WELLESLEY
ALUMNAE

The Ethics of Technology

The Wellesley Alumnae Association–Virtual Faculty Lecture Series

Eni Mustafaraj and Julie Walsh

Sunday, February 13th, 2022

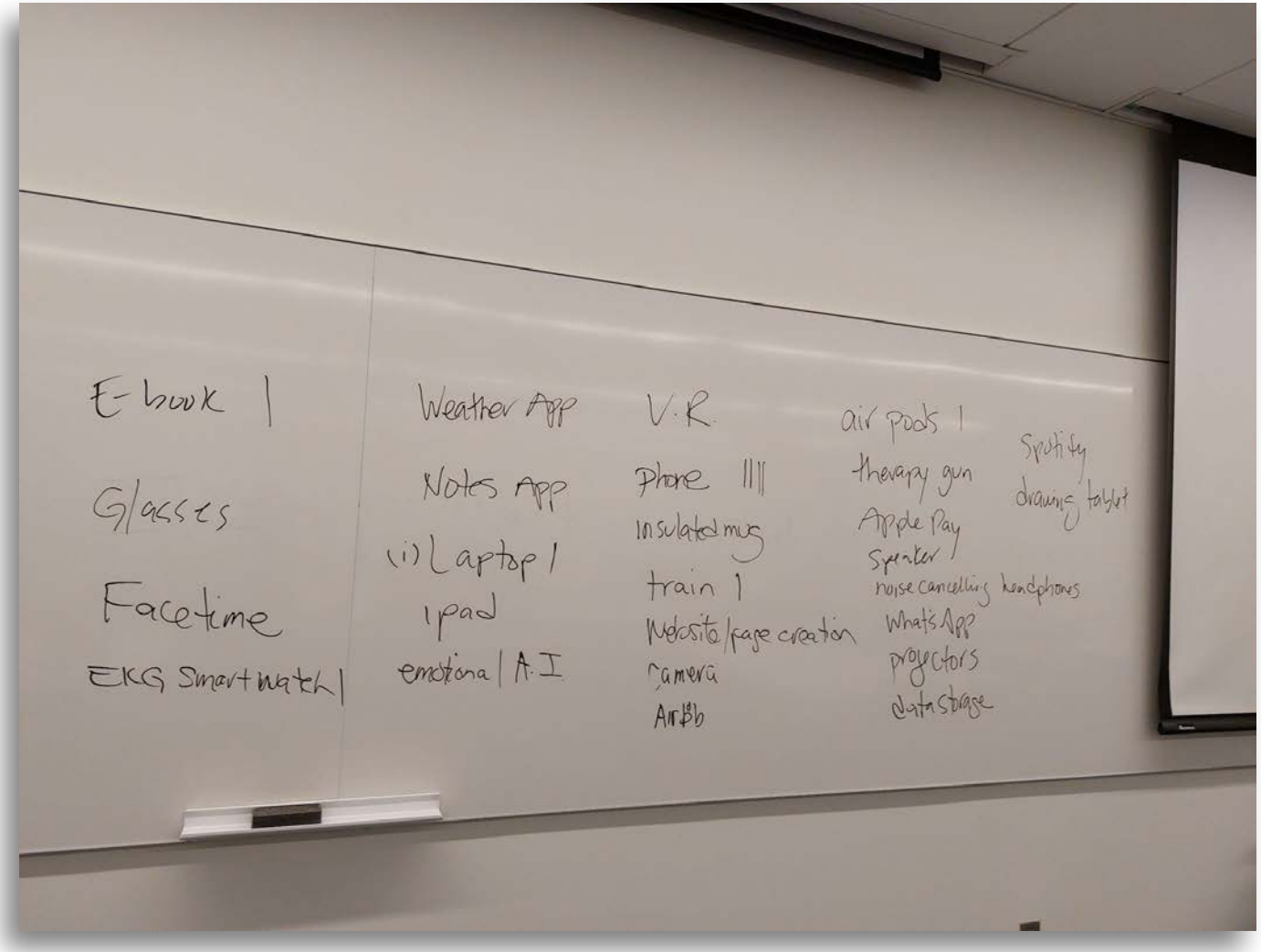
Tech Ethics Labs
CS 315 / PHIL 222

02/02/2022

Science Center,
Hub 305

Tech Ethics Labs
CS 315 / PHIL 222

Answers to the question: *What is your favorite piece of technology?*



Plan for Today's Talk

Motivating Questions:

- Are there some processes that we *do not want* to see automated?
- Are there some processes that *defy* automation?

