NRC·CNRC

NRC.CANADA.CA

AI Ethics in a research context: current approaches and future challenges

Kathleen Fraser, AI Researcher, NRC Digital Technologies
Terrence Stewart, Senior Research Officer, NRC Digital Technologies
Stephen Downes, Senior Research Officer, NRC Digital Technologies
Margaret McKay, Program Leader Digital Privacy & Security, NRC

CAREB 2024

Part 1: (14:10 – 14:55 EST)

a) How should REB members expect to encounter AI ethics questions? (M. McKay)
b) Understanding the ethics-relevant qualities of AI (T. Stewart)
c) Overview of leading AI ethics approaches and perspectives (S. Downes)
d) Practical examples of AI ethics challenges (K. Fraser)

Part 2: (15:05 – 15:50 EST)

a) Breakout sessions: Applying AI ethics principles for REB reviews in specific disciplines (facilitated small group discussions)
b) Current and future AI ethics challenges for REBs, resources available (S. Downes)

Part 1(a)

REB Encounters with AI Ethics:
an introduction

Margaret McKay
Digital Technologies Research Centre
National Research Council Canada

AI Ethics in REB Context

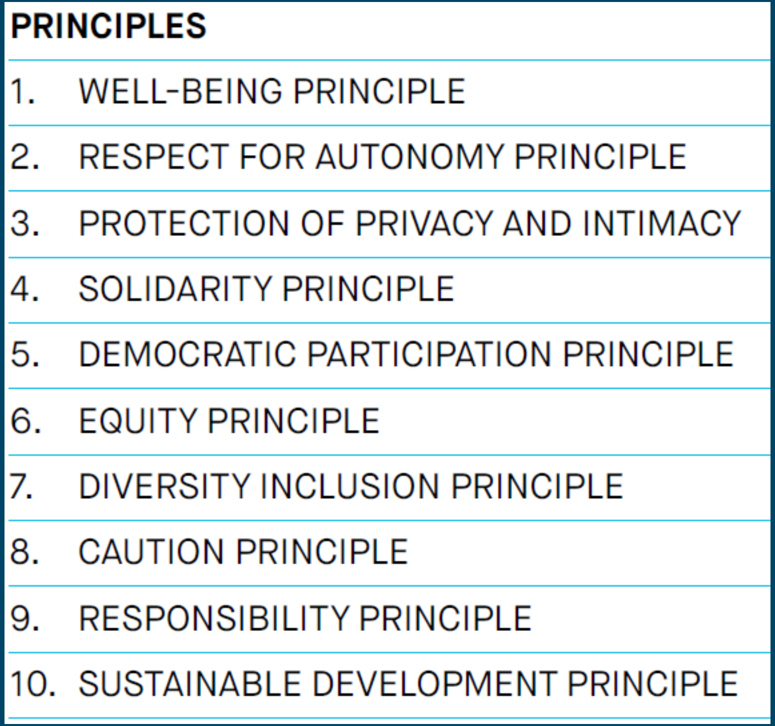
Sources of Risk

Impact on Research

	Research Data	AI Models/Tools
Research Participant Well-being	Privacy, consent, use of publicly-accessible data	Direct harm to participant, privacy risks
Proportionality (Value of Research)	Flawed data leads to flawed results	Inappropriate AI tool use leads to flawed results

3 EXAMPLE APPROACHES:
AI ETHICS/ RESPONSIBLE AI FRAMEWORKS & PRINCIPLES

The Montreal Declaration For Responsible AI Development



PRINCIPLES

1. WELL-BEING PRINCIPLE
2. RESPECT FOR AUTONOMY PRINCIPLE
3. PROTECTION OF PRIVACY AND INTIMACY
4. SOLIDARITY PRINCIPLE
5. DEMOCRATIC PARTICIPATION PRINCIPLE
6. EQUITY PRINCIPLE
7. DIVERSITY INCLUSION PRINCIPLE
8. CAUTION PRINCIPLE
9. RESPONSIBILITY PRINCIPLE
10. SUSTAINABLE DEVELOPMENT PRINCIPLE

The Montréal Declaration for responsible AI development has three main objectives:

1. Develop an ethical framework for the development and deployment of AI;

2. Guide the digital transition so everyone benefits from this technological revolution;

3. Open a national and international forum for discussion to collectively achieve equitable, inclusive, and ecologically sustainable AI development.

https://declarationmontreal-iaresponsable.com/wp-content/uploads/2023/04/UdeM_Decl-IA-Resp_LA-Declaration-ENG_WEB_09-07-19.pdf

OECD Principles for responsible stewardship of trustworthy AI

Principles:

i. Inclusive growth, sustainable development and well-being
ii. Human-centred values and fairness
iii. Transparency and explainability
iv. Robustness, security and safety
v. Accountability.

Recommendations:

…a) Governments should consider long-term public investment, and encourage private investment, in **research and development**, including interdisciplinary efforts, **to spur innovation in trustworthy AI** that focus on challenging technical issues **and on AI-related social, legal and ethical implications and policy issues.**
b) Governments should also consider public investment and encourage private investment in **open datasets** that are representative and **respect privacy and data protection** to support an environment for AI research and development that is **free of inappropriate bias** and to **improve interoperability** and **use of standards**.

https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Hiroshima AI Process

We, the Leaders of the Group of Seven (G7), stress the innovative opportunities and transformative potential of advanced Artificial Intelligence (AI) systems, in particular, foundation models and generative AI. We also recognize the need to manage risks and to protect individuals, society, and our shared principles including the rule of law and democratic values, keeping humankind at the center….

https://www.mofa.go.jp/ecm/ec/page5e_000076.html

Eleven Principles and Actions, including:

1. Take appropriate measures throughout the development of advanced AI systems, including prior to and throughout their deployment and placement on the market, to identify, evaluate, and mitigate risks across the AI lifecycle.
4. Work towards responsible information sharing and reporting of incidents among organizations developing advanced AI systems including with industry, governments, civil society, and academia.
8. Prioritize research to mitigate societal, safety and security risks and prioritize investment in effective mitigation measures.
11. Implement appropriate data input measures and protections for personal data and intellectual property

SO…HOW DOES ALL THIS FIT WITH THE TCPS2?

Applying Guidance – Proportionality:

TCPS2 requires a proportional approach to ethics review. This includes balancing the foreseeable risks, as well as the potential benefits, of the proposed research.

- **Data issues:** AI-enabled research which uses biased or otherwise flawed training or validation data can offer little to no evidence of potential benefits.
- **AI Issues:** A research team using third-party AI tools without adequate AI-specific expertise can inadvertently obtain seemingly convincing but actually erroneous results which can delay advances in the field and potentially lead to wider harms.
- **Security issues (data and / or AI model):** where data and models are not adequately protected form hostile interference, even a well-designed project can be manipulated to produce useless or even harmful results.

