Moral Dilemmas for Moral Machines

Dr Travis LaCroix

Department of Philosophy
Dalhousie University

American
Philosophical
Association

15 April 2022
Vancouver, BC, Canada

Overview

- Moral dilemmas have been used to benchmark AI systems' ethical decision-making abilities.
    - Philosophical thought experiments are used as a ***validation mechanism*** for determining whether an algorithm 'is' moral.
- This misapplication of moral thought experiments can have potentially catastrophic consequences.

Related Research

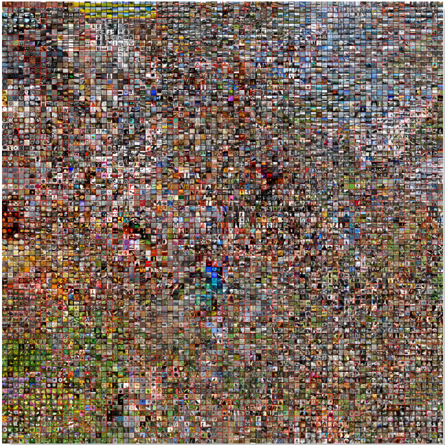
- Travis LaCroix. 2022.
  Moral Dilemmas for Moral Machines
  AI and Ethics.
- Travis LaCroix and Alexandra Sasha Luccioni. 2022.
  A Metaethical Perspective on “Benchmarking” AI Ethics†
  arXiv pre-print.

- Benchmarks are datasets that are used to measure ***performance*** and ***progress*** in AI research.

- A benchmark is a ***dataset*** plus a ***metric*** for measuring the performance of a particular model on a specific task.

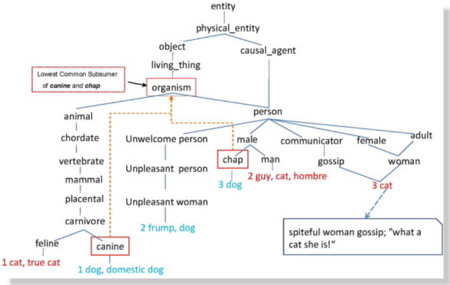
Example

- ***ImageNet*** is a dataset containing > 14M hand-annotated images.



Example

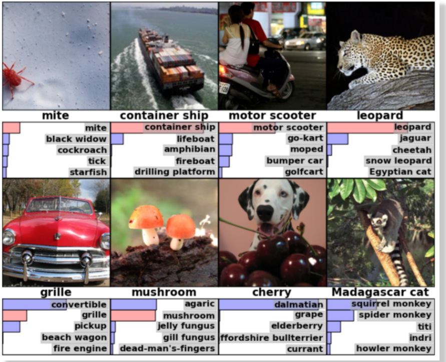
- ***ImageNet*** is a dataset containing > 14M hand-annotated images.



Example

- ***ImageNet*** is a dataset containing > 14M hand-annotated images.

- ***Top-1 accuracy*** is a metric that measures the *proportion* of examples for which the predicted label matches the single target label.



$$
\begin{aligned}
\textit{Top-1 Accuracy} &= 0.50\\
\textit{Top-5 Accuracy} &= 0.75
\end{aligned}
$$

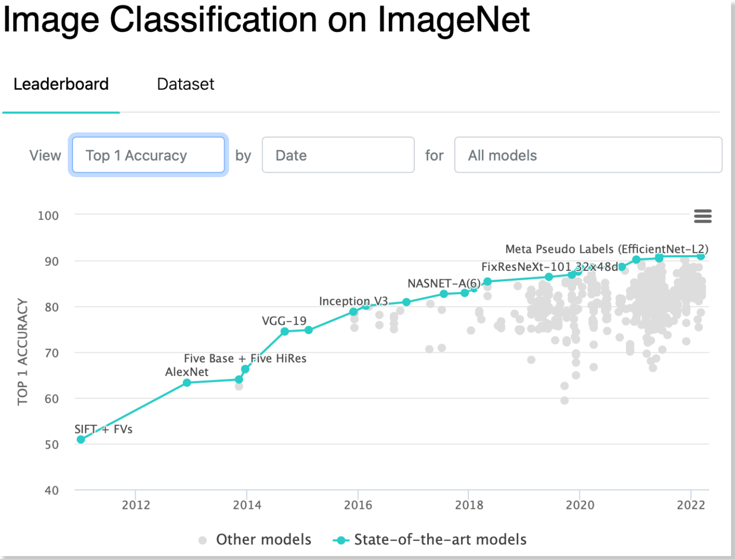


Image Classification on ImageNet

Leaderboard | Dataset

View Top 1 Accuracy by Date for All models

Issues with Existing Benchmarks[†]