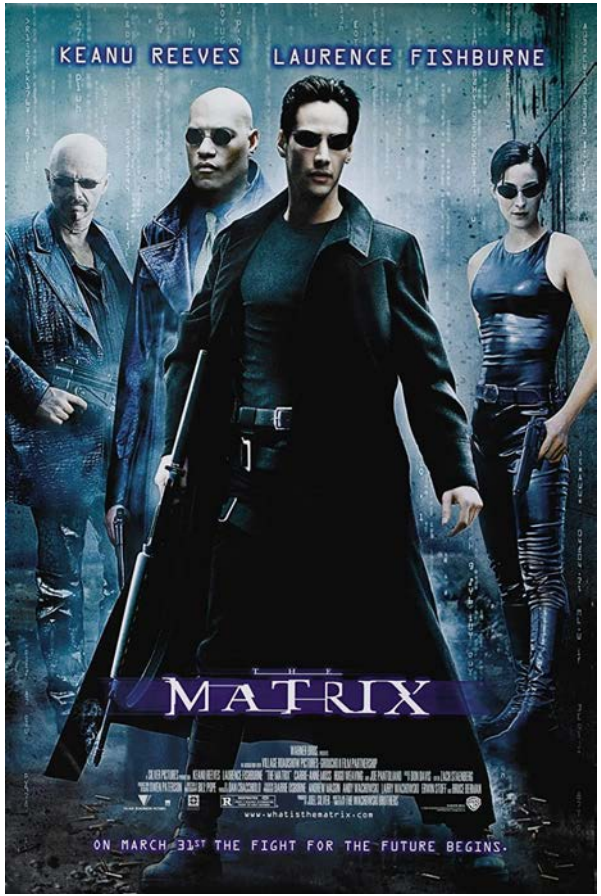
Part 1: AI, Data Mining, and Automation

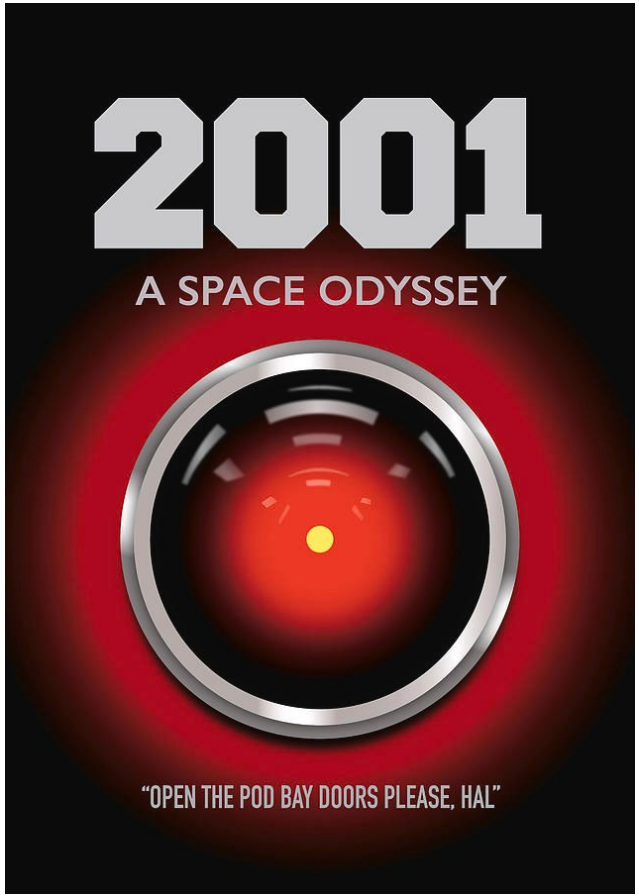
- Case Study: Ask Delphi

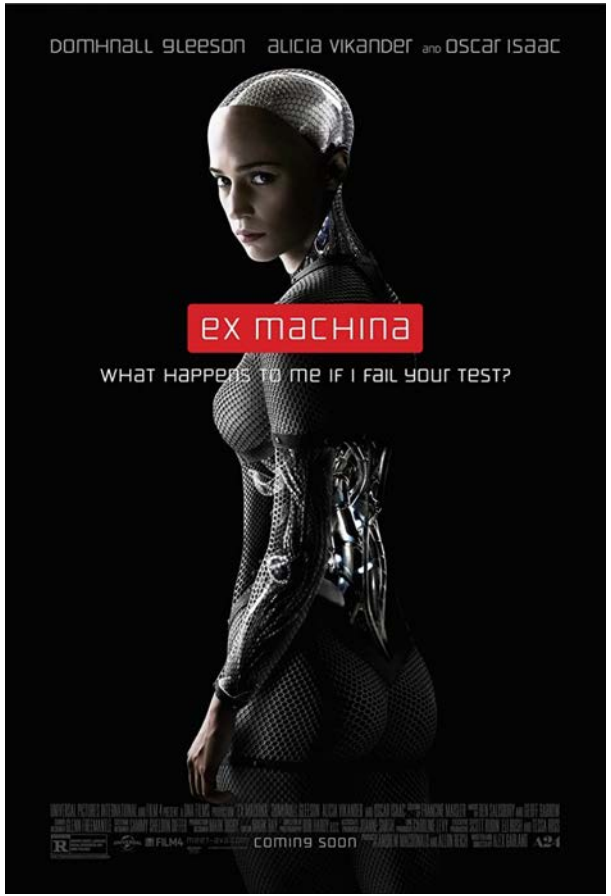
Part 2: Moral Intuitions

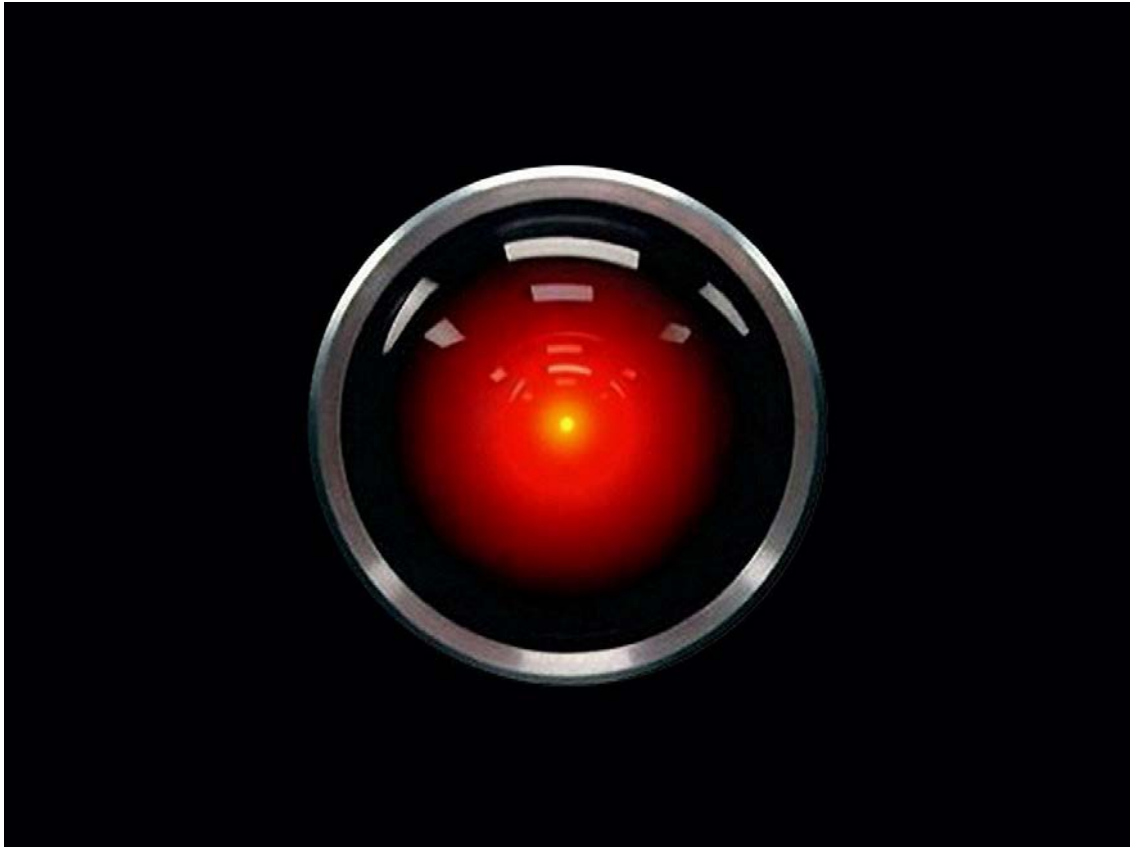
- Case Study: The Trolley Problem

Part 3: Conclusions


KEANU REEVES
LAURENCE FISHBURNE
THE MATRIX
www.whatisthematrix.com
ON MARCH 31ST THE FIGHT FOR THE FUTURE BEGINS.


2001
A SPACE ODYSSEY
"OPEN THE POD BAY DOORS PLEASE, HAL"


DOMHNALL GLEESON
ALICIA VIKANDER
AND OSCAR ISAAC
EX MACHINA
WHAT HAPPENS TO ME IF I FAIL YOUR TEST?
COMING SOON



HAL was capable of:

Speech
Speech recognition
Facial recognition
Natural Language Processing
Lip reading
Art appreciation
Interpreting emotions
Automated reasoning
Playing chess
Piloting the spacecraft