Applying Guidance – Research Participant Risks

Data Collection and Storage:

- Consent, repository requirements, third-party sourced data
- Security of data during collection, use, and storage

AI Model Privacy Risks:

- Potential for model leakage of sensitive information

AI Model Risks:

- Where the research relates to the interaction of the human with an AI-enabled system, system bias, accuracy, etc. may impact research subject well-being

Points to Consider:

1. Responsible AI / AI Ethics texts tend to focus primarily on broader societal impacts…yet they remain relevant to REBs as well.
2. The value of research involving or relying on the development or re-training of an AI model or system is impacted by risks that the model produced will produce inaccurate or biased results – REB and broader societal interests intersect here.
3. The selection of research participants and other data sources (human or non-human) can impact the quality of the resulting AI model and therefore the value of the research
4. Direct harms to research participants can come from conventional privacy issues, and also from weaknesses in the AI model developed (e.g. leakage).
5. Ethical priorities can differ between individuals, cultures, etc. Differences of opinion on questions of AI ethics are common (even within this panel).

EXAMPLES:

ALIGNMENT BETWEEN PRINCIPLES OF MONTREAL DECLARATION ON RESPONSIBLE AI DEVELOPMENT

AND

CORE REB PRINCIPLES

TCPS2 Concerns in Montreal Declaration Terms

(i) Risks Arising from Data Used in Research:

Risks to Proportionality - Value of the Research:

- Equity: Bias - over/under representation, historical bias, labeling bias / off-shored ethics, etc.
- Prudence: threats to AI model integrity - data poisoning, label flipping, etc.

Research Participant Risks:

- Privacy: Consent, consent to secondary use, blanket vs. broad consent, limits on collection and use of publicly accessible data
- Privacy: Security of data at rest and in movement (e.g. during AI training)

(ii) Risks Arising from AI Models & Tools:

Research Participant Risks:

- Prudence: Risks of AI models leaking sensitive training data

Risks to Proportionality of the Value of the Research:

- Well-Being: Competence of research team in the use of the proposed AI tool
- Well-Being: Risks of biased / inaccurate, or otherwise inappropriate model outputs which undercut the intended research value of the project

Part 1(b)

Understanding the ethics-relevant qualities of AI

Terry Stewart
Digital Technologies Research Centre
National Research Council Canada

What is AI?

What is AI, practically?

A system that takes in input and produces corresponding output

What is AI, practically?

A system that takes in input and produces corresponding output

cat

cat

dog

dog

What is AI, practically?

A system that takes in input and produces corresponding output

cat

cat

dog

dog

My favourite animal is a	dog
I like to	go
I like to go	for
I like to go for	a
I like to go for a	walk

What is AI, practically?

A system that takes in input and produces corresponding output

cat

cat

dog

dog

My favourite animal is a	dog
I like to	go
I like to go	for
I like to go for	a
I like to go for a	walk

cat

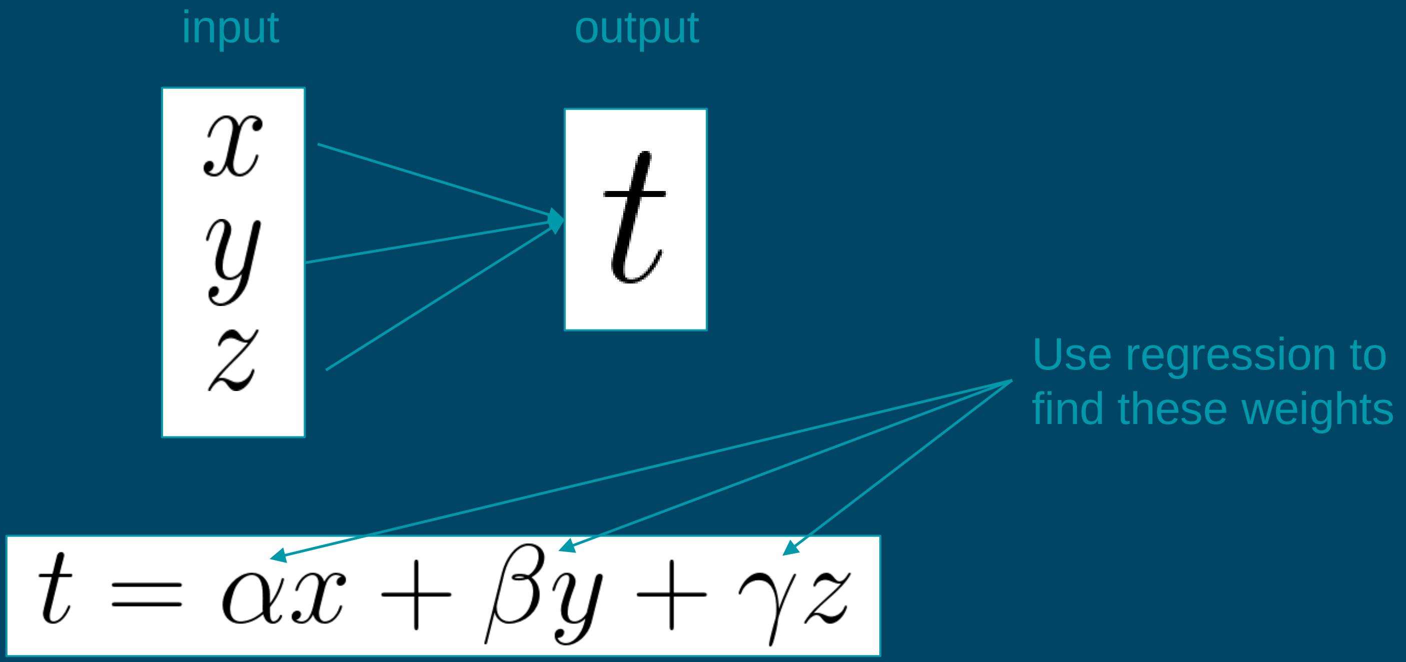
cat

dog

dog

How does AI work?

Regression, scaled up



input

x
y
z

output

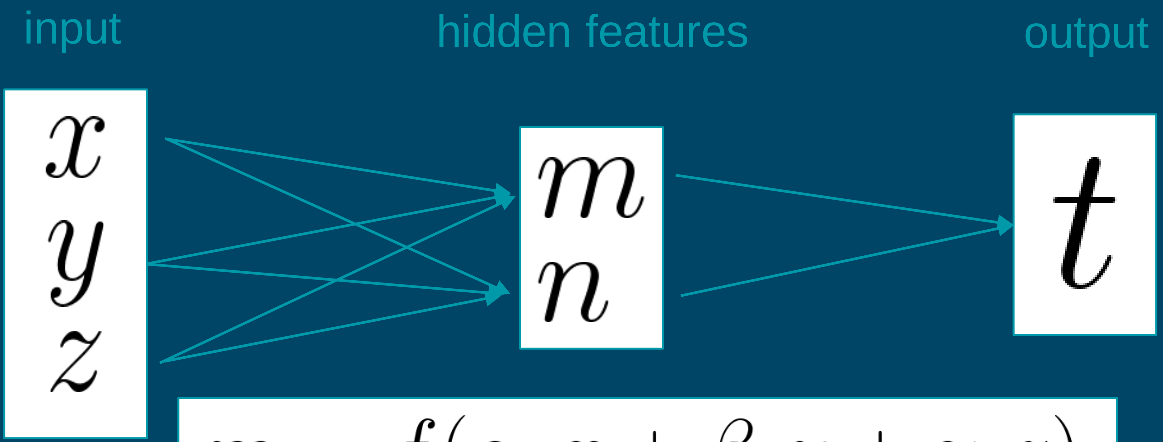
t

Use regression to find these weights

$$t = \alpha x + \beta y + \gamma z$$

How does AI work?

