MACHINE LEARNING IN THE OPERATING ROOM

FRANK RUDZICZ

TODAY

1. Some risks of deep machine learning
   a) Mitigation of those risks with 'Explainable AI'
2. Potential of deep machine learning in the operating room
3. Implications for practice

UNIVERSITY OF TORONTO

STOKING THE HYPE: TO ERR IS HUMAN?

- **Humans** are notoriously **bad** with information.
  - Patients **misread** or **miscommunicate** their own symptoms.
    - Nearly **half** of American adults have difficulty understanding and acting upon health information (IOM, 2004).
  - Faulty memory; skill obsolescence; cognitive biases; cognitive/time limitations; **recency biases**; other human biases.
    - *Diagnoses* correlate with advertising and media exposure.

- Winters *et al.* (2012) showed that ~40,500 patients die in ICU, in the USA, each year due to misdiagnosis.

http://www.nap.edu/openbook.php?record_id=10883&page=1

Winters *et al.* (2012) Diagnostic errors in the intensive care unit: a systematic review of autopsy studies. *BMJ Qual Saf* 2012;**21**:894-902

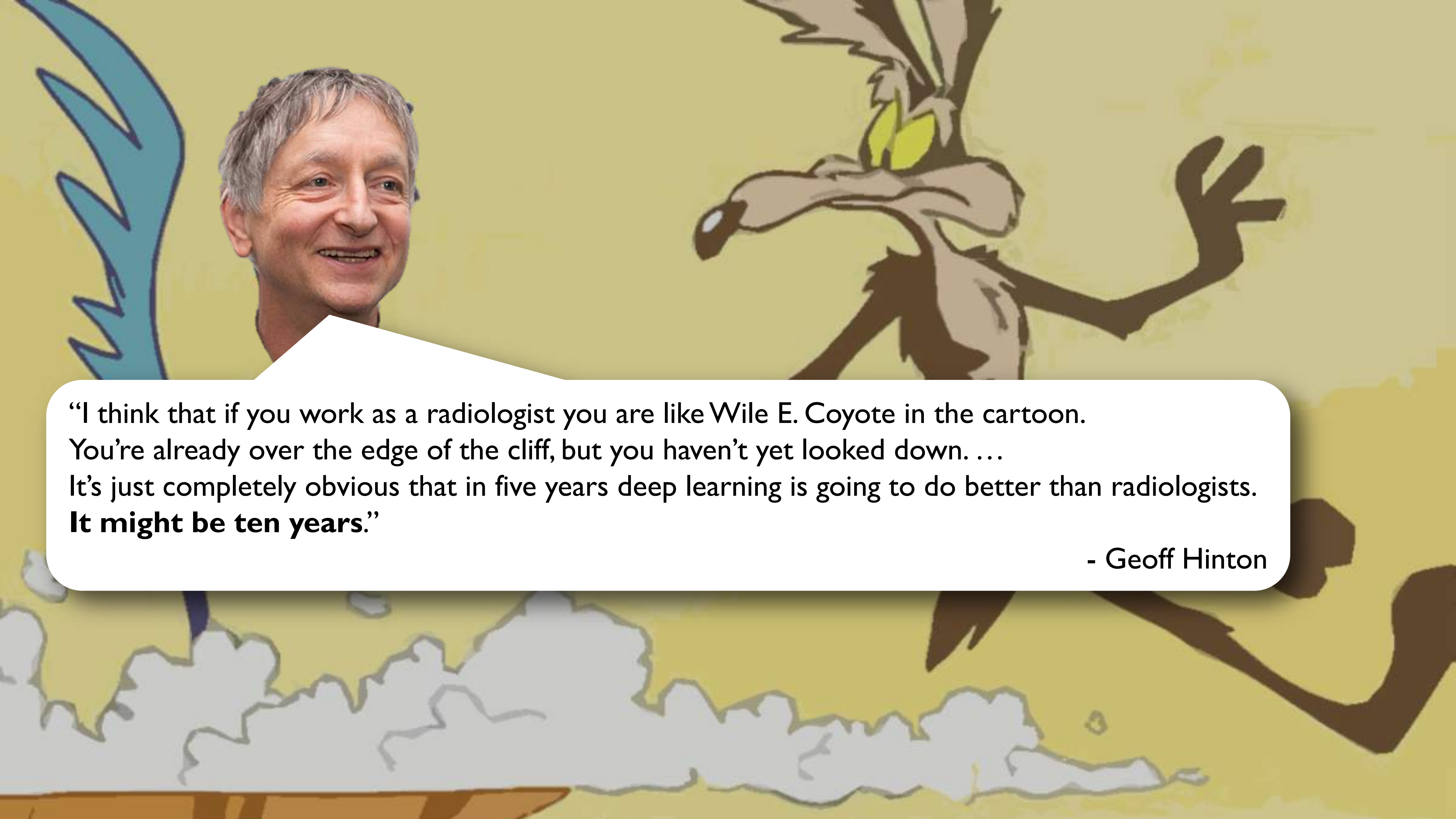
STOKING THE HYPE: TO ERR IS HUMAN?

