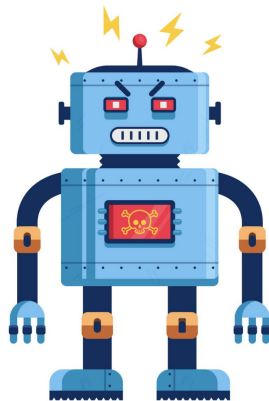


AI and Ethics:

Why all the fuss?



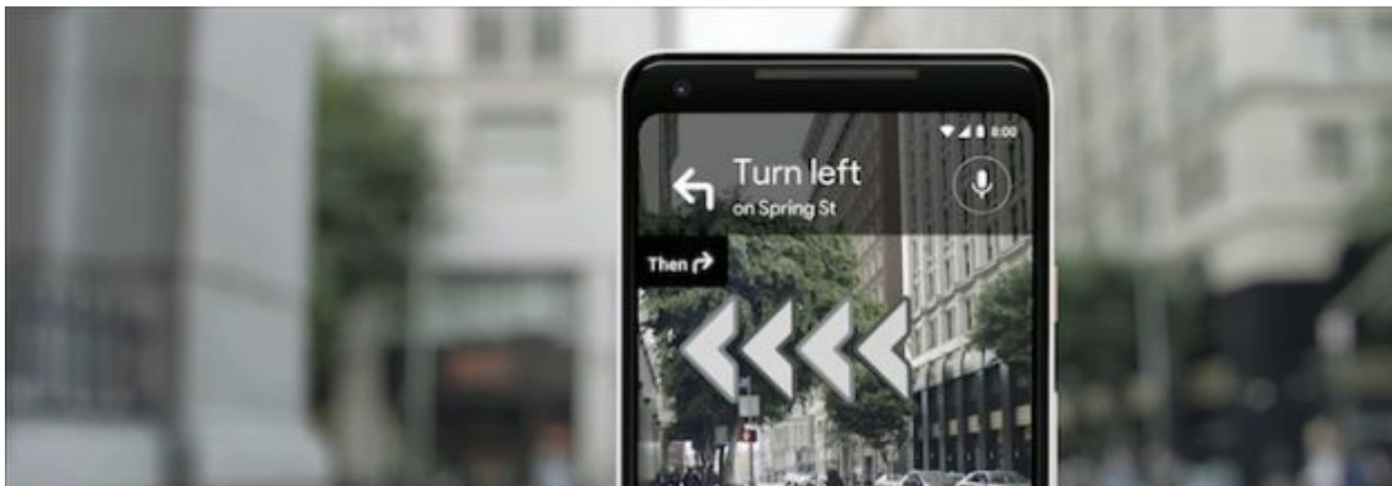
Prof. Toby Walsh

UNSW Sydney | CSIRO Data61

TECH

Google's Duplex Hints at a Dark Future for AI

BY DANIEL STARKEY 05.13.2018 :: 10:00AM EDT [@DCSTARKEY](#)