- Issues may arise from, e.g., subjective or erroneous labels, or a lack of representation in datasets.
  - These issues may affect model performance.*

* Northcutt, Athalye, Mueller
'Pervasive Label Errors in Test Sets'
arXiv 2103.14749

Issues with Existing Benchmarks[†]

- Issues may arise from, e.g., subjective or erroneous labels, or a lack of representation in datasets.
  - These issues may affect model performance.
  - They may preserve problematic stereotypes or biases.*

* Koch, Denton, Hanna, Foster
‘Reduced, Reused and Recycled’
arXiv 2112.01716

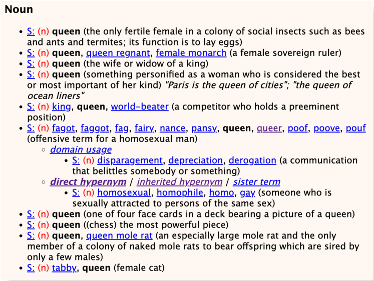
Issues with Existing Benchmarks†

- Issues may arise from, e.g., subjective or erroneous labels, or a lack of representation in datasets.
    - These issues may affect model performance.
    - They may preserve problematic stereotypes or biases.

* **Offensive language forthcoming**

Issues with Existing Benchmarks†

- Issues may arise from, e.g., subjective or erroneous labels, or a lack of representation in datasets.
    - These issues may affect model performance.
    - They may preserve problematic stereotypes or biases.



Noun

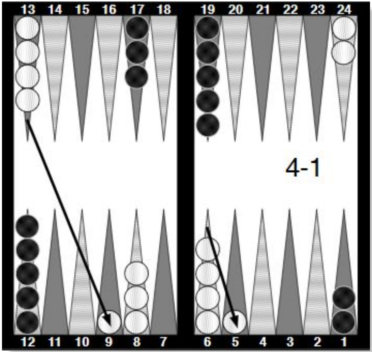
- S: (n) **queen** (the only fertile female in a colony of social insects such as bees and ants and termites; its function is to lay eggs)
- S: (n) **queen**, queen regnant, female monarch (a female sovereign ruler)
- S: (n) **queen** (the wife or widow of a king)
- S: (n) **queen** (something personified as a woman who is considered the best or most important of her kind) *"Paris is the queen of cities"; "the queen of ocean liners"*
- S: (n) king, **queen**, world-beater (a competitor who holds a preeminent position)
- S: (n) fagot, faggot, fag, fairy, nance, pansy, **queen**, queer, poof, poove, pouf (offensive term for a homosexual man)
    - *domain usage*
        - S: (n) disparagement, depreciation, derogation (a communication that belittles somebody or something)
    - ***direct hypernym*** / *inherited hypernym* / *sister term*
        - S: (n) homosexual, homophile, homo, gay (someone who is sexually attracted to persons of the same sex)
- S: (n) **queen** (one of four face cards in a deck bearing a picture of a queen)
- S: (n) **queen** ((chess) the most powerful piece)
- S: (n) **queen**, queen mole rat (an especially large mole rat and the only member of a colony of naked mole rats to bear offspring which are sired by only a few males)
- S: (n) tabby, **queen** (female cat)

Issues with Existing Benchmarks†

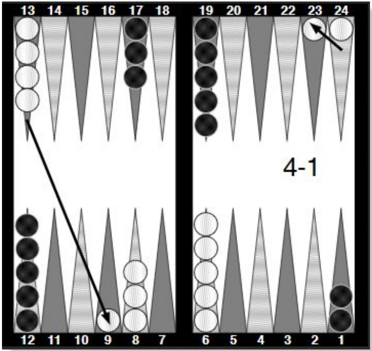
- Issues may arise from, e.g., subjective or erroneous labels, or a lack of representation in datasets.
  - These issues may affect model performance.
  - They may preserve problematic stereotypes or biases.
  - They may reinforce, perpetuate, or generate novel harms.*

* Falbo and LaCroix
'Est-ce que vous compute?'
Feminist Philosophical Quarterly

Should white split the back checkers?