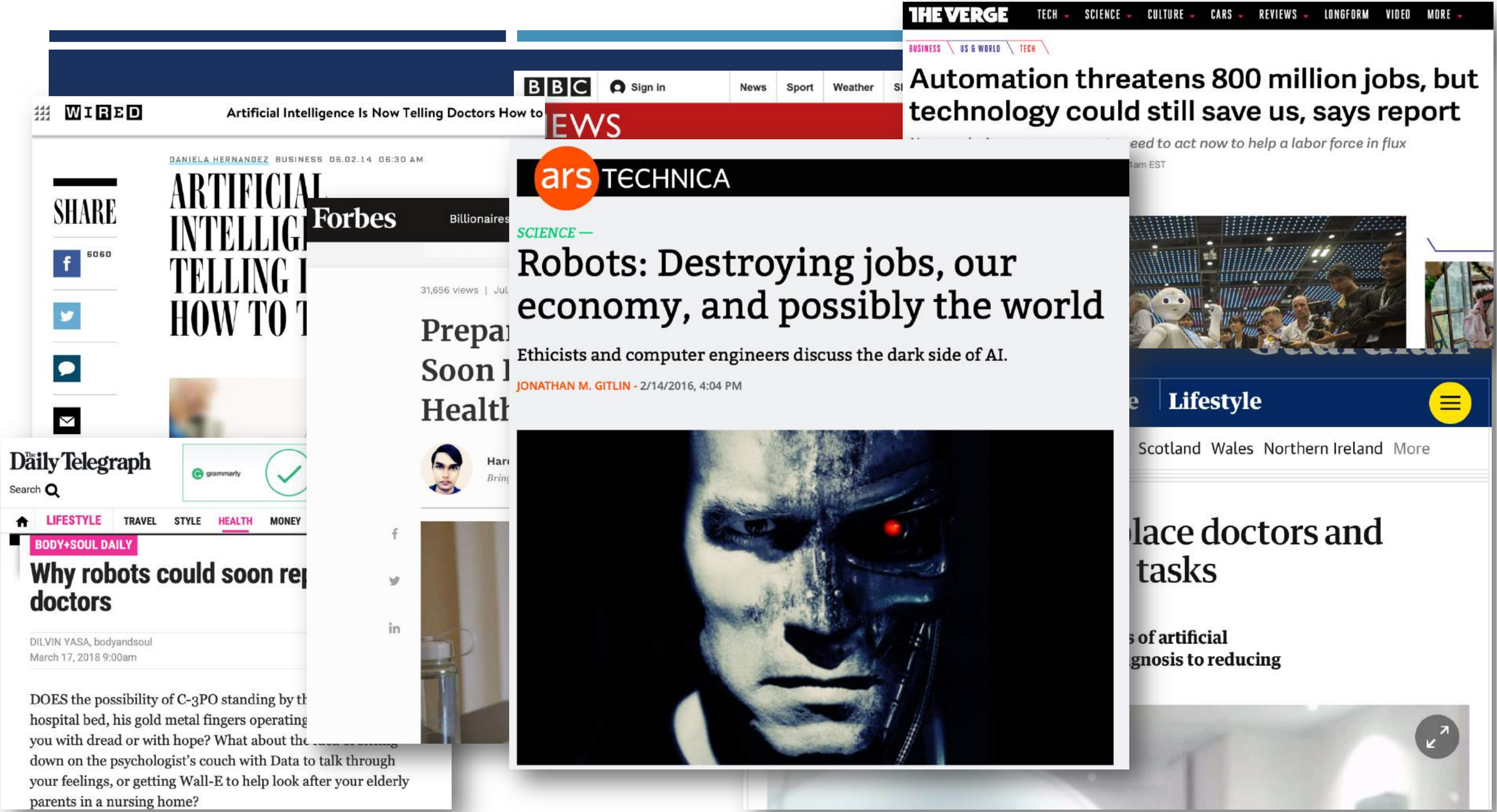
- Graber et al. (2005) studied one hundred cases of **diagnostic error** involving internists …
  - **Cognitive factors** contributed to 74% of cases.
    - Most common cause: 'premature closure'.

- Eddy (1990) showed top surgeons descriptions of surgical problems and asked: *Should the patient have surgery?*
  - 50% said **Yes**, 50% said **No**.
  - 40% gave conflicting answers upon retesting.

UNIVERSITY OF TORONTO

Graber et al. (2005) Diagnostic Error in Internal Medicine. Arch Intern Med., 165(13):1493-1499
Eddy (1990) The Challenge. JAMA, 263(2):287-290. http://jama.jamanetwork.com/article.aspx?articleid=380215


"I think that if you work as a radiologist you are like Wile E. Coyote in the cartoon.
You're already over the edge of the cliff, but you haven't yet looked down. …
It's just completely obvious that in five years deep learning is going to do better than radiologists.
It might be ten years."
- Geoff Hinton

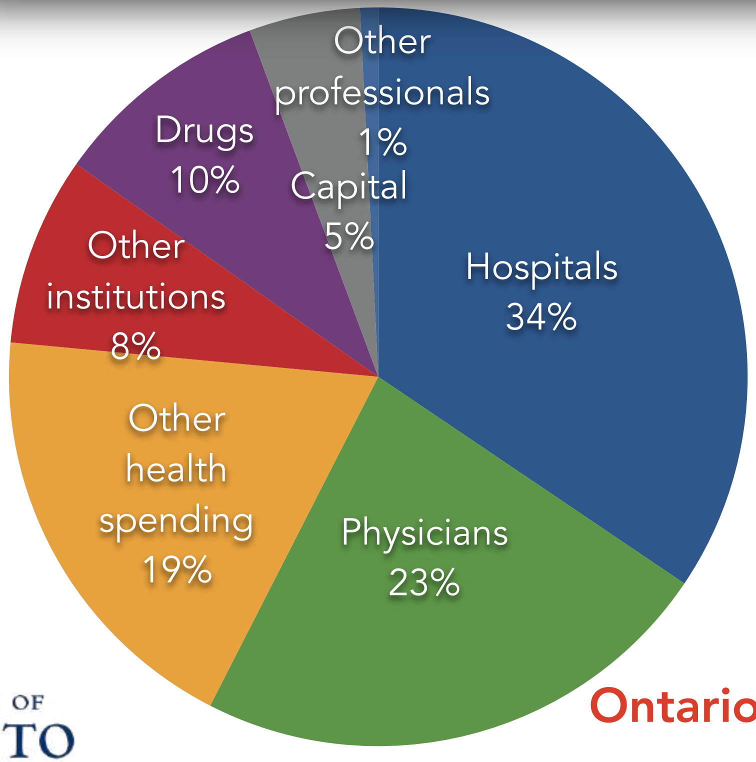


THE VERGE
TECH SCIENCE CULTURE CARS REVIEWS LONGFORM VIDEO MORE
BUSINESS US & WORLD TECH
Automation threatens 800 million jobs, but technology could still save us, says report
eed to act now to help a labor force in flux
BBC
Sign in
News
Sport
Weather
EWS
WIRED
Artificial Intelligence Is Now Telling Doctors How to
DANIELA HERNANDEZ BUSINESS 06.02.14 06:30 AM
SHARE
6060
ARTIFICIAL INTELLIG TELLING HOW TO
Forbes
Billionaires
31,656 views | Jul
Prepa Soon Healt
ars TECHNICA
SCIENCE —
Robots: Destroying jobs, our economy, and possibly the world
Ethicists and computer engineers discuss the dark side of AI.
JONATHAN M. GITLIN - 2/14/2016, 4:04 PM
Lifestyle
Scotland Wales Northern Ireland More
lace doctors and tasks
s of artificial
gnosis to reducing
The Daily Telegraph
Search
grammarly
LIFESTYLE TRAVEL STYLE HEALTH MONEY
BODY+SOUL DAILY
Why robots could soon rep doctors
DILVIN YASA, bodyandsoul
March 17, 2018 9:00am
DOES the possibility of C-3PO standing by th hospital bed, his gold metal fingers operating you with dread or with hope? What about the down on the psychologist's couch with Data to talk through your feelings, or getting Wall-E to help look after your elderly parents in a nursing home?