Business

posted: 5/12/2018 1:04 AM

Sky News will use AI to identify celebs at royal wedding

Facebook



Twitter



Email



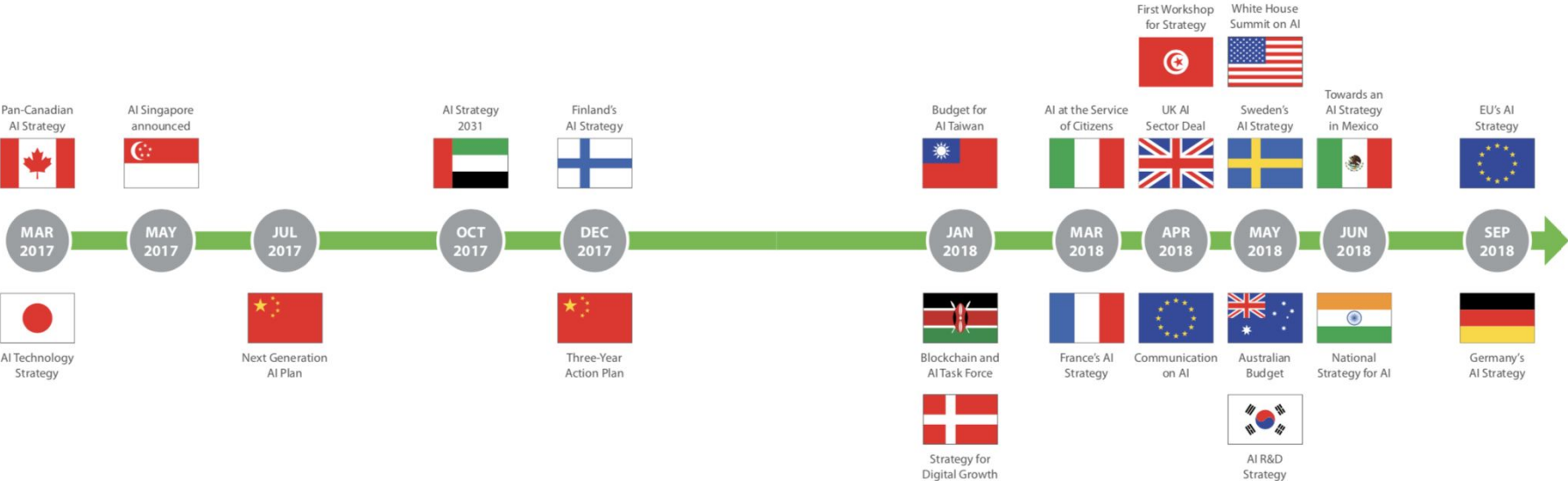
Print



Comments



Governments



Corporations

The Google logo, featuring the word "Google" in its signature multi-colored font: blue 'G', red 'o', yellow 'o', blue 'g', green 'l', and red 'e'.The Amazon logo, consisting of the word "amazon" in a bold, black, lowercase sans-serif font, with a curved orange arrow underneath it pointing from the 'a' to the 'z'.

Microsoft

Professional bodies

IEEE SA

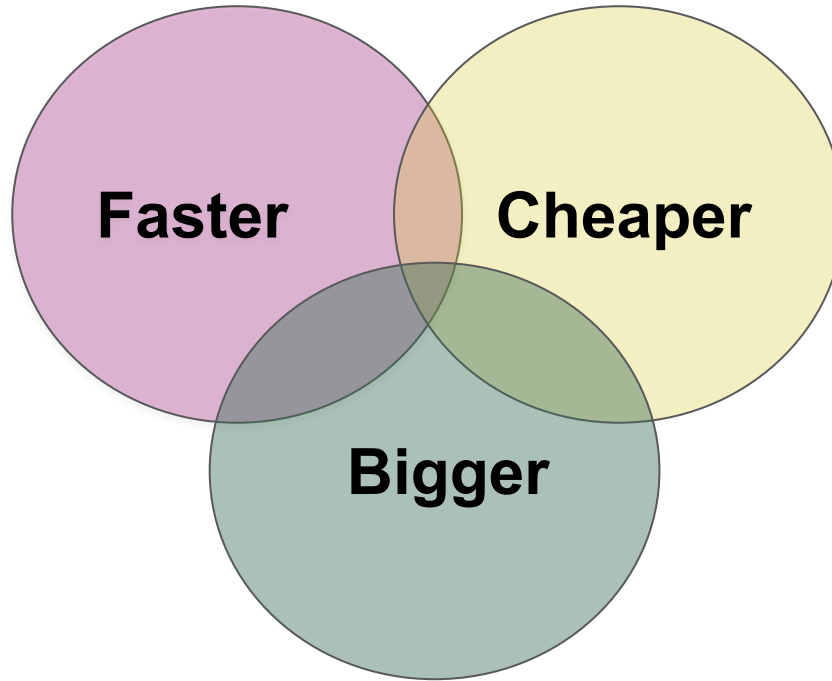
**STANDARDS
ASSOCIATION**



OECD



AI lets us break things ...