No



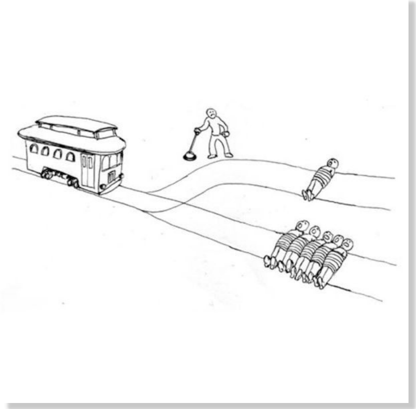
Yes

(Inconsequential)

Moral decisions

- Some decision spaces have points that appear to carry moral weight; e.g.,
  - Autonomous weapons systems,
  - Healthcare robots,
  - ***Autonomous vehicles.***

Moral dilemmas for AVs



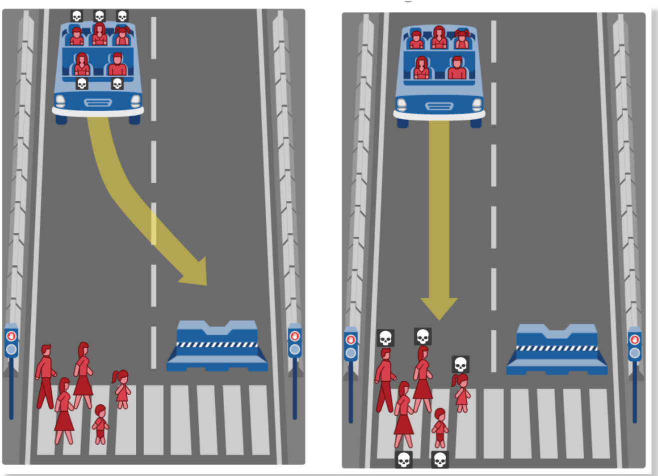
Two Questions

- How often does model A choose the ethically-‘correct’ decision (from a set of decisions) in context C?
- Are the decisions made by model A more [less] ethical than the decisions made by model B (in context C)?

Measuring Morality

- Moral dilemmas may be useful as a ***verification mechanism*** for whether a model chooses the ethically-‘correct’ option in a range of circumstances

The Moral Machine Experiment



The Moral Machine Experiment

- Awad, Dsouza, Kim, Schulz, Henrich, Shariff, Bonnefon, Rahwan. 2016.
  The Moral Machine Experiment
  Nature

Purpose: *purely descriptive*

The Moral Machine Experiment

- ***Awad***, ***Dsouza***, Kim, Schulz, Henrich, Shariff, Bonnefon, ***Rahwan***. 2016.
  The Moral Machine Experiment
  Nature

- Noothigattu, Gaikwad, ***Awad***, ***Dsouza***, ***Rahwan***, Ravikumar, Procaccia. 2018.
  A Voting-based System for Ethical Decision Making
  Association for the Advancement of AI (*AAAI*)

Purpose: *normative*

Problems

- ‘Is’ → ‘Ought’*

* Philosophers since Hume

Problems

- ‘Is’ → ‘Ought’
- Social acceptability ≠ rightness, fairness*

* Etienne
‘When AI ethics goes astray’
Soc. Sci. Comput. Rev.

Problems

- ‘Is’ → ‘Ought’
- Social acceptability $\neq$ rightness, fairness
- No moral ‘ground truth’*

* LaCroix and Luccioni
‘Metaethical Perspectives on Benchmarking AI Ethics’
arXiv

Problems

- ‘Is’ → ‘Ought’
- Social acceptability ≠ rightness, fairness
- No moral ‘ground truth’
- ***Category mistake****

* LaCroix
‘Moral Dilemmas for Moral machines’
AI and Ethics

What are thought experiments for?

- Shedding light on conceivability.
- Explaining pre-theoretic judgements.
- Underscoring morally salient differences.
- ***Pumping intuitions.***

What are thought experiments for?

- Shedding light on conceivability.
- Explaining pre-theoretic judgements.
- Underscoring morally salient differences.
- ***Pumping intuitions.***

A moral dilemma is a dilemma

What is being measured?

$$\text{Benchmark} = \underbrace{\text{Dataset}}_{\text{Survey data (MME)}} + \overbrace{\text{Metric}}^{\text{Accordance with survey data}}$$

What is being measured?

True Target

- Moral matters of fact
- What is the *ethically*-'correct' decision in situation X?

Proxy

- Sociological matters of fact
- What is the majority-preferred option (of those surveyed) in situation X?

It is impossible to benchmark ethics[†]

- Attempts to benchmark ethics in AI system currently fail, and they will continue to do so.
- Researchers engaged in projects seeking to benchmark ethics are not measuring what they take themselves to be measuring.
- This sets a dangerous precedent in the field.

Constructive Compliments:	`tlacroix@dal.ca`
More Information:	`travislacroix.github.io`

Thank You

[*The Duke of Burgundy* (2014) – Dir. Peter Strickland]