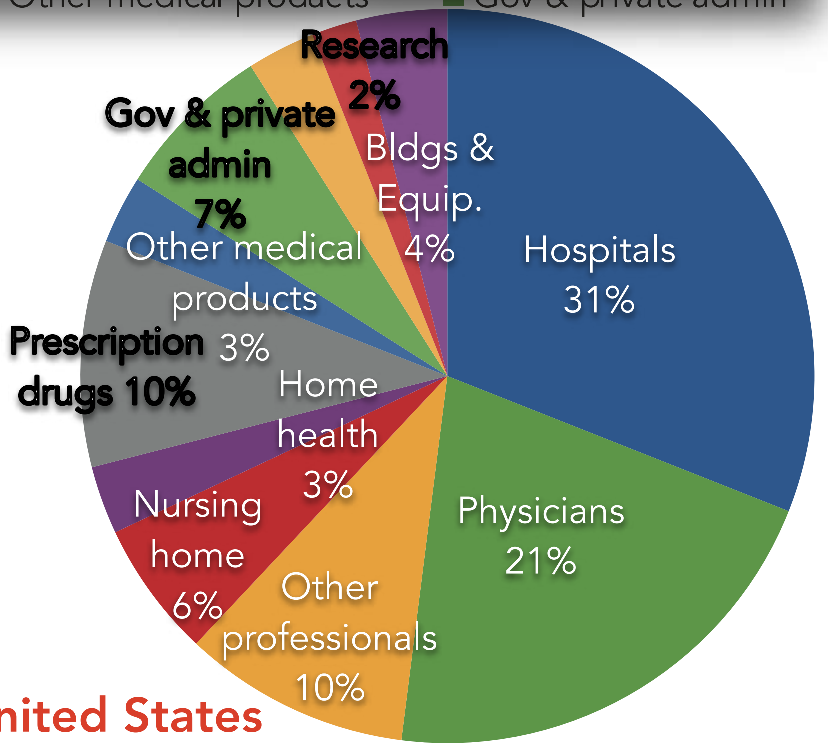
THE REAL FUTURE

WHERE WILL CHANGE HAPPEN?

"from a solely hospital-centred system [towards] a community [primary care] system"
Premier Kathleen Wynne, 2017



Ontario



United States

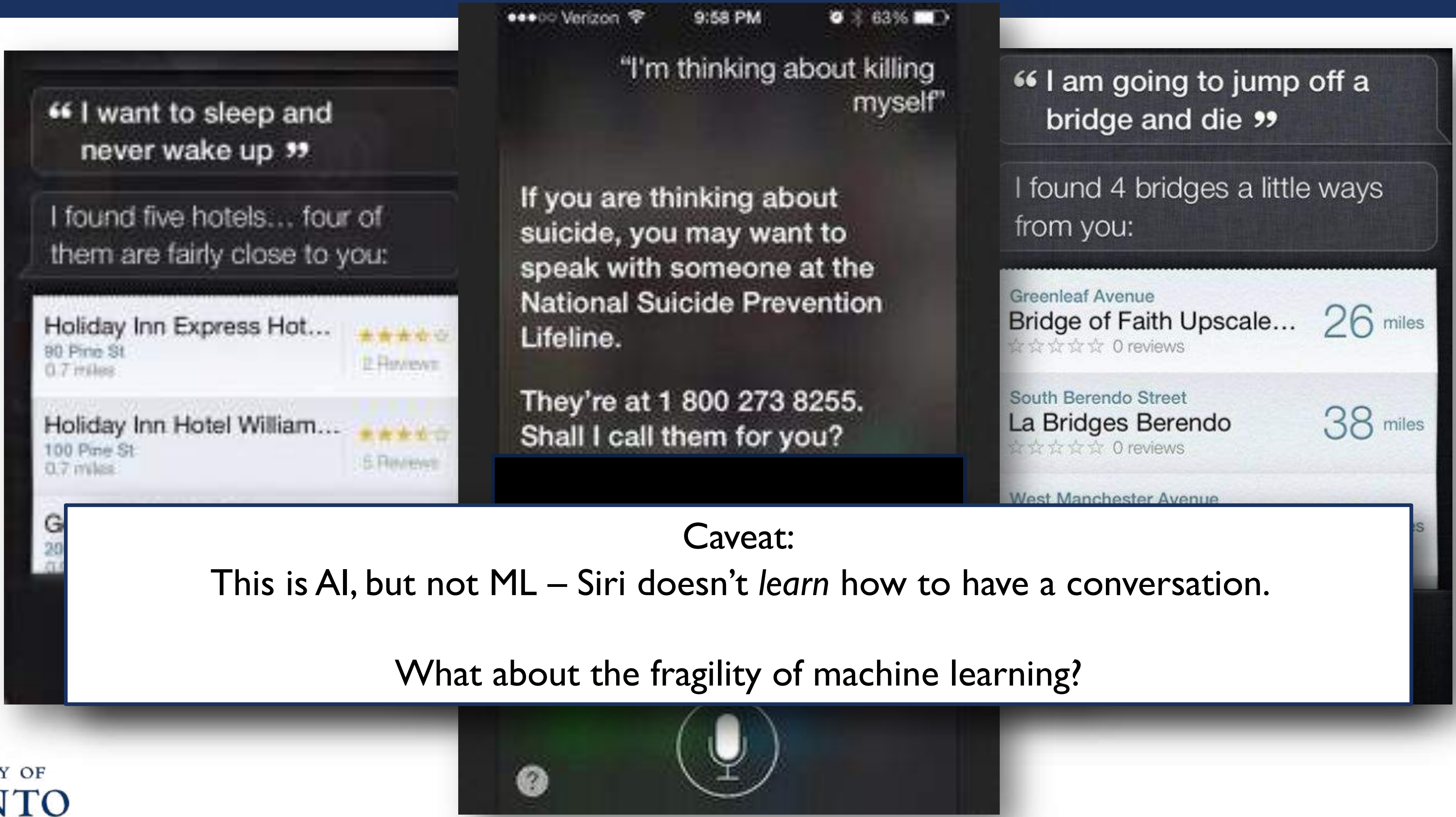
National Health Expenditure Trends, Canadian Institute for Health Information, 2010

