diagnosis physician error: a machine learning approach to low-value health care" *arXiv preprint* [3]Khan et al., "Clinical Diagnosis—Is There Any Other Type?" *JAMA Int Med* 2020

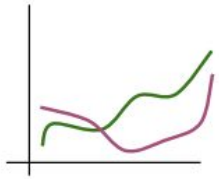
Algorithm Development



- Algorithms are not neutral[1]
- Risks of particular model choices may put some patients more at risk than others
  - Confounding: finding non-causal, associationist patterns in data
  - Particularly regarding social determinants, systematic racism
- Feature selection, tuning parameters, performance metrics, group fairness definitions

[1]Chen et al., ”Ethical machine learning in health care” 2020 *arXiv preprint*;

Post-Deployment Considerations



Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing

Inioluwa Deborah Raji*
Partnership on AI
deb@partnershiponai.org

Andrew Smart*
Google
andrewsmart@google.com

Rebecca N. White
Google

Margaret Mitchell
Google

Timnit Gebru
Google

Ben Hutchinson
Google

Jamila Smith-Loud
Google

Daniel Theron
Google

Parker Barnes
Google

- Most important: **auditing** and oversight[1]
  - Continuous quality improvement[2]
- Promote accountability, identify targets for improvement
- Real-world evaluation through evidence-gathering paradigms
- Requires considerations of scale and actionability

[1]Chen et al., "Ethical machine learning in health care" 2020 *arXiv preprint*; [2]McCradden et al., "Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning" *JAMIA* 2020; Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In Conference on Fairness, Accountability, and Transparency (FAT* '20)

Is prediction unfairness a harm *per se*?

- Harm (non-maleficence): refraining from providing ineffective or negative treatments, violating trust, or causing undue hardship
- Discrepant performance of models is inevitable – but it's not *necessarily* a harm until someone is negatively impacted
  - E.g., doctors redressed some of the algorithmic bias in Obermeyer et al., 2019

RESEARCH ARTICLE

ECONOMICS

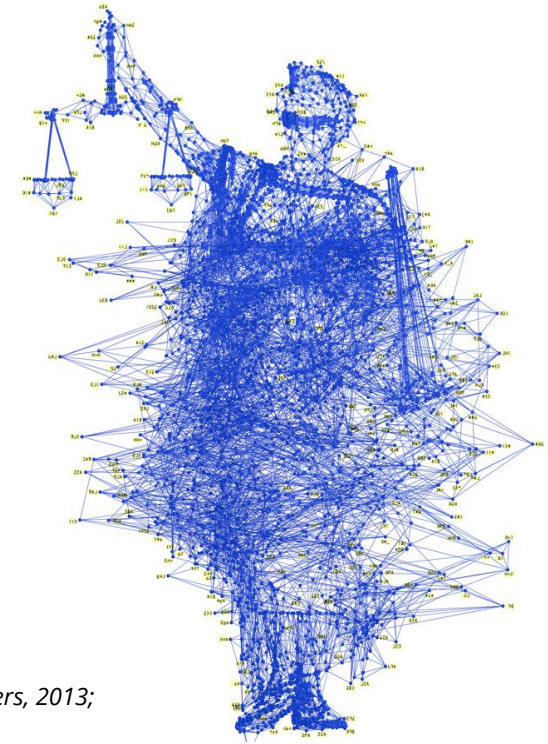
Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Can we prevent harm with algorithmic fairness?

- Fairness in ML[1] = loss of accuracy
- Types of potential computational solutions[2]
  1) Data-based methods[3,4,5]
  2) Model-based methods[6,7]
  3) Post-hoc methods[8,9]

...but which, when, why, & how?



[1]Hutchison & Mitchell, 2019 arXiv; [2]Suresh & Guttag, 2018 arXiv; [3]Hajian & Domingo-Ferrer, 2013; [4]Kamiran, Zliobaite & Calders, 2013; [5]Kamiran & Calders, 2012; [6]Zafar et al., 2017; [7]Zemel et al., 2013; [8]Hardt et al., 2016; [9]Corbett-Davies & Goel, 2018

Can we prevent harm with algorithmic fairness?

COMMENT | VOLUME 2, ISSUE 5, E221-E223, MAY 01, 2020

Ethical limitations of algorithmic fairness solutions in health care machine learning