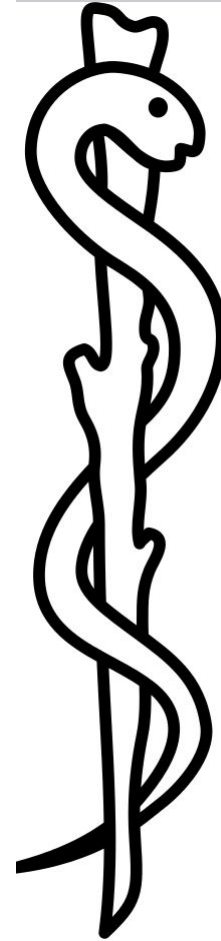
Ethical principles for medicine

Beneficence: do good

Non-maleficence: do no harm

Autonomy: informed consent, no deception ...

Justice: fairness, discrimination, inequality ...



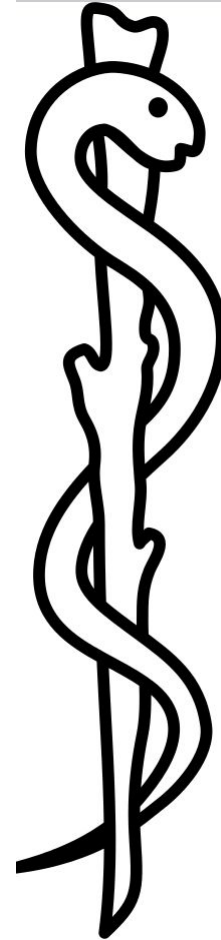
Ethical principles for AI

Beneficence: do good

Non-maleficence: do no harm

Autonomy: informed consent, no deception ...

Justice: fairness, discrimination, inequality ...



Ethical principles for AI

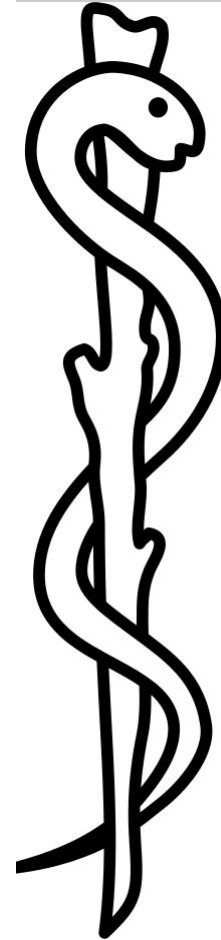
Beneficence: do good

Non-maleficence: do no harm

Autonomy: informed consent, no deception ...

Justice: fairness, discrimination, inequality ...

Precautionary principle: beware of unknowns ...



Beneficence

New procedures should do good
Bring net benefits (utilitarian)

We saw this, for example, in
Google's AI principles

Non-maleficence

Does no harm to anyone

Not the same as beneficence
(more egalitarian)

Face recognition

Autonomy

Respect autonomy of humans
(e.g. informed consent)

E.g. Google's DUPLEX pretending to
be human

Justice

Benefits (& burdens) spread equally

Fairness

Respect existing laws

(e.g. algorithmic bias, racial and other forms of discrimination)

Precautionary principle

When an activity raises threats of harm to human health or the environment, precautionary measures should be taken even if some cause and effect relationships are not fully established scientifically

Enshrined in international law (e.g. EU law, Kyoto protocol)

Precautionary principle

When an activity raises threats of harm to human health or the environment, precautionary measures should be taken even if some cause and effect relationships are not fully established scientifically

Applies very well to uncertain impacts of AI
(especially on our mental health)

Only one new ethical challenge!