Department of Health and Human Services, 2011

SYMPATHY FROM THE ANVIL

UNIVERSITY OF TORONTO

SYMPATHY FROM THE ANVIL

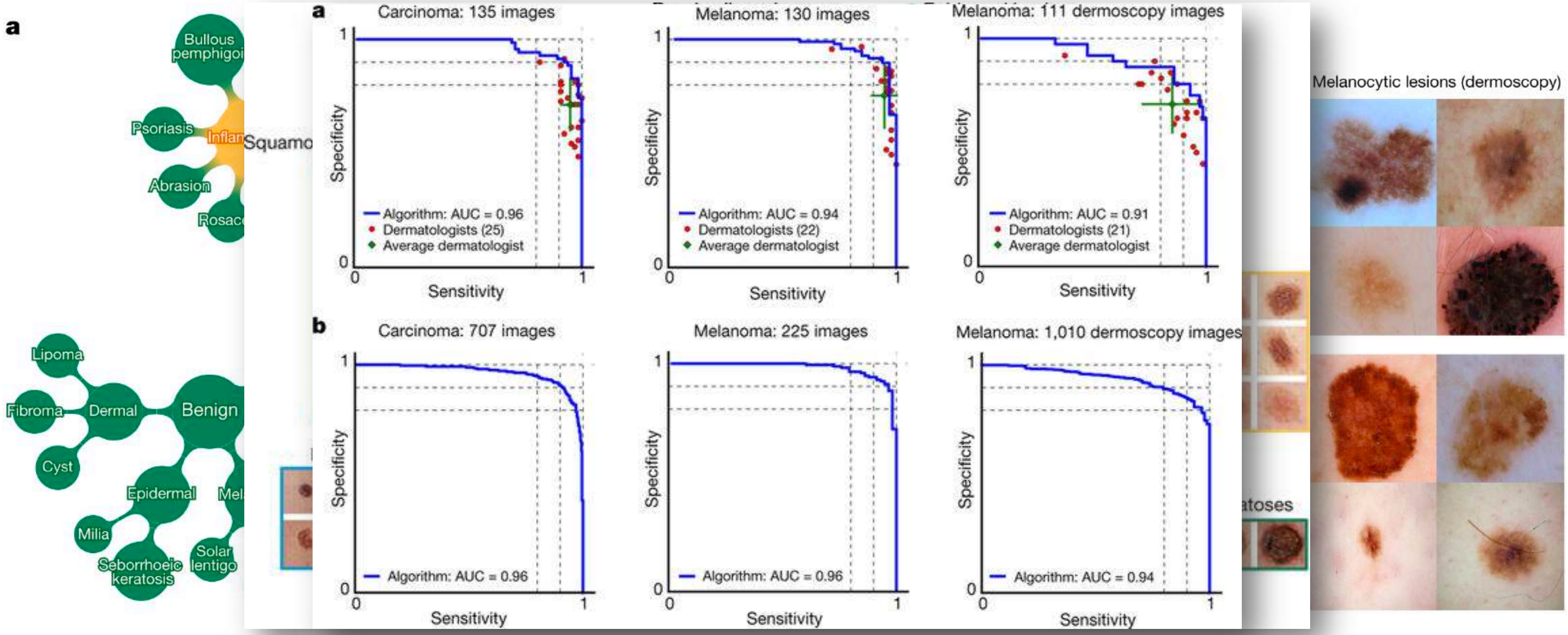


Caveat:

This is AI, but not ML – Siri doesn't *learn* how to have a conversation.

What about the fragility of machine learning?

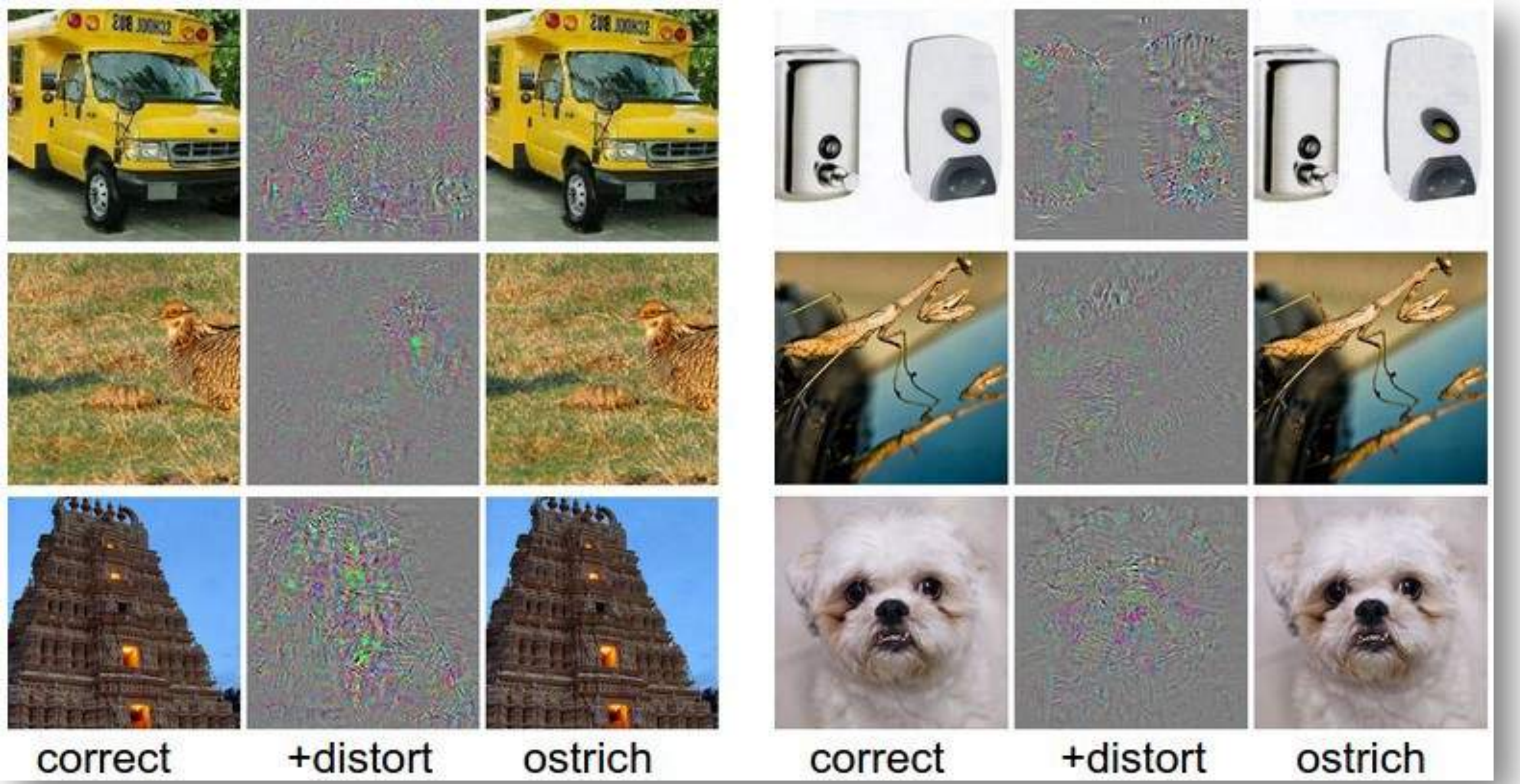
AUTOMATED DIAGNOSES



Trained with 129,450 clinical images
Tested against 21 certified dermatologists.

Esteva A, Kuprel B, Novoa RA, *et al.* (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**:115-118

NEURAL NETWORKS CAN BE FOOLED



Nguyen A, Yosinski J, Clune J. (2015) Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *Proc. of IEEE CVPR*. 427–36.

RISKS OF ARTIFICIAL INTELLIGENCE

THE BRITTLENESS OF SAFETY

1. There *is* a risk that AI in the wrong hands, or in those of a select few will:
   a) perform tasks that may not be 'globally optimal', or
   b) change the nature of work in unexpected, adverse ways.

2. A *bigger* risk is that AI in the *right* hands will:
   a) lazily be given goals that are too abstract,
   b) find a 'trick' to achieve those goals that we don't understand, and
   c) result in unexpected, uninterpretable behaviour

CONCRETE PROBLEMS IN AI SAFETY

Techniques to promote *safe* use of AI that are not always followed, e.g.:

1. Avoiding **negative side effects**
2. Avoiding **reward hacking**
3. Ensuring **scalable oversight**
4. Ensuring robustness to **distributional shift**
5. Ensuring **safe exploration**

UNIVERSITY OF TORONTO

Amodei D, Olah C, Steinhardt J, *et al.* (2016) Concrete Problems in AI Safety. arXiv:1606.06565v2, pp 1–29. doi:1606.06565

AVOID NEGATIVE SIDE EFFECTS

1. Include an '*impact regularizer*' that penalizes change to the environment.
    1. But how does the system represent change?
2. *Penalize influence*.
    1. I.e., limit the amount/scope of resources available
    2. But how does the system represent empowerment?
    3. Do you penalize the AI if it *can* take an action, or if it *does*?