Regression, scaled up



$$
\begin{aligned}
m &= f(\alpha_1 x + \beta_1 y + \gamma_1 z)\\
n &= f(\alpha_2 x + \beta_2 y + \gamma_2 z)\\
t &= \lambda m + \omega n
\end{aligned}
$$

Adjust first set of weights so that the second set does a good job (“backprop”)

Relevance to ethics

When finding the weights, we are optimizing for something

- Usually mean squared error (MSE) or categorization error (accuracy)
- “loss function” or “objective”

“When learning is involved and you pick some objective function to optimize like error you should never expect to get for free anything that you didn't explicitly state in the objective and you shouldn't expect to avoid any behavior that you didn't specify should be explicitly avoided.”

“Because if you're searching some complicated model space looking for the lower error and there's some little corner of the model space where you can even incrementally infinitesimally improve your error at the expense of some social norm machine learning is going to go for that corner because that's what it does”

michael kearns + aaron roth

the ethical algorithm

the science of socially aware algorithm design

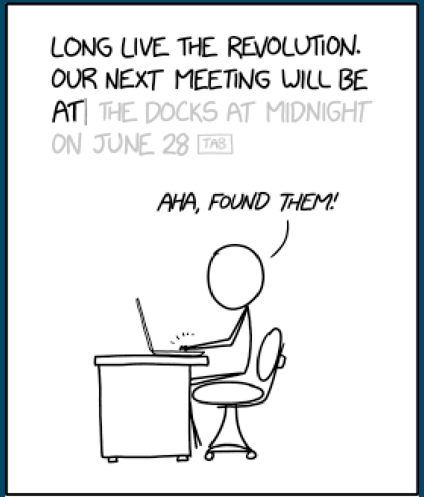
Relevance to ethics

Privacy

- Where does this data come from?

Security

- Can you recover the original training data?


LONG LIVE THE REVOLUTION. OUR NEXT MEETING WILL BE AT| THE DOCKS AT MIDNIGHT ON JUNE 28 TAB
AHA, FOUND THEM!

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

https://xkcd.com/2169

Relevance to ethics

Bias

input	output
My favourite animal is a	dog
My favourite animal is a	cat
My favourite animal is a	dog
My favourite animal is a	cat
My favourite animal is a	dog

Relevance to ethics

Bias

input	output
My favourite animal is a	dog
My favourite animal is a	cat
My favourite animal is a	dog
My favourite animal is a	cat
My favourite animal is a	dog

Accuracy if output dog 60% of the time and cat 40% of the time

$$0.6 \times 0.6 + 0.4 \times 0.4 = 0.52$$

Relevance to ethics

Bias

input	output
My favourite animal is a	dog
My favourite animal is a	cat
My favourite animal is a	dog
My favourite animal is a	cat
My favourite animal is a	dog

Accuracy if output dog 60% of the time and cat 40% of the time

$$\boxed{0.6 \times 0.6 + 0.4 \times 0.4 = 0.52}$$

Accuracy if output dog 100% of the time

$$\boxed{1 \times 0.6 + 0 \times 0.4 = 0.6}$$

Relevance to ethics

Bias

input	output
My favourite animal is a	dog
My favourite animal is a	cat
My favourite animal is a	dog
My favourite animal is a	cat
My favourite animal is a	dog

Accuracy if output dog 60% of the time and cat 40% of the time

$$0.6 \times 0.6 + 0.4 \times 0.4 = 0.52$$

Accuracy if output dog 100% of the time

$$1 \times 0.6 + 0 \times 0.4 = 0.6$$

Optimizing for accuracy can make bias in the training data even worse!

Part 1(c)

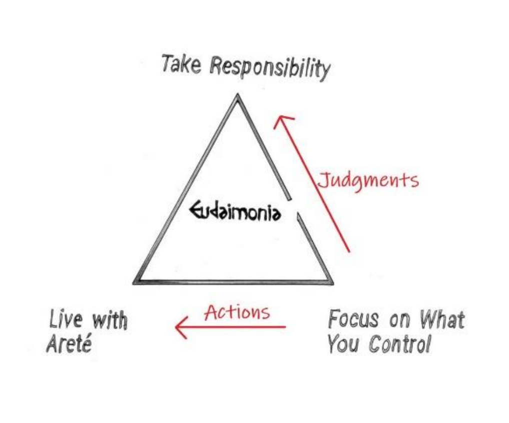
Overview of leading AI ethics approaches and perspectives

Stephen Downes
Digital Technologies Research Centre
National Research Council Canada

Leading AI ethics approaches and perspectives

Character and Virtue

- Cultivation of inherent ethical qualities such as honesty
- Arete "Be all you can be"
- Between deficiency, excess



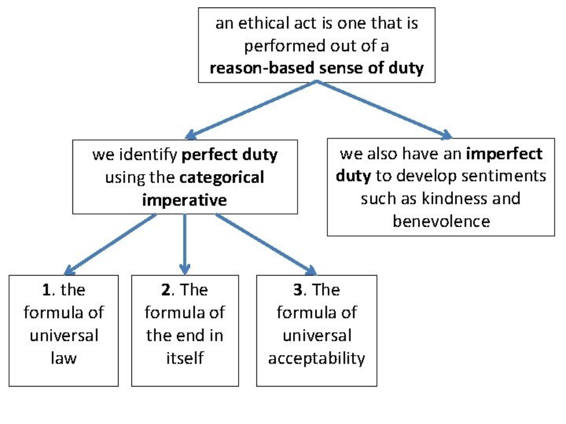
Character and Virtue in AI Ethics:

- Issues around the impact on the individual, for example, loss of skills, loss of critical reflection, loss of sense of right and wrong
- Virtue-based responses: standards of professional conduct (e.g. CFA Institute, ACM code, Nolan principles, AMA)
- Values-based: "values are academic freedom, scholarly excellence, mutual respect, collaboration, integrity" (Folan, 2020)

Leading AI ethics approaches and perspectives

Duty and Deontology

- Inherent value of humans
- Categorical Imperative – “What if everyone did that?”