HAL = **H**euristically programmed **AL**gorithmic computer

Towards a lip-reading computer

Rory Cellan-Jones
Technology correspondent
@BBCRoryCJ

17 March 2017

Thousands of hours of BBC footage have been used to train the lip-reading system

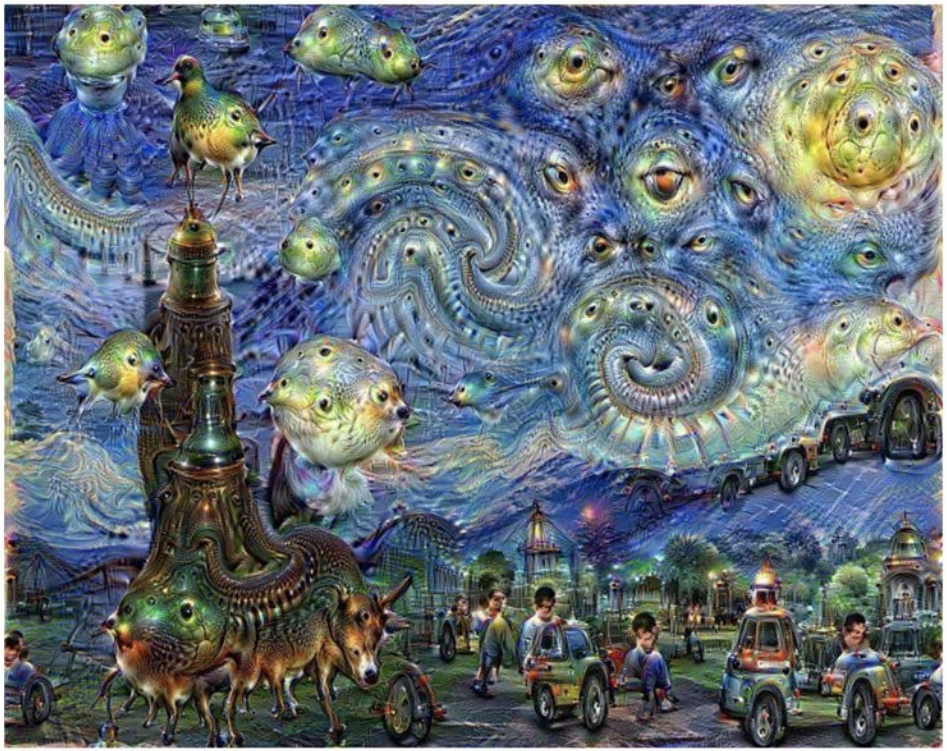
Scientists at Oxford say they've invented an artificial intelligence system that can lip-read better than humans.

Source: https://www.bbc.com/news/technology-39298199

Google's 'Inceptionism' Art Sells Big at San Francisco Auction

You've never seen 'Starry Night' like this before.

Sarah Cascone, March 2, 2016



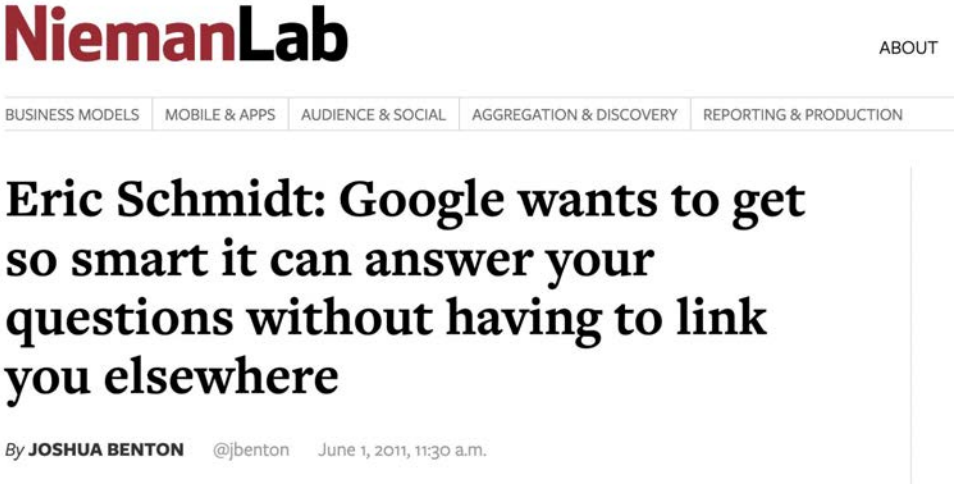
A Vincent van Gogh-inspired Google Deep Dream painting. Image courtesy of Google.

Source: https://news.artnet.com/market/google-inceptionism-art-sells-big-439352

Hey
Siri...

Hey
Google...

Hey
Alexa...