UNIVERSITY OF TORONTO

Amodei D, Olah C, Steinhardt J, *et al.* (2016) Concrete Problems in AI Safety. arXiv:1606.06565v2, pp 1–29. doi:1606.06565

AVOID 'REWARD HACKING'

1. *<u>Abstract rewards</u>*. Avoid the curse of dimensionality, especially with misbehaving numerical dimensions.
2. *<u>Avoid Goodhart' Law.</u>* ("when a metric is used as a target, it ceases to be a good metric").
    1. E.g., avoid this logic: "if I increase prescriptions, patient admissions decrease, ∴ maximize prescriptions!"

UNIVERSITY OF TORONTO

Amodei D, Olah C, Steinhardt J, *et al.* (2016) Concrete Problems in AI Safety. arXiv:1606.06565v2, pp 1–29. doi:1606.06565

SCALABLE OVERSIGHT & DISTRIBUTIONAL SHIFT

1. A model trained on few examples might not scale well.
2. A model trained to regress to the mean, might not capture rare events
3. *Active learning* may help.
   1. Continuously rely on human consensus and input; validate 'difficult' data.
4. A model must acknowledge its own ignorance, and resist shifting its parameters too hastily.
   1. See '*canary deployment*' methodology (e.g., in KubeFlow) regarding '*safe exploration*'

UNIVERSITY OF TORONTO

Amodei D, Olah C, Steinhardt J, *et al.* (2016) Concrete Problems in AI Safety. arXiv:1606.06565v2, pp 1–29. doi:1606.06565

THE WANTS AND NEEDS OF EXPLAINABLE AI

- We ***want*** ML to be explainable:
  - To identify and remove bias to promote **safety**
  - To leverage **domain expertise** and induce **new knowledge**
  - To ensure **generalizability** and **consistency**
  - To audit and **trust** the system

- We ***need*** ML to be explainable:
  - For regulatory approval process (e.g., FDA)
  - For the 'right to explanation' (e.g., GDPR)

UNIVERSITY OF TORONTO

CONCRETE PROBLEMS IN AI SAFETY

Techniques to promote *safe* use of AI that are not always followed, e.g.:

1. Avoiding **negative side effects**
2. Avoiding **reward hacking**
3. Ensuring **scalable oversight**
4. Ensuring robustness to **distributional shift**
5. Ensuring **safe exploration**
6. … Ensuring decisions are **explainable**?

UNIVERSITY OF TORONTO

Amodei D, Olah C, Steinhardt J, *et al.* (2016) Concrete Problems in AI Safety. arXiv:1606.06565v2, pp 1–29. doi:1606.06565

IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received August 5, 2018, accepted September 4, 2018, date of publication September 17, 2018, date of current version October 12, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2870052

Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)

AMINA ADADI AND MOHAMMED BERRADA
Computer and Interdisciplinary Physics Laboratory, Sidi Mohammed Ben Abdellah University, Fez 30050, Morocco
Corresponding author: Amina Adadi (amina.adadi@gmail.com)

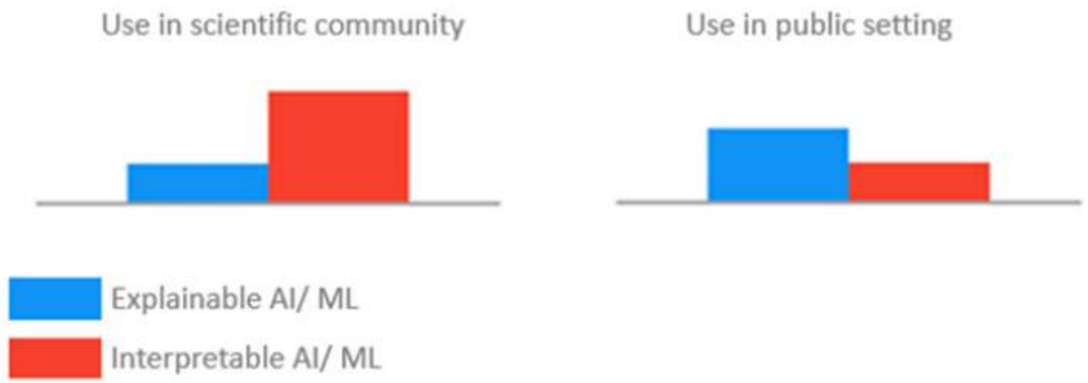


FIGURE 3. Google trends result for comparing the use of "Explainable" and "Interpretable" according to the context.