an ethical act is one that is performed out of a **reason-based sense of duty**

we identify **perfect duty** using the **categorical imperative**

we also have an **imperfect duty** to develop sentiments such as kindness and benevolence

1. the formula of universal law

2. The formula of the end in itself

3. The formula of universal acceptability

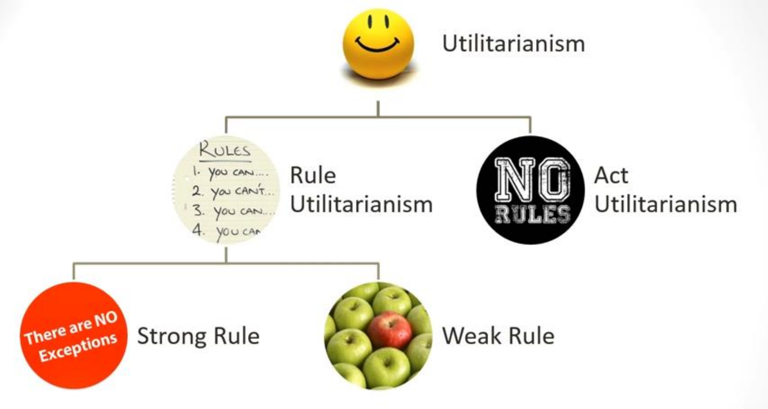
Duty and Deontology in AI Ethics:

- Issues based on rights and agency, including surveillance, tracking, anonymity, and privacy generally, human decisions
- Response based on worth & dignity of each human (NEA, 1975)
- Role and responsibility, e.g. duty to “the pursuit and dissemination of knowledge…” (SFU Code)
- Informed consent (eauchamp and Childress, Helsinki, Belmont)

Leading AI ethics approaches and perspectives

Consequentialism

- The greatest good for the greatest number of people
- The happiness principle
- Comparing acts versus rules



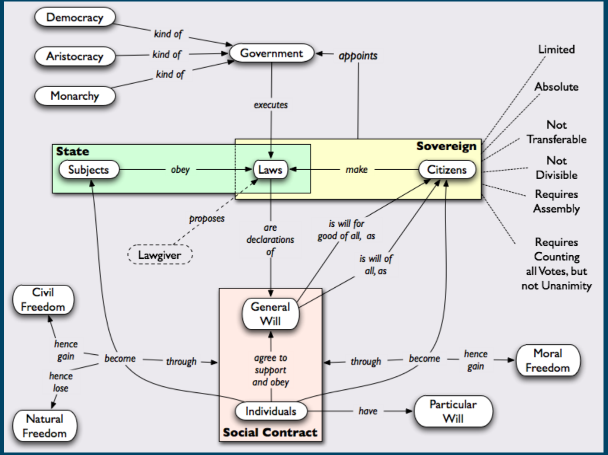
Consequentialism in AI Ethics:

- Issues related to undesirable outcomes, including bias and prejudice, misinterpretation
- Benefit-based response: “safety, health, and welfare of the public” (IEEE Code)
- Reduction of Harm (HHS Common Rule)
- Risk-based approaches

Leading AI ethics approaches and perspectives

Social Contract Theory

- Ethics as agreement for the greater good
- Rawls: ‘original position’
- Justice as fairness



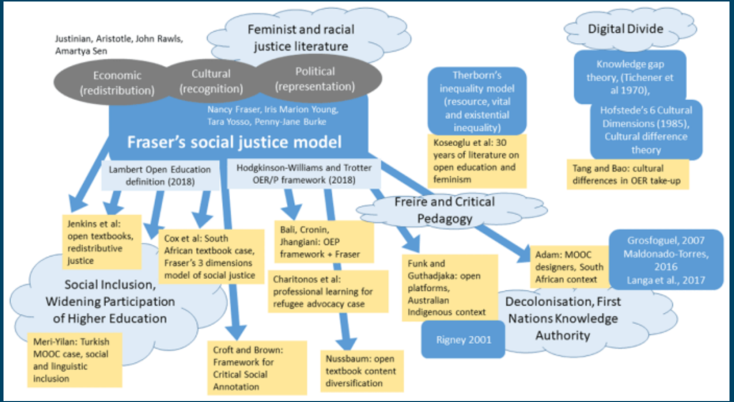
Social Contract Theory in AI Ethics

- Issues that undermine decision-making and democracy, for example, content manipulation, micro-targeting, discrimination,
- Society-based responses based in responsibility (Code Soleil)
- Promotion of public trust (Ontario College of Teachers)
- Constitutional and rights-based approaches

Leading AI ethics approaches and perspectives

Care and Community

- Ethics as based in personal relationships, interaction
- Based in a sense of ethics rather than rules & principles



Justinian, Aristotle, John Rawls, Amartya Sen

Feminist and racial justice literature

Economic (redistribution)

Cultural (recognition)

Political (representation)

Nancy Fraser, Iris Marion Young, Tara Yosso, Penny-Jane Burke

Fraser's social justice model

Lambert Open Education definition (2018)

Hodgkinson-Williams and Trotter OER/P framework (2018)

Jenkins et al: open textbooks, redistributive justice

Cox et al: South African textbook case, Fraser's 3 dimensions model of social justice

Bali, Cronin, Jhangiani: OEP framework + Fraser

Charitonos et al: professional learning for refugee advocacy case

Social Inclusion, Widening Participation of Higher Education

Meri-Yilan: Turkish MOOC case, social and linguistic inclusion

Croft and Brown: Framework for Critical Social Annotation

Nussbaum: open textbook content diversification

Therborn's inequality model (resource, vital and existential inequality)

Koseoglu et al: 30 years of literature on open education and feminism

Digital Divide

Knowledge gap theory, (Tichener et al 1970),

Hofstede's 6 Cultural Dimensions (1985), Cultural difference theory

Tang and Bao: cultural differences in OER take-up

Freire and Critical Pedagogy

Funk and Guthadjaka: open platforms, Australian Indigenous context

Adam: MOOC designers, South African context

Grosfoguel, 2007 Maldonado-Torres, 2016 Langa et al., 2017

Decolonisation, First Nations Knowledge Authority

Rigney 2001

Care and Community in AI Ethics:

- Issues of alienation and dehumanization, including the creation of a climate of mistrust, fear and anxiety, loss of social cohesion
- Relationship-based response: "“a privileged relationship exists between members and students" (BC Teachers Federation)
- Benefit to subject as perceived by the subject (Cdn Nurses Assn)

Leading AI ethics approaches and perspectives

- Summary of Issues - https://ethics.mooc.ca/all_issues.htm

- Approaches to ethics https://ethics.mooc.ca/cgi-bin/page.cgi?module=9

- The Ethical Codes Reader https://docs.google.com/document/d/1mv9VxbIyGvBaFwHSvtyN1a3iwAKws6SZDGuDTlYMZ48/edit