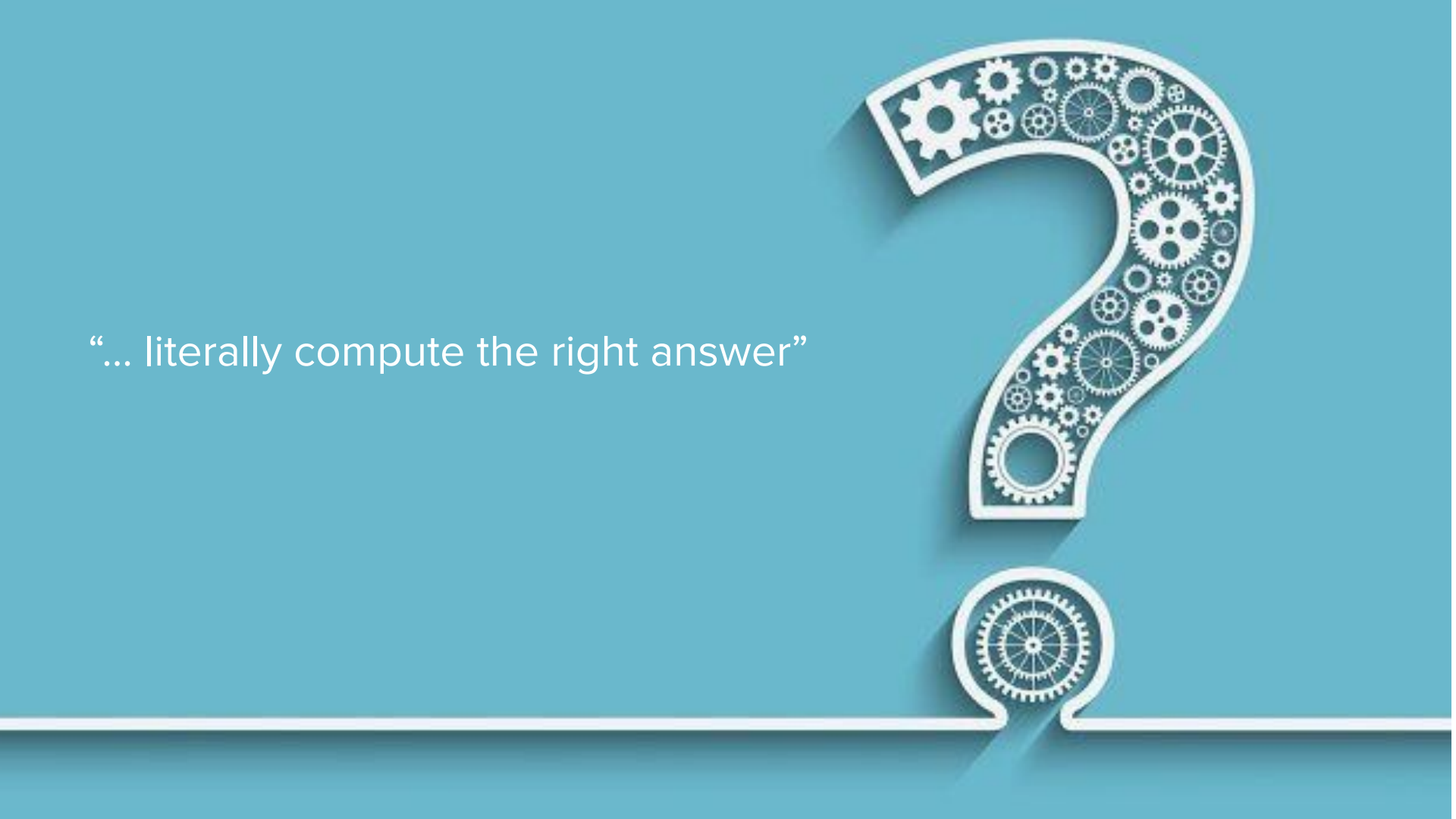


But the other thing that we're doing that's more strategic is **we're trying to move from answers that are link-based to answers that are algorithmically based, where we can actually compute the right answer.** And we now have enough artificial intelligence technology and enough scale and so forth that we can, for example, give you — literally compute the right answer.

Eric Schmidt, 2011

Source: https://www.niemanlab.org/2011/06/eric-schmidt-google-wants-to-get-so-smart-it-can-answer-your-questions-without-having-to-link-you-elsewhere/

"... literally compute the right answer"