Melissa D McCradden • Shalmali Joshi • Mjaye Mazwi • James A Anderson

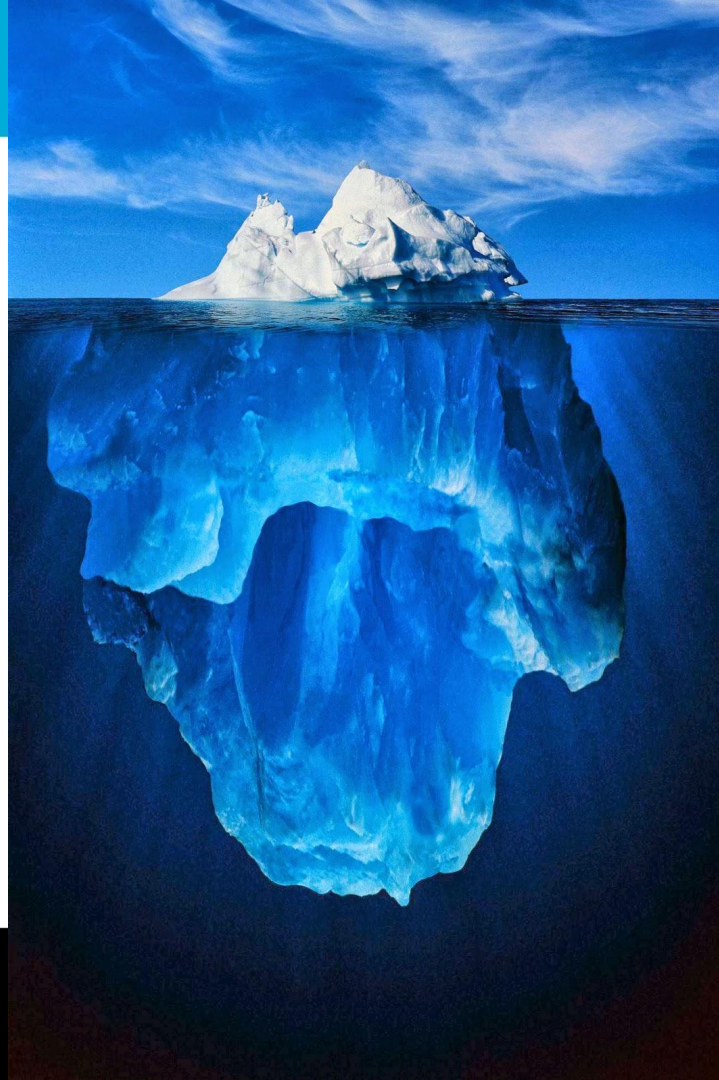
Open Access • Published: May, 2020 • DOI: https://doi.org/10.1016/S2589-7500(20)30065-0 • Check for updates

THE LANCET
Digital Health

- More than just technical choices → requires ethical analysis centred on impacted persons
- Constrained by:
  - What we know about the causal nature of unfairness patterns (epistemic)
  - How we will evaluate the model's real-world performance (empirical)
- Guided by the desired intent of the model's use at the point-of-care

Bias: Key Points

- We see only the tip of the iceberg, but not the history underneath
  - Value of multi-disciplinary insights
- Resist the notion that one can ‘solve’ bias through machine learning
- Use the insights to make informed model choices
- Aim for justice and accountability instead



“Ethics? Whose ethics?”

Criticisms of AI Ethics

Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning

Daniel Greene
College of Information Studies
University of Maryland
dgreene1@umd.edu

Anna Lauren Hoffman
The Information School
University of Washington
alho@uw.edu

Luke Stark
Microsoft Research Montreal
luke.stark@microsoft.com

Abstract

This paper uses frame analysis to examine recent high-profile values statements endorsing ethical design for artificial intelligence and machine learning (AI/ML). Guided by insights from values in design and the sociology of business ethics, we uncover the grounding assumptions and terms of debate that make some conversations about ethical design possible while forestalling alternative visions. Vision statements for ethical AI/ML co-opt the language of some critics, folding them into a limited, technologically deterministic, expert-driven view of what ethical AI/ML means and how it might work.

RESEARCH-ARTICLE OPEN ACCESS

You Can't Sit With Us: Exclusionary Pedagogy in AI Ethics Education

Authors: Inioluwa Deborah Raji, Morgan Klaus Scheuerman, Razvan Amironesei Authors Info & Affiliations

Publication: FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency • March 2021 • Pages 515–525 • https://doi.org/10.1145/3442188.3445914

However, we claim that the current AI ethics education space relies on a form of "exclusionary pedagogy," where ethics is distilled for computational approaches, but there is no deeper epistemological engagement with other ways of knowing that would benefit ethical thinking or an acknowledgement of the limitations of uni-vocal computational thinking. This results in indifference, devaluation, and a lack of mutual support between CS and humanistic social science (HSS), elevating the myth of technologists as "ethical unicorns" that can do it all, though their disciplinary tools are ultimately limited.

Why does ethics matter in science and research?

Practical take-aways

1. IDEA framework can help you think through ethical issues

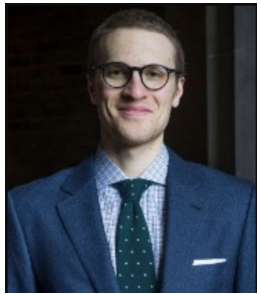
2. Reach out for multi-disciplinary perspectives on your projects

3. Frame your thinking around the intended use and impact of models, then think through the data, design choices, and outputs

4. Evidence over explainability!

Anna Goldenberg, Mjaye Mazwi, Victor Sami
AIM Initiative - The Hospital for Sick Children

Armando Lorenzo, Lauren Erdman, Mandy Rikard, Marta Skreta
Department of Urology - The Hospital for Sick Children



Erik Drysdale, AIM

Shalmali Joshi, Harvard

Danny Eytan, Azzy Assadi, Peter Laussen, Anusha Jegatheeswaran, Mjaye Mazwi, Andrew Goodwin, Robert Greer, Sebastian Goodfellow

Thank you!