UNIVERSITY OF TORONTO

EXPLANATIONS BY LOCAL EXAMPLES

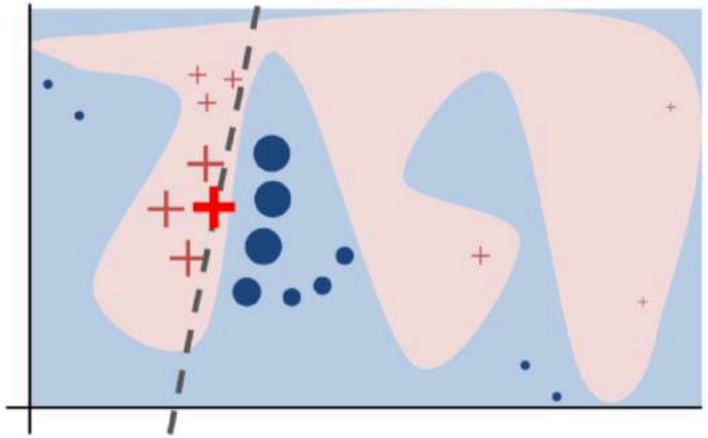


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

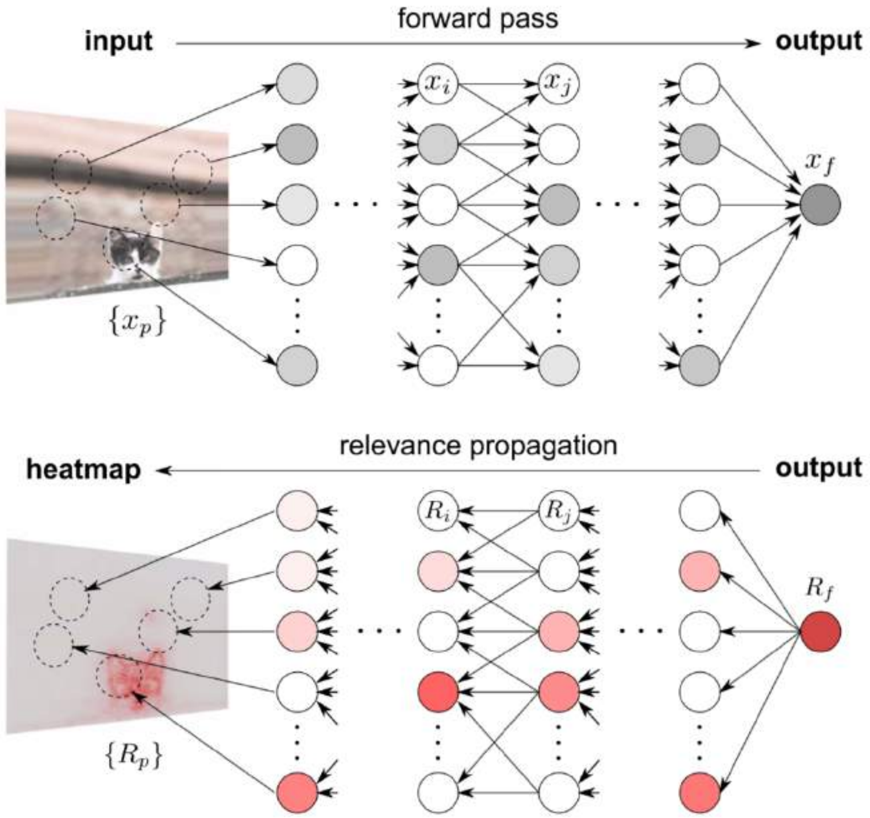
Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K
  $\mathcal{Z} \leftarrow \{\}$
  for $i \in \{1, 2, 3, ..., N\}$ **do**
    $z'_i \leftarrow sample\_around(x')$
    $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
  end for
  $w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ $\triangleright$ with z'_i as features, $f(z)$ as target
  return w

UNIVERSITY OF TORONTO

Ribeiro MT, Singh S, Guestrin C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. 2016. doi:10.1145/1235

EXPLANATIONS BY RELEVANCE



First-order Taylor decomposition

$$f(\mathbf{x}) = f(\tilde{\mathbf{x}}) + \left(\frac{\partial f}{\partial \mathbf{x}} \Big|_{x=\tilde{x}} \right)^{\top} \cdot (\mathbf{x} - \tilde{\mathbf{x}}) + \varepsilon = 0 + \sum_p \underbrace{\frac{\partial f}{\partial x_p} \Big|_{x=\tilde{x}} \cdot (x_p - \tilde{x}_p)}_{R_p(\mathbf{x})} + \varepsilon,$$

$$R_j = \left(\frac{\partial R_j}{\partial \{x_i\}} \Big|_{\{\tilde{x}_i\}^{(j)}} \right)^{\top} \cdot (\{x_i\} - \{\tilde{x}_i\}^{(j)}) + \varepsilon_j = \sum_i \underbrace{\frac{\partial R_j}{\partial x_i} \Big|_{\{\tilde{x}_i\}^{(j)}} \cdot (x_i - \tilde{x}_i^{(j)})}_{R_{ij}} + \varepsilon_j,$$

Deep Taylor decomposition of ‘relevance’ at neuron *j*

Montavon G, Lapuschkin S, Binder A, et al. Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognit 2017;65:211–22.

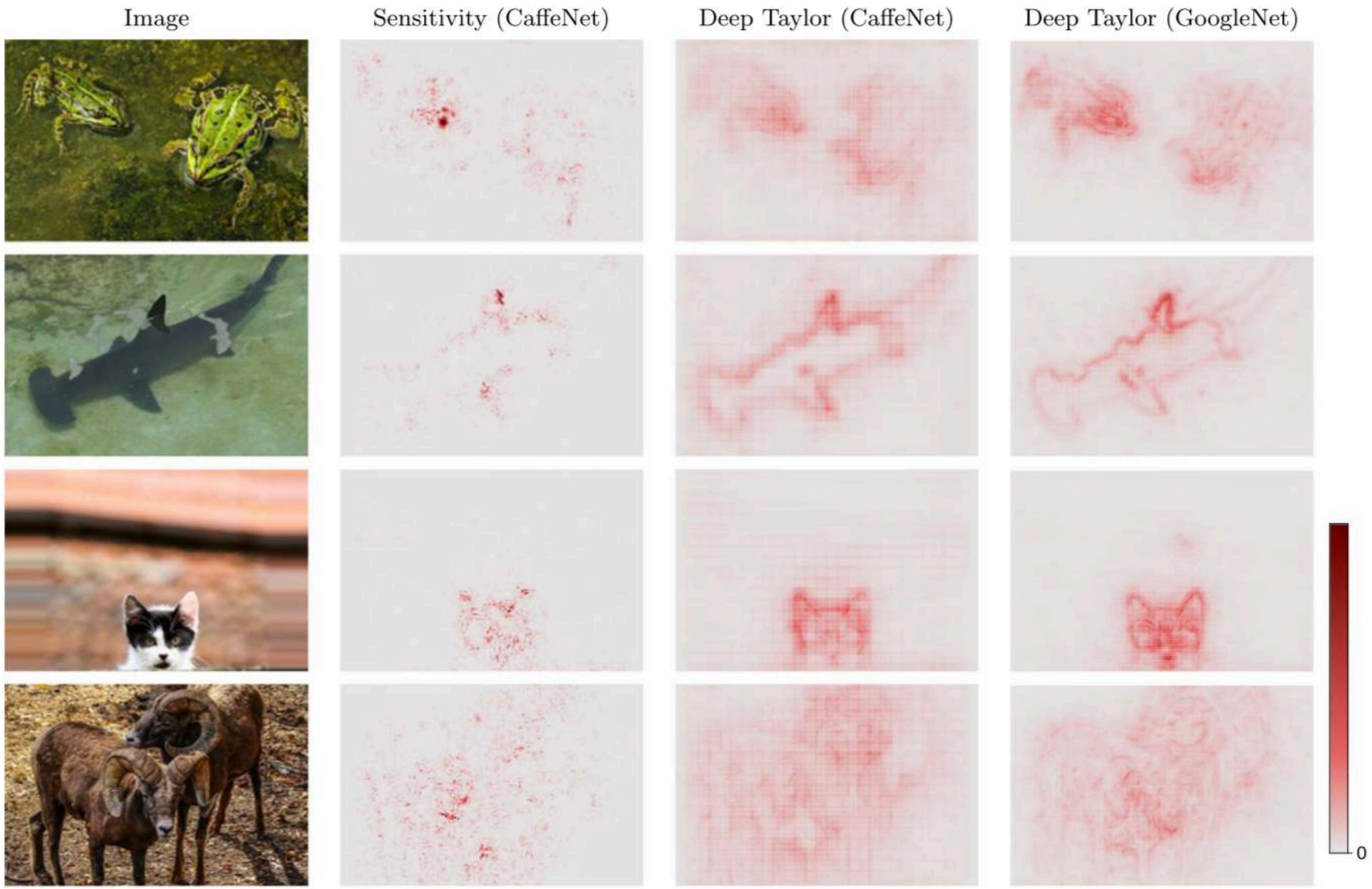


Fig. 7. Images of different ILSVRC classes ("frog", "shark", "cat", and "sheep") given as input to a deep network, and displayed next to the corresponding heatmaps. Heatmap scores are summed over all color channels of the image.