- MetaEthics https://ethics.mooc.ca/cgi-bin/page.cgi?presentation=41

Part 1(d)

Real-World Examples

Kathleen C. Fraser

Digital Technologies Research Centre
National Research Council Canada

AI Ethics in REB Context

Sources of Risk

Impact on Research

	Research Data	AI Models/Tools
Research Participant Well-being	Privacy, consent, use of public data	Direct harm to participant, privacy risks
Proportionality (Value of Research)	Flawed data leads to flawed results	Use of an inappropriate third-party AI tool leads to flawed results

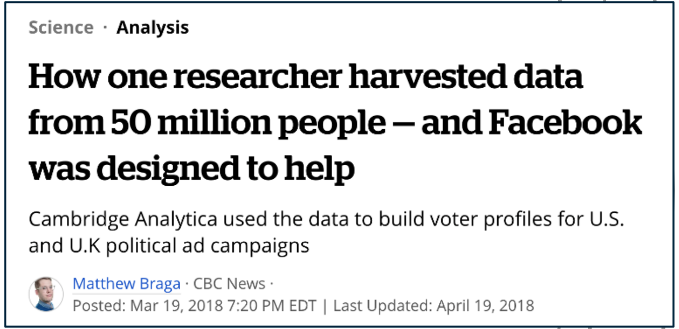
Research Data – Participant Risks

Questions: Have users **consented** for their data to be used for this purpose? Would they consider it a breach of **privacy**?

Example: Consent to participate

Cambridge Analytica scandal:

- Data originally collected by Dr. Aleksandr Kogan, Cambridge University.
- Online quiz for psychology research, common in his department
- Collected from the quiz-taker (who had “consented”) *and their friends*
- Data was collected according to Facebook’s Terms of Service



Science · **Analysis**

How one researcher harvested data from 50 million people – and Facebook was designed to help

Cambridge Analytica used the data to build voter profiles for U.S. and U.K political ad campaigns

Matthew Braga · CBC News ·
Posted: Mar 19, 2018 7:20 PM EDT | Last Updated: April 19, 2018

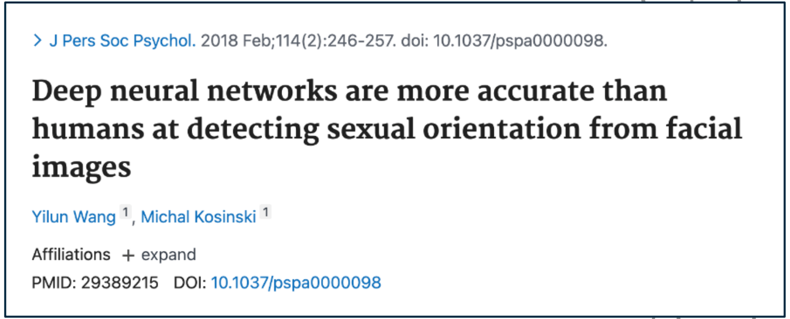
Does clicking a box online count as “informed consent”?

Can the user consent on behalf of their friends?

Example: Data from the public domain

"AI Gaydar" controversy:

- Led by Dr. Michal Kosinski, Stanford University
- Trained a machine learning classifier to distinguish between straight/gay men and women on the basis of a **photo of their face**
- Scraped images and sexual orientation data from **dating sites**



> J Pers Soc Psychol. 2018 Feb;114(2):246-257. doi: 10.1037/pspa0000098.

Deep neural networks are more accurate than humans at detecting sexual orientation from facial images

Yilun Wang [1], Michal Kosinski [1]

Affiliations + expand

PMID: 29389215 DOI: 10.1037/pspa0000098

Was the data in the public domain?

Was there a reasonable expectation of privacy?

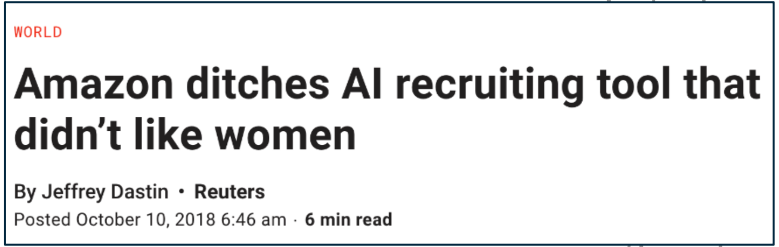
Research Data: Proportionality Risks

Questions: Are there sources of **bias** in the data which will limit the **value (potential benefits)** of the research?

Example: Biased Data

Amazon algorithmic hiring tool

- Replicated company’s **historical bias** for hiring men, rather than women
- Note: **“applicant gender”** was *not* an input to the algorithm
- Algorithm learned to down-vote resumes mentioning all-women colleges, women’s sports teams
- Also learned to up-vote resumes with words like ***executed*** and ***captured***, which were used more frequently by men



WORLD

Amazon ditches AI recruiting tool that didn’t like women

By Jeffrey Dastin • **Reuters**

Posted October 10, 2018 6:46 am · **6 min read**

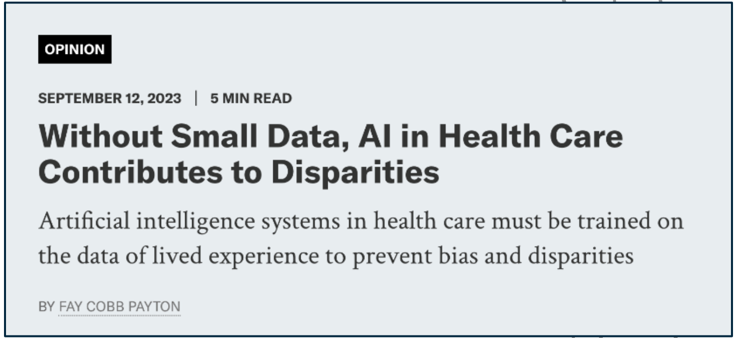
What is the likelihood of systemic bias in the training data?

Can that bias be removed or mitigated?

Example: Incomplete Data

AI tool for diabetes management

- Designing treatment plans for lower socioeconomic status groups
- Trained on **medical records** (lab tests, diagnosis codes, etc.)
- Does *not* take into account **social determinants of health**, transportation options to medical centre, food insecurity, employment, etc.



OPINION

SEPTEMBER 12, 2023 | 5 MIN READ

Without Small Data, AI in Health Care Contributes to Disparities

Artificial intelligence systems in health care must be trained on the data of lived experience to prevent bias and disparities

BY FAY COBB PAYTON

Will an AI model trained on this data be capable of benefiting participants and society?

AI Models/Tools: Participant Risks

Questions: How does the AI tool use **participant data**? What is the risk that participant data will be **leaked/exposed** by the trained model?