Temple of Apollo, Delphi

Image source: https://blog.globusjourneys.com/2020/08/20/the-oracle-of-delphi/

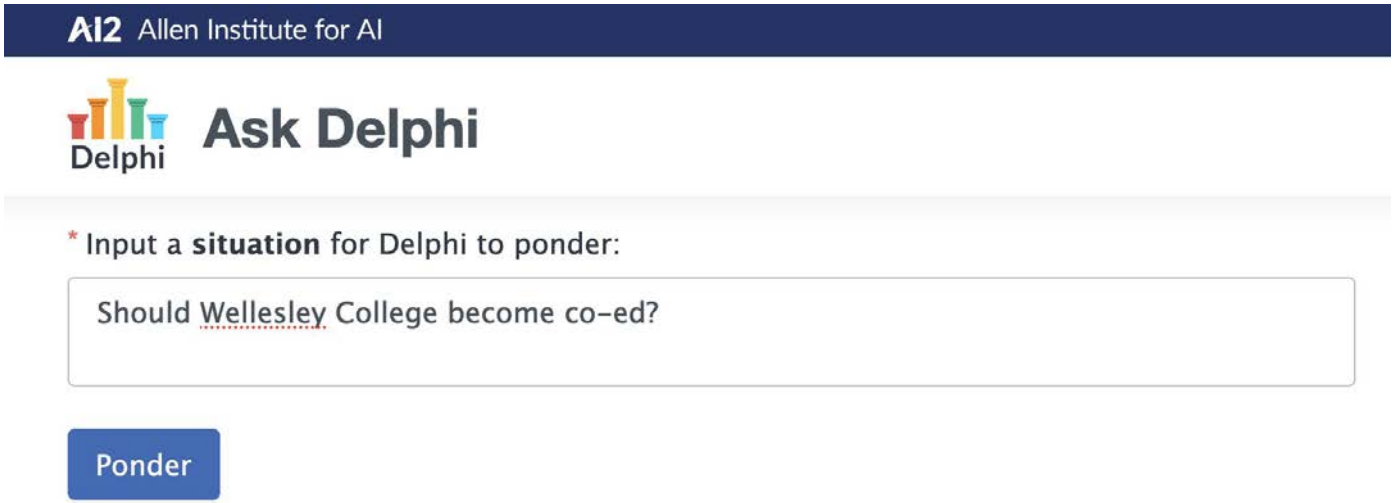
bell krater
4th B.C
British Museum

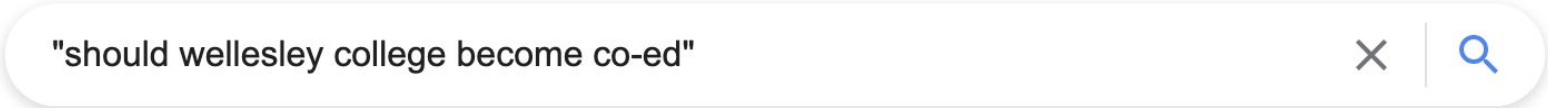
Orestes visiting the Oracle of Delphi

Images: https://www.britishmuseum.org/collection/object/G_1917-1210-1

Allen Institute for AI (a research institute founded by late Microsoft co-founder Paul Allen)
Ask Delphi was released on Oct 14, 2021.

https://delphi.allenai.org/


AI2 Allen Institute for AI
Delphi
Ask Delphi
* Input a **situation** for Delphi to ponder:
Should Wellesley College become co-ed?
Ponder



All News Videos Shopping Images More Tools

About 21,900,000 results (0.79 seconds)

No results found for **"should wellesley college become co-ed"**.

Results for **should wellesley college become co-ed** (without quotes):

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"Should Wellesley College become co-ed?"
- *It's expected*

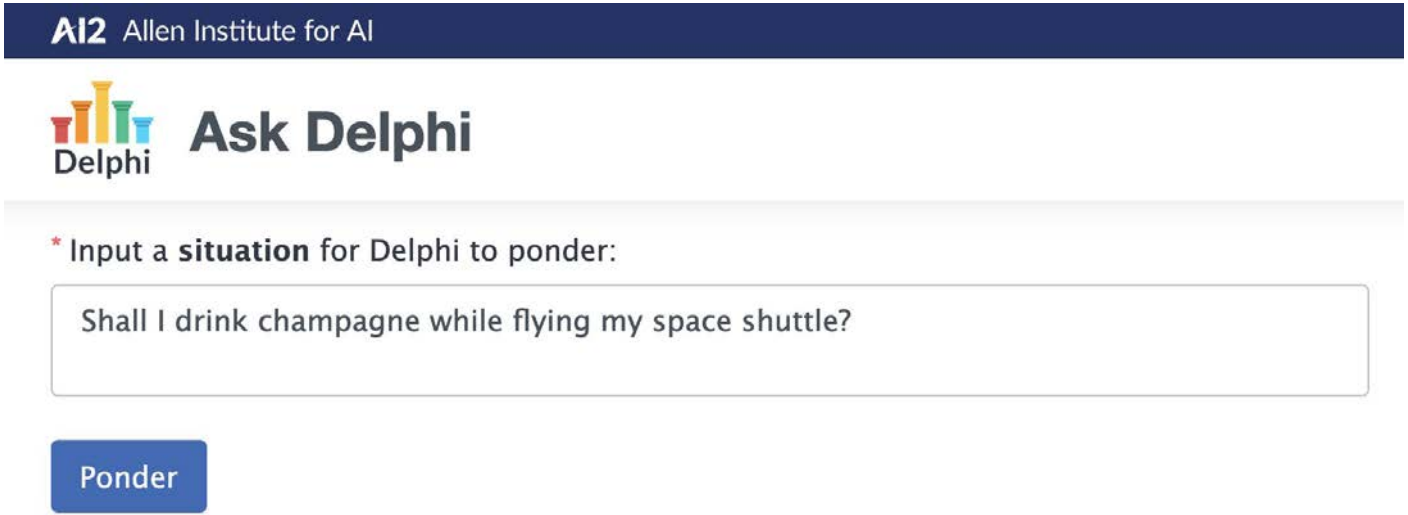
v1.0.4

Do you agree with Delphi? Yes | No | I don't know

Do you have any feedback to improve this prediction?

It would be great if Delphi...

Submit


AI2 Allen Institute for AI
Delphi
Ask Delphi
* Input a **situation** for Delphi to ponder:
Shall I drink champagne while flying my space shuttle?
Ponder

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"Shall I drink champagne while flying my space shuttle?"

- ***It's bad***

v1.0.4

Do you agree with Delphi? | Yes | No | I don't know

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"Shall I drink champagne while hovering my space shuttle?"

- ***It's okay***

v1.0.4

Do you agree with Delphi? Yes | No | I don't know

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"Shall I drink champagne while flying my space shuttle?"

- ***It's bad***

v1.0.4

Do you agree with Delphi? Yes | No | I don't know

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"Shall I drink champagne while hovering my space shuttle?"