UNIVERSITY OF TORONTO

Montavon G, Lapuschkin S, Binder A, et al. Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognit 2017;65:211–22.

U.S. FOOD & DRUG ADMINISTRATION

September 11, 2018

subject to the following special controls:

1. Clinical [testing] under anticipated conditions of use must demonstrate…:
   1. The ability to obtain an ECG of sufficient quality for display and analysis; and
   2. **The performance characteristics of the detection algorithm as reported by sensitivity and either specificity or positive predictive value**.
2. **Software verification, validation, and hazard analysis must be performed**. Documentation must include a characterization of the technical specifications of the software, including the **detection algorithm and its inputs and outputs**.
3. Non-clinical performance testing must **validate** detection algorithm **performance using a previously adjudicated data set**.
4. Human factors and usability testing must demonstrate the following:
   1. The user can correctly use the device based solely on reading the device labeling; and
   2. **The user** can correctly **interpret the device output and understand when to seek medical care**.
5. …

FDA concludes that this device should be classified into Class II. This order, therefore, classifies the ECG App, and substantially equivalent devices of this generic type, into Class II under the generic name electrocardiograph software for over-the-counter use.

FDA identifies this generic type of device as:

U.S. Food & Drug Administration
10903 New Hampshire Avenue
Silver Spring, MD 20993
www.fda.gov

UNIVERSITY OF TORONTO

RIGHT TO EXPLANATION

- EU General Data Protection Regulation (enacted 2016), extends the automated decision-making rights in the **1995 Data Protection Directive** to provide a **right to an explanation**, in Recital 71:

> **The data subject should have the right** not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based **solely** on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention.
>
> …
>
> [S]uch processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view**, to obtain an explanation of the decision** reached after such assessment and to challenge the decision.

- Note: recitals are *not* binding
- However, to pretend that explainability won't be a part of AI in practice is to 'play make believe'.

UNIVERSITY OF TORONTO

XAI FOR SURGERY

Surgical Innovation

September 11, 2019

Explainable Artificial Intelligence for Safe Intraoperative Decision Support

Lauren Gordon, MD, MSc[1,2]; Teodor Grantcharov, MD, PhD[1,2]; Frank Rudzicz, PhD[1,3]

Author Affiliations

JAMA Surg. 2019;154(11):1064-1065. doi:10.1001/jamasurg.2019.2821

Machine Learning Website

Intraoperative adverse events are a common and important cause of surgical morbidity.[1,2] Strategies to reduce adverse events and mitigate their consequences have traditionally focused on surgical education, structured communication, and adverse event management. However, until now, little could be done to anticipate these events in the operating room. Advances in both data capture in the operating room and explainable artificial intelligence (XAI) techniques to process these data open the way for real-time clinical decision support tools that can help surgical teams anticipate, understand, and prevent intraoperative events.

UNIVERSITY OF TORONTO

Gordon L, Grantcharov T, Rudzicz F. (2019) Explainable Artificial Intelligence for Safe Intraoperative Decision Support. JAMA Surg. **154**(11):1064-1065. . doi:10.1001/jamasurg.2019.2821