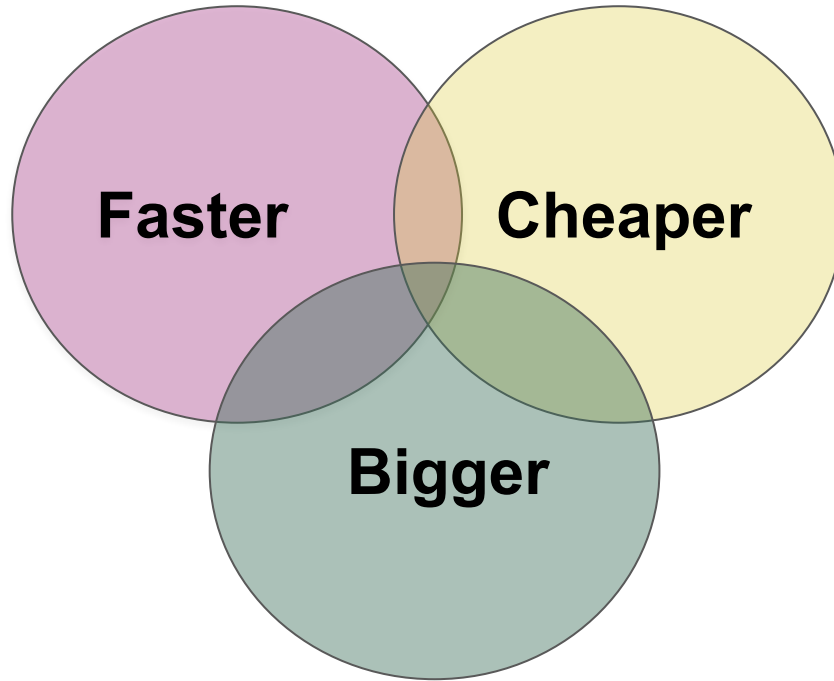
Who is accountable?