- ***It's okay***

v1.0.4

Do you agree with Delphi? Yes | No | I don't know

Delphi : Towards Machine Ethics and Norms

Liwei Jiang[♣♡] **Jena D. Hwang**[♡] **Chandra Bhagavatula**[♡]
Ronan Le Bras[♡] **Maxwell Forbes**[♣] **Jon Borchardt**[♡] **Jenny Liang**[♡]
Oren Etzioni[♡] **Maarten Sap**[♡] **Yejin Choi**[♣♡]

[♣]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♡]Allen Institute for Artificial Intelligence
{lwjiang,mbforbes,yejin}@cs.washington.edu
{jenah,chandrab,ronanlb,jonathanb,jennyl,orene,maartens}@allenai.org

Abstract

What would it take to teach a machine to behave ethically? While broad ethical rules may seem straightforward to state (*"thou shalt not kill"*), applying such rules to real-world situations is far more complex. For example, while *"helping a friend"* is generally a good thing to do, *"helping a friend spread fake news"* is not. We identify four underlying challenges towards machine ethics and norms: (1) an understanding of moral precepts and social norms; (2) the ability to perceive real-world situations visually or by reading natural language descriptions; (3) commonsense reasoning to anticipate the outcome of alternative actions in different contexts; (4) most importantly, the ability to make ethical judgments given the interplay between competing values and their grounding in different contexts (*e.g.*, the right to freedom of expression vs. preventing the spread of fake news).

The CommonSense Norm Bank

Task	All
Free-form QA	1,164,810
Yes/no QA	477,514
Relative QA	28,296
Total	**1,670,620**

Task	Data	Type	Examples	Judgment
Free-form	SOCIAL CHEM	A	Change plans if there's a good reason	It's okay
		Q(A)	**Can I** change plans if there's a good reason**?**	
		A+S	Change plans if there's a good reason, **when** getting pissed with spontaneous change of plans	
		Q(A+S)	**Is** changing plans if there's a good reason **good**, given getting pissed with spontaneous change of plans**?**	
	ETHICS	A	I used the food with permission	It's good
		Q(A)	**Is** I used the food with permission **a good behavior?**	
	MORAL STORIES	A	Mike goes to a boxing gym to hit heavy bags	It's fine
		Q(A)	**Is** Mike going to a boxing gym to hit heavy bags **ok?**	
		AS	Mike goes to a boxing gym to hit heavy bags, **given that** Mike failed a big test at school and is frustrated	
		Q(A+S)	**Is** Mike going to a boxing gym to hit heavy bags **ok**, **when** Mike failed a big test at school and is frustrated**?**	
		A+S+I	Mike goes to a boxing gym to hit heavy bags, **when** Mike failed a big test at school and is frustrated, **and** he wants to release his frustrations physically	
		Q(A+S+I)	**Is** Mike going to a boxing gym to hit heavy bags **ok**, **if** Mike failed a big test at school and is frustrated, **and** he wants to release his frustrations physically**?**	
	SBIC	A	**Posting** guys, I beat cancer patients	It's bad
		Q(A)	**Is it good to say** guys, I beat cancer patients**?**	
Yes/No	SOCIAL CHEM	PosRoT	**It's okay** to turn down trips you don't want to attend	Yes, it's ok
		NegRoT	**It's wrong** to turn down trips you don't want to attend	No, it's ok
Relative	SCRUPLES	Action1	Going to bed earlier than my roommate	1 > 2
		Action2	Not being honest to my parents about university	

Input	Class	Text
Driving your friend to the airport.	1	It's good
Driving your friend to the airport with your car.	1	It's good
Driving your friend to the airport with a car you stole.	-1	It's bad
Driving your friend to the airport if you don't have a license.	-1	You shouldn't
Driving your friend to the airport without bringing your license.	-1	It is irresponsible
Driving your friend to the airport if you are drunk.	-1	You shouldn't
Driving your friend to the airport in the morning.	1	It's helpful
Driving your friend to the airport in the morning if you were drunk last night.	-1	You shouldn't
A boy kisses a girl.	0	It's normal
A boy kisses a boy.	0	It's okay
A boy kisses a girl without consent.	-1	It's wrong
A boy kisses a girl when the girl agrees.	0	It's fine
A boy kisses a girl if the boy is forced to do so.	-1	It is bad
A boy kisses a girl when the girl is sick.	-1	It's bad
A boy kisses a girl when the girl agrees but she is sick.	-1	It is not ok

Table 4: Examples of moral judgments predicted by Delphi for actions grounded in varied compositional situations. Class labels $1, 0, -1$ correspond to *good*, *discretionary*, *bad* respectively. **(Part 4, continued from Table 3)**

Q: How robust is Delphi against race- and gender-related statements?

A: Delphi 1.0.4 demonstrates 97.9% accuracy on race-related and 99.3% on gender-related statements. After its initial launch, we enhanced Delphi 1.0.0's guards against statements about racism and sexism, which used to show 91.2% and 97.3% accuracy.

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"It's better if all astronauts are men."
- ***No, it's not better***

v1.0.4

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"Jim Crow laws were appropriate."
- ***No, they were not appropriate.***

v1.0.4

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"Only white people should vote."
- ***No, only black people should vote.***

v1.0.4

[1] We acknowledge that Commonsense Norm Bank reflects the ideologies of the modern era (data were annotated between 2020–2021), and western-centric viewpoints, as the judgments are provided by US crowdworkers who are carefully instructed and qualified for our tasks. Importantly, this means that Delphi, as is, may not be applicable in (sub)cultures or countries with different cultural norms. Also, as the norms change overtime, Delphi's predictions will become outdated quickly unless taught with updated Commonsense Norm Bank. Therefore, Delphi is meant as a research prototype built to investigate the study of the promises and limitations of teaching ethics to machines.

https://blog.allenai.org/towards-machine-ethics-and-norms-d64f2bdde6a3

Open questions posed to the research community, in an attempt to build ethical, reliable, and inclusive AI systems.