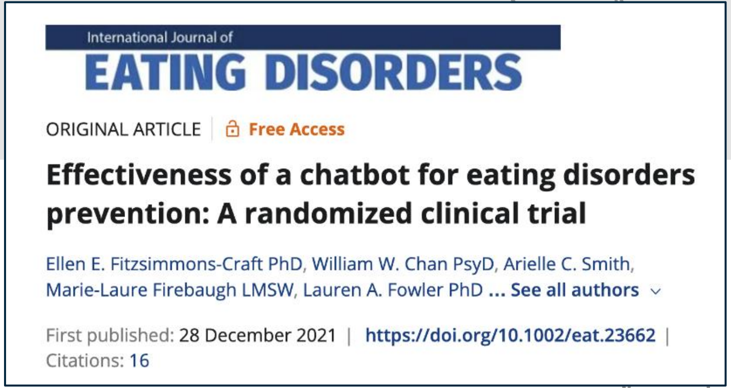
Example: Risk to Well-being

National Eating Disorders Association (US)

- Nonprofit advocacy group shut down helpline after 20 years, replaced with chatbot
- Used **third-party mental health chatbot**
- Tested in clinical trial with university researchers, found to reduce short-term ED risk
- At some point, chatbot company made **“systems upgrade”** to incorporate more AI in responses
- Resulting chatbot gave diet advice, promoted caloric deficits, etc.

Do the potential benefits outweigh the risks?

Do the researchers ultimately control the technology?



International Journal of

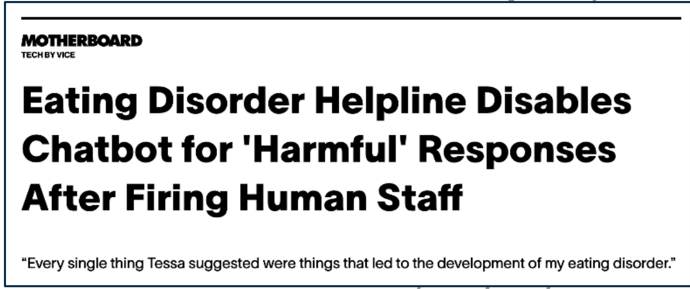
EATING DISORDERS

ORIGINAL ARTICLE | Free Access

Effectiveness of a chatbot for eating disorders prevention: A randomized clinical trial

Ellen E. Fitzsimmons-Craft PhD, William W. Chan PsyD, Arielle C. Smith, Marie-Laure Firebaugh LMSW, Lauren A. Fowler PhD ... See all authors

First published: 28 December 2021 | https://doi.org/10.1002/eat.23662 | Citations: 16



MOTHERBOARD
TECH BY VICE

Eating Disorder Helpline Disables Chatbot for 'Harmful' Responses After Firing Human Staff

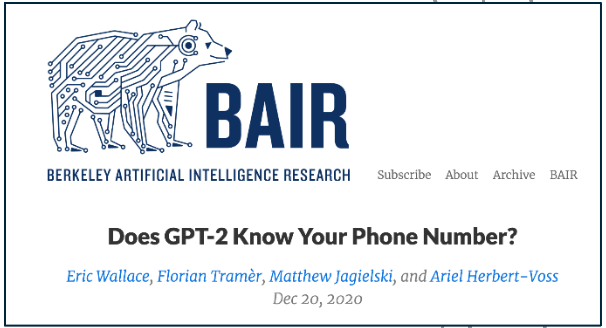
“Every single thing Tessa suggested were things that led to the development of my eating disorder.”

Example: Risk to Privacy

Berkley study extracting personal info from GPT-2

- Developed a method to uncover outputs that had been “memorized” from the training data
- Much of it was from news data, Wikipedia articles, and advertising
- Some of it included **personally-identifiable information**, including addresses and phone numbers
- *Note: training data, not input data*

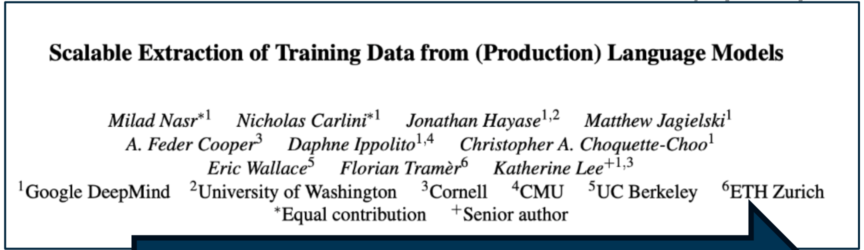
When training a new AI model, can researchers ensure that none of their participants’ data will be memorized?



Does GPT-2 Know Your Phone Number?

Eric Wallace, Florian Tramèr, Matthew Jagielski, and Ariel Herbert-Voss

Dec 20, 2020



Scalable Extraction of Training Data from (Production) Language Models

Milad Nasr*[1] Nicholas Carlini*[1] Jonathan Hayase[1,2] Matthew Jagielski[1]
A. Feder Cooper[3] Daphne Ippolito[1,4] Christopher A. Choquette-Choo[1]
Eric Wallace[5] Florian Tramèr[6] Katherine Lee+[1,3]
[1]Google DeepMind [2]University of Washington [3]Cornell [4]CMU [5]UC Berkeley [6]ETH Zurich
*Equal contribution +Senior author

Including Llama, Falcon, GPT-3.5 turbo

AI Models/Tools: Proportionality Risks

Questions: How well has the AI tool been **validated for the proposed purpose?** What is the likelihood that the AI model will produce **inaccurate output**, and what will be the effect on the research? Does the research team have the **expertise** to use and understand the AI tool?

Example: Tool Not Validated for Population

My own research: webcam based eye-tracking in elderly population

- Cognitive assessment tool that involves eye-tracking (REB approved)
- Using **a third-party AI model** for the eye-tracking
- Worked great .. until tested with older adults!
- Turns out oldest person in the training data was **41 years old**
- Fortunately: our team has the AI expertise to implement personalized re-training

Has the AI model been validated on the population of interest?

If not, can the research team adapt/retrain?

Future Considerations

Most academic AI research is not reviewed by REB

Many societal concerns about AI are not under the mandate of REB

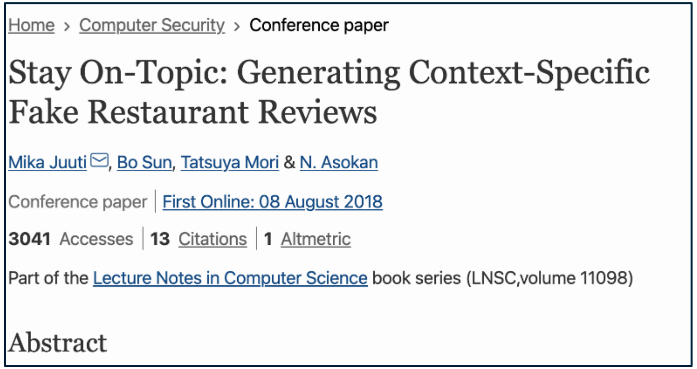
"[There is] a large gap between the relevant concerns that follow from AI research and those that fall under the purview of IRBs. Issues like **dual use of data**, **worker displacement, unrepresentative training data** and **excluding stakeholders** from project design and deployment remain unreviewed and often unmitigated." [Waeiss, 2023]