OR BLACK BOX®

Video, audio of OR

Endoscopic camera and wearable tech

Patient physiology

Environmental factors

Devices

Assessments of risks/hazards

Analysis of safety threats and resilience

Assessment of surgical team performance

Assessment of surgical technology

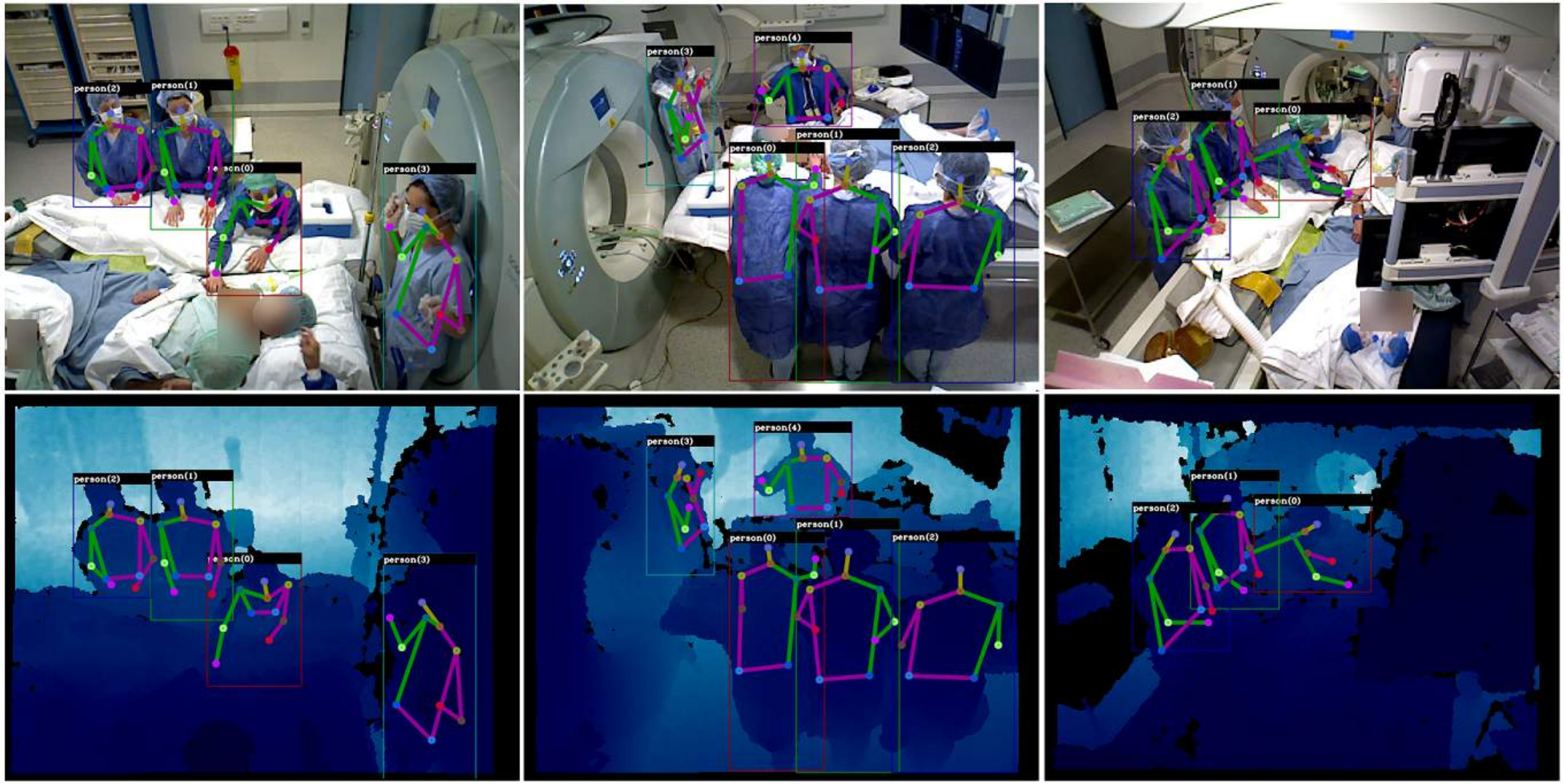
Assessment of efficiency

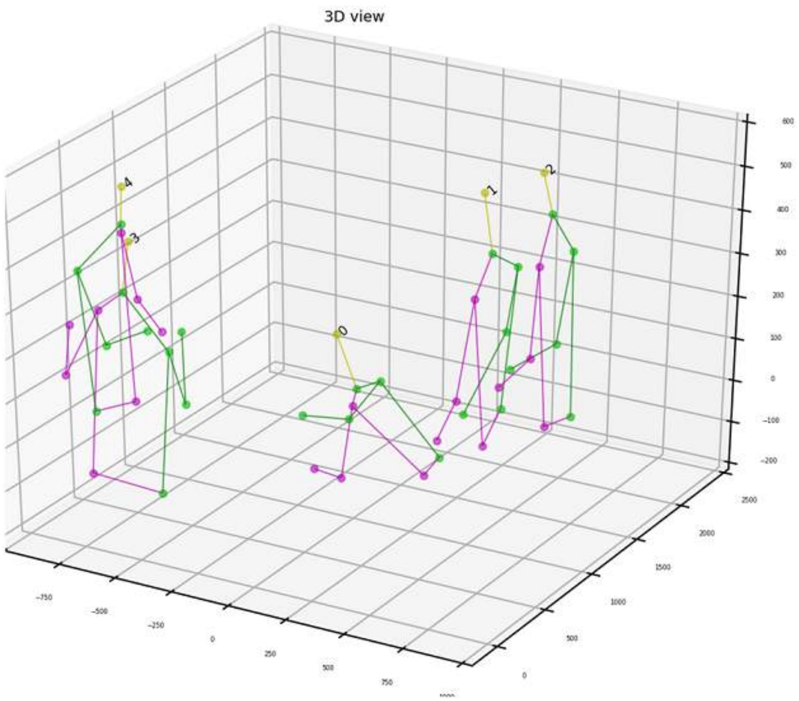
INPUT

ANALYSIS

OUTPUT

TEAM
EFFICIENCY AND NON-TECHNICAL SKILLS





SURGICAL SAFETY
TECHNOLOGIES

Privacy versus artificial intelligence in medicine

Taryn J Rohringer (BMSc)[1]; Akshay Budhkar (BASc)[2,5]; Frank Rudzicz (PhD)[2,3,4,5]

[1]Faculty of Medicine, University of Toronto, Medical Sciences Building, 1 King's College Circle, Toronto, ON, Canada, M5S 1A8.

The same month that GDPR came into effect, Canada issued new guidance for the Personal Information Protection and Electronic Documents Act (PIPEDA) … subsection 5(3) of PIPEDA states that "An organization may collect, use or disclose personal information only for purposes that **a reasonable person** would consider are appropriate in the circumstances." Given that consensus has not been widely achieved with regards to the details of surveillance of this type (e.g., what risks to personal information are necessary, given the technology, to achieve some perceived benefit to the person involved), it is not yet clear what a "**reasonable person** would consider appropriate."

and potential re-identification of patient data. This paper outlines these challenges and suggests some open questions and potential solutions. Given recent news of companies overstepping their bounds in the pursuit of patient data to train their systems, and new regulations around privacy of those data, this discussion is especially pertinent. Here, we suggest that a common good can be achieved in which data can be kept private while also useful for artificial intelligence in the practice of medicine.

Introduction

Recent advances in artificial intelligence (AI) have accelerated their use in healthcare, from remote monitoring and wearables to clinical decision support.[1]

(PIPEDA) detailing guidelines for obtaining meaningful consent, and against "[s]urveillance by an organization through audio or video functionality of the individual's own device."[4] More specifically, subsection 5(3) of PIPEDA states that "An organization may collect, use or disclose personal information only for purposes that a reasonable person would consider are appropriate in the circumstances." Given that consensus has not been widely achieved with regards to the details of surveillance of this type (e.g., what risks to personal information are necessary, given the technology, to achieve some perceived benefit to the person involved), it is not yet clear what a "reasonable person would consider appropriate."

As AI is increasingly integrated into clinical practice, various challenges will persist (e.g. data acquisition, reporting, and re-identification) and these emphasize a potential struggle between patient privacy and the promise of these systems.

Challenges to Data Acquisition

Personal health data is extremely valuable; for example, the $6 billion acquisition of Medco Containment Services by Merck was

UNIVERSITY OF TORONTO

OR BLACK BOX®

DE-IDENTIFIED BY DESIGN

SURGICAL SAFETY
TECHNOLOGIES

The following de-identification demostration was done in a simulation environment

ETHICS OF ARTIFICIAL INTELLIGENCE IN SURGERY 14

Frank Rudzicz and Raeid Saqur

HIGHLIGHTS

- The 4 key principles of biomedical ethics from a surgical context are autonomy, nonmaleficence, beneficence, and justice.
- Implications of fairness and the taxonomy of algorithmic bias in artificial intelligence (AI) system design are important factors in the ethics of AI.
- The ethical paradigm shifts as the degree of autonomy in AI agents evolves.
- Ethics in AI is dynamic, and continuous revisions are needed as AI evolves.

INTRODUCTION

Surgery manifests in an intense form of practical ethics. The practice of surgery often forces unique ad hoc decisions based on contextual intricacies in the moment, which are not typically captured in broad, top-down, or committee-approved guidelines. Surgical ethics are principled, of course, but also pragmatic. They are also replete with moral contradictions and uncertainties; the introduction of novel technology into this environment can potentially increase those challenges.

A discussion about ethics is often a discussion about choice. Wall et al[1] defined an ethical problem as "when an agent must choose between mutually exclusive options, both of which either have equal elements of right and wrong, or are perceived as equally obligatory. The essential element that distinguishes an ethical problem from a tragic situation is the element of choice." Moreover, choosing between options often involves identifying factors by which those options are *not* exactly equal, and the method one uses to weigh these factors can draw upon a set of ethical

https://arxiv.org/abs/2007.14302

UNIVERSITY OF TORONTO

AI IN HEALTHCARE

- AI and ML are maturing to a point where they can be put into practice.
- There is a strong pull in healthcare for automation, generally, and for tools to improve safety, specifically and objectively.
- As our tools are designed to improve safety in healthcare,
  we must also ensure that the tools themselves are safe.

UNIVERSITY OF TORONTO


First we shape our tools,
and thereafter our tools
shape us.
- *Marshall McLuhan*
Thank you