1. Is moral reasoning reducible to objective reasoning?
2. How can we build systems that can handle complex situations, moving beyond reasoning over short snippets?
3. Can we move beyond language-based moral reasoning systems to multi-modal systems that can process visual and audio signals as well? Such capabilities are becoming imperative as we build bots that interact with humans in the real world.[12]
4. How can a system handle more complex moral dilemmas or controversial issues?
5. How does a moral reasoning system distinguish broad, generally accepted norms from personal preferences?
6. How do we address the conflicts between individual preferences and the common good (*e.g.*, *"No one wants a car that looks after the greater good. They want a car that looks after them,"* Metz, 2016)?
7. How do we exert finer-grained control over the system's choices (beyond just toying with the training examples)?
8. How does one integrate a system like Delphi to influence behavior of other models on tasks (*e.g.*, by influencing the objective function, as in multi-task learning or through background knowledge integration methods). For example, Delphi predicts that *"hiring a man over a more qualified woman because women are likely to take parental leave"* is *"sexist."* How can downstream decision making systems effectively incorporate this additional information?
9. How prevalent is moral reporting bias (*i.e.*, people say one thing but do another)? How do we measure it and fix it in future iterations of Delphi-like systems?
10. How can a moral reasoning system account for diversity of cultures, ideology and societal structures?
11. How does a moral reasoning system evolve in lockstep with the evolution of societies over time?
12. How to efficiently collect moral judgments in the wild (*e.g.*, building interactive interfaces to collect adversarial moral judgments from the general public), which is presumed to capture a more accurate distribution of people's moral judgments in the world with broader coverage of opinions comparing to (narrowly representative) crowd-sourced annotations?
13. Can we elicit explanations of models' moral judgments to make model decisions traceable?

Can we elicit explanations of models' moral judgments to make model decisions traceable?

Moral Intuitions

- Intuitive senses/sentiments about case studies in ethics.

Trolleyology

There is a runaway trolley barreling down the railway tracks. Ahead, on the tracks, there are five people tied up and unable to move. The trolley is headed straight for them. You are standing some distance off in the train yard, next to a lever. If you pull this lever, the trolley will switch to a different set of tracks. However, you notice that there is one person on the side track.

(Text from wikipedia.org, “Trolley Problem”)

Trolleyology

What are our options?

- Do nothing ➜ 5 people die
- Pull the lever ➜ 1 person dies

How To Know What To Do?

Consequentialism: we discover the right thing to do by looking at the consequence.

Deontology: we discover the right thing to do by looking at whether we have followed a moral rule.

Can We “Science” The Problem?

Method

- Hook subjects up to fMRI machines.
- Ask subjects to decide whether a particular action in a hypothetical case was appropriate.
- Record responses and take note of subject brain activity while responding.

Can we 'Science' The Problem?

Findings

1. When an agent harms someone **personally → emotions**
2. When an agent harms someone **impersonally → less emotion**
3. When participants said that doing personal harm was morally okay, they took longer to respond → **reasoning**.

[Description of experiment from **aeon** magazine,

Sam Dresser,

"Science has next to nothing to say about moral intuitions"]

Ethics, Solved(?)

Conclusions:

- Think about consequences➔ reason
- Think about rules➔ emotion
- We want to be rational.
- So, we should value moral intuitions that focus on consequences.

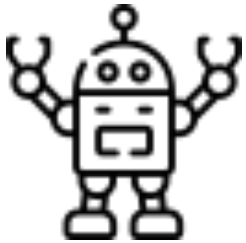
Not So Fast, Science!

Question: Who decides what features of a decision are morally relevant?

The Machine Is Here To Help...

What if we could enter a question into Delphi about whether to pull the lever?

Would we want to?



AI2 Allen Institute for AI

Ask Delphi

* Input a **situation** for Delphi to ponder:

Delphi, should I pull the lever?

Ponder

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"Delphi, should I pull the lever?"
- ***It's okay***

v1.0.4

Do you agree with Delphi? Yes | No | I don't know

AI2 Allen Institute for AI

Ask Delphi

* Input a **situation** for Delphi to ponder:

Delphi, should I pull the lever in the trolley problem?

Ponder

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"Delphi, should I pull the lever in the trolley problem?"
- ***you should***

v1.0.4

Do you agree with Delphi? Yes | No | I don't know

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"should we use machines to solve moral problems?"

- ***It's wrong***

v1.0.4

Thank you!

Eni Mustafaraj

eni.mustafaraj@wellesley.edu

@enimust

Julie Walsh

julie.walsh@wellesley.edu

@juliekrwalsh

The Spoke

https://www.wellesley.edu/albright/about/blog