https://www.adalovelaceinstitute.org/blog/ai-research-ethics-collective-problem/

Thank you

kathleen.fraser@nrc-cnrc.gc.ca

Example: long-term consequences of the research



Home › Computer Security › Conference paper

Stay On-Topic: Generating Context-Specific Fake Restaurant Reviews

Mika Juuti, Bo Sun, Tatsuya Mori & N. Asokan

Conference paper | First Online: 08 August 2018

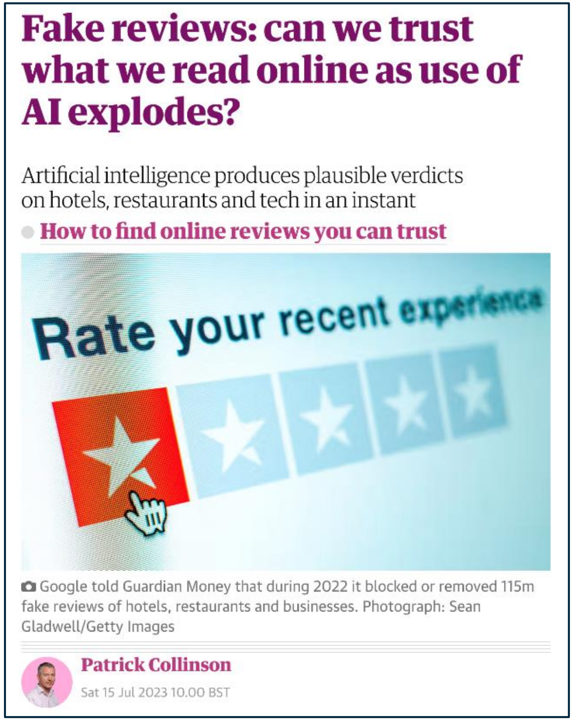
3041 Accesses | **13** Citations | **1** Altmetric

Part of the Lecture Notes in Computer Science book series (LNSC,volume 11098)

Abstract

2018

2023



Fake reviews: can we trust what we read online as use of AI explodes?

Artificial intelligence produces plausible verdicts on hotels, restaurants and tech in an instant

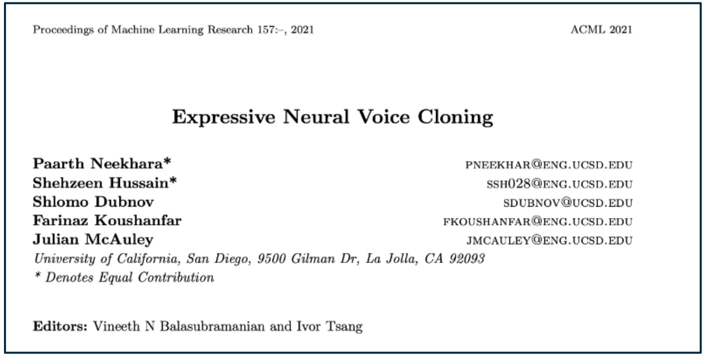
● **How to find online reviews you can trust**

Google told Guardian Money that during 2022 it blocked or removed 115m fake reviews of hotels, restaurants and businesses. Photograph: Sean Gladwell/Getty Images

Patrick Collinson

Sat 15 Jul 2023 10.00 BST

Example: long-term consequences of the research



Proceedings of Machine Learning Research 157:–, 2021 ACML 2021

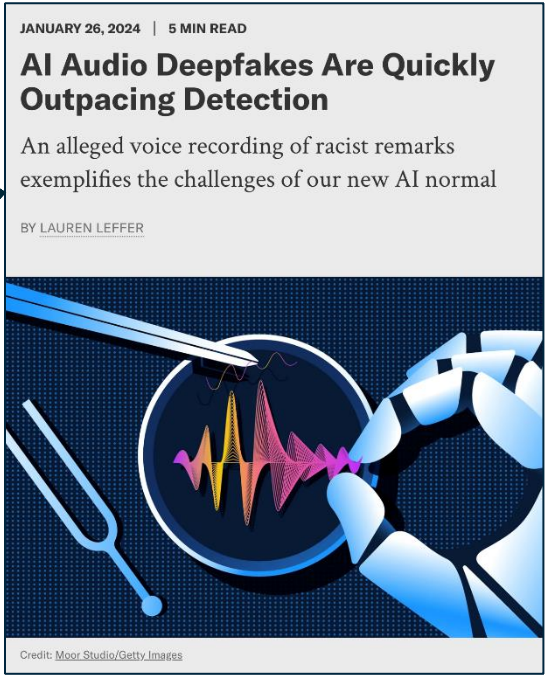
Expressive Neural Voice Cloning

Paarth Neekhara* PNEEKHAR@ENG.UCSD.EDU
Shehzeen Hussain* SSH028@ENG.UCSD.EDU
Shlomo Dubnov SDUBNOV@UCSD.EDU
Farinaz Koushanfar FKOUSHANFAR@ENG.UCSD.EDU
Julian McAuley JMCAULEY@ENG.UCSD.EDU
University of California, San Diego, 9500 Gilman Dr, La Jolla, CA 92093
** Denotes Equal Contribution*

Editors: Vineeth N Balasubramanian and Ivor Tsang

2021

2024



JANUARY 26, 2024 | 5 MIN READ

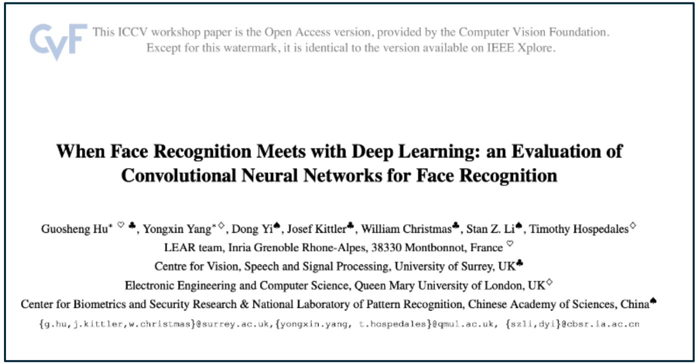
AI Audio Deepfakes Are Quickly Outpacing Detection

An alleged voice recording of racist remarks exemplifies the challenges of our new AI normal

BY LAUREN LEFFER

Credit: Moor Studio/Getty Images

Example: long-term consequences of the research

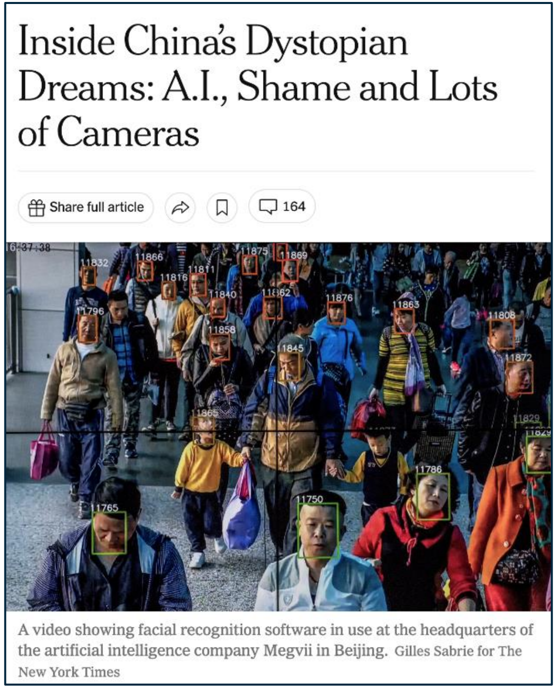


When Face Recognition Meets with Deep Learning: an Evaluation of Convolutional Neural Networks for Face Recognition

Guosheng Hu[*♡♣], Yongxin Yang[*◇], Dong Yi[♠], Josef Kittler[♣], William Christmas[♣], Stan Z. Li[♠], Timothy Hospedales[◇]
LEAR team, Inria Grenoble Rhone-Alpes, 38330 Montbonnot, France [♡]
Centre for Vision, Speech and Signal Processing, University of Surrey, UK[♣]
Electronic Engineering and Computer Science, Queen Mary University of London, UK[◇]
Center for Biometrics and Security Research & National Laboratory of Pattern Recognition, Chinese Academy of Sciences, China[♠]
{g.hu,j.kittler,w.christmas}@surrey.ac.uk,{yongxin.yang, t.hospedales}@qmul.ac.uk, {szli,dyi}@cbsr.ia.ac.cn

2015

2018



Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras

A video showing facial recognition software in use at the headquarters of the artificial intelligence company Megvii in Beijing. Gilles Sabrie for The New York Times

Break

…please return by 15:05 EDT

Small Group discussions:

- You will be assigned to a small group (virtual room) room, please click to enter
- If you are comfortable, please turn on cameras for the small group discussions
- Assign a note-keeper and spokes person (to present your group’s thoughts in the plenary)
- Notes should be entered on the Google document for your group # (see chat for links)

Discussion Questions (same for all groups):

A. What areas of AI Ethics do you feel:

i. Are likely to apply most often in your REB work?

ii. Raise the trickiest questions for analysis in the context of the kind of proposals your REB sees?

B. Based on the structure and principles provided in the Montreal Declaration and the discussion today, what are the 3 to 5 key questions which you feel you will need to ask about many / most REB proposals you see which involve AI in some way?

C. (If time permits) What supports or processes do you feel would help REBs like yours address these challenges?

Montreal Declaration Links:

English: https://declarationmontreal-iaresponsable.com/wp-content/uploads/2023/04/UdeM_Decl-IA-Resp_LA-Declaration-ENG_WEB_09-07-19.pdf
Français: https://declarationmontreal-iaresponsable.com/wp-content/uploads/2023/01/UdeM_Decl_IA_Resp_LA_Declaration_FR_web_4juin2019.pdf

Links to Group Notes Pages:

Group 1:
https://docs.google.com/document/d/17SmU2qFcvrIs7ix1skSYrM0NyD7uwiFVRwpJu9fKrVk/edit?usp=sharing

Group 2:
https://docs.google.com/document/d/1vP2dnZTMpXurwNet3eBY4AUvdxckHAJ-odAdKJi90Rk/edit?usp=sharing

Group 3:
https://docs.google.com/document/d/1T90gz96u833fqOjL5AKPeMKsE6fvWE3_u7HPMbccNkY/edit?usp=sharing

Group 4:
https://docs.google.com/document/d/1WG8N5OWAns546DCqfJ40rDpPgBfanD_-6yjRlTGWO8c/edit?usp=sharing

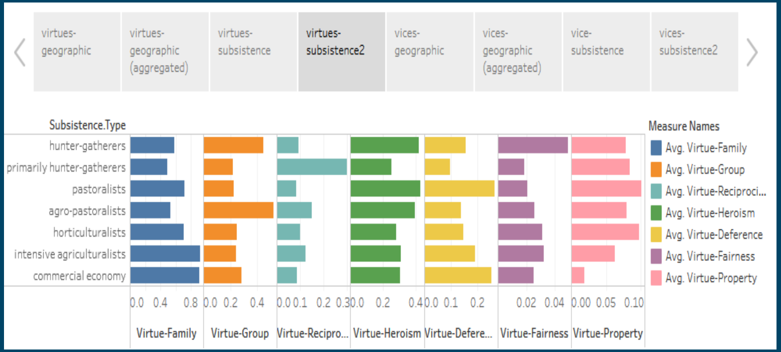
Part 2(b)

Current and future AI ethics challenges for REBs

Current and future AI ethics challenges for REBs

Challenging some oft-held ideas about ethics

- Universalism – the idea that there is (or should be) one set of ethics that applies to all
- Normativity – the idea that ethics is prescriptive and describes attitudes we *should have or* how we *should* behave
- Cognitivism – the idea that moral principles can be expressed as a series of knowable rules



Virtues and vices around the world
https://public.tableau.com/profile/mark.alfano#!/vizhome/Virtuesandvicesfromtheperspectivesof256cultures/Virtuesandvicesaroundtheworld

Current and future AI ethics challenges for REBs

MetaEthics

What is the basis for ethics?

- Types of ethics: descriptive, normative, analytic
- Relativism
  - Cultural, agent, speaker
  - Descriptive, metaethical, normative
- Non-cognitivism – "views moral discourse as a way to express attitudes towards certain actions."
- Realism: naturalism and non-naturalism, objectivism
- Rational choice vs intuition vs sentiment
  - emotivism, prescriptivism, expressivism
- Basis (motivation): deity, power, reason, agreement, sentiment

Current and future AI ethics challenges for REBs

Who Owns Ethics

- Scientific Virtues
  - “scientists invoke theoretical virtues explicitly, albeit rather infrequently, when they talk about models”
- Business Ethics
  - “real concerns and real-world problems of the vast majority of managers”
- Silicon Valley Ethics
  - Metcalf, Moss & Boyd: “broader and longer-standing industry commitments to meritocracy, technological solutionism, and market fundamentalism”

https://philpapers.org/archive/MIZTVI.pdf
https://hbr.org/1993/05/whats-the-matter-with-business-ethics
https://datasociety.net/library/owning-ethics-corporate-logics-silicon-valley-and-the-institutionalization-of-ethics/

Thankyou

Digital Privacy and Security program - National Research Council Canada

Programme de vie privée et sécurité numériques - Conseil national de recherches Canada

NRC.DPSec-VPSN.CNRC@nrc-cnrc.gc.ca