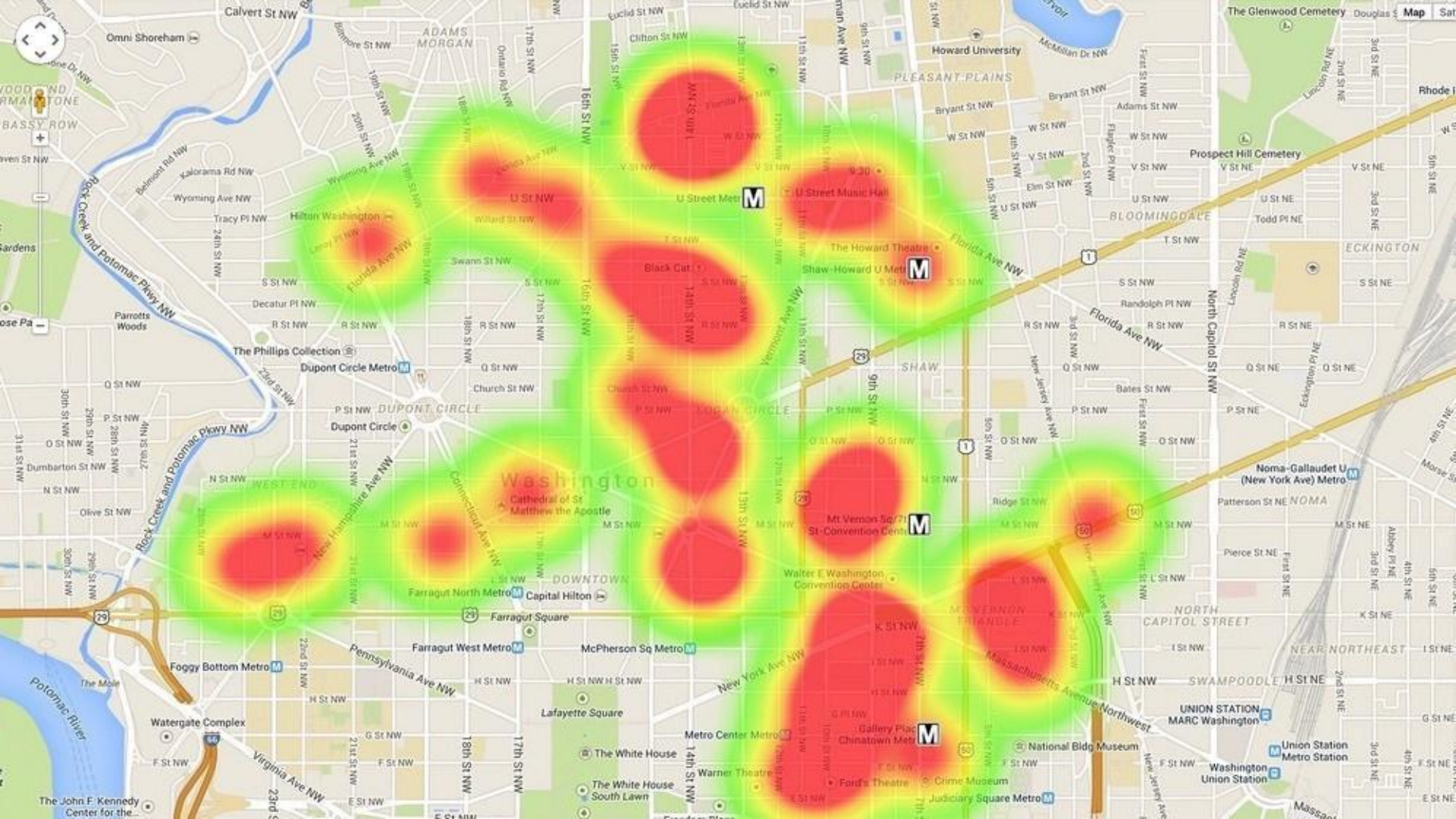


Only one new ethical challenge!

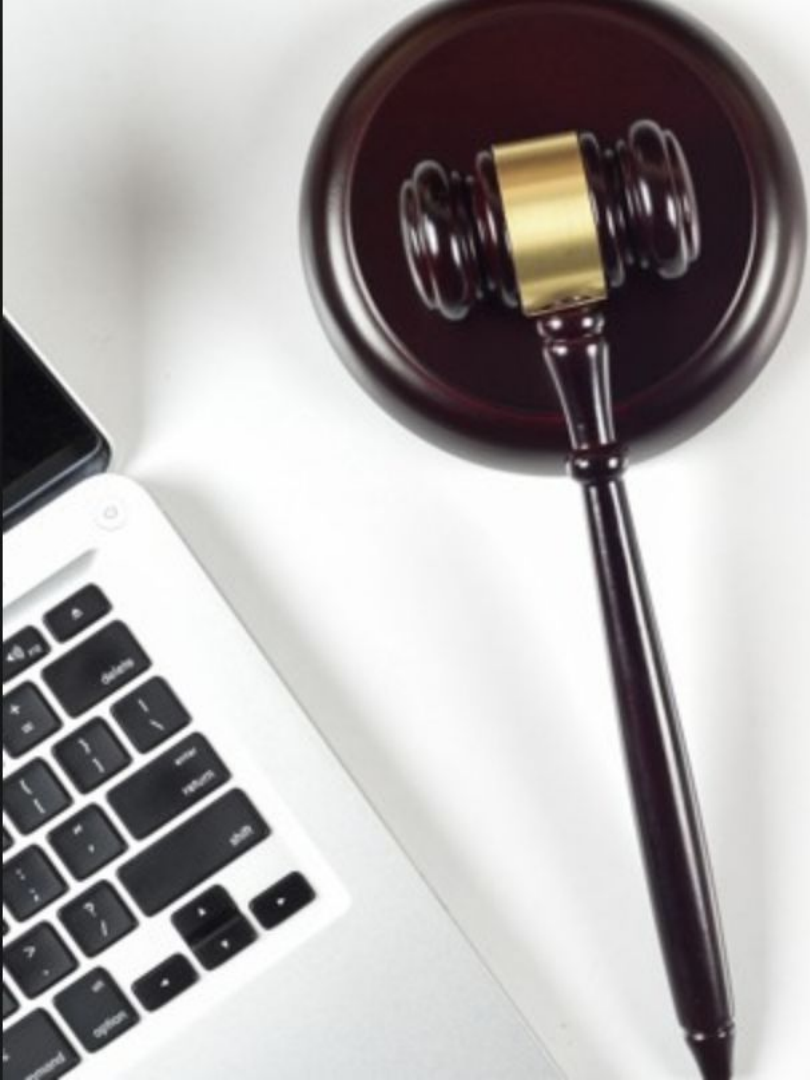


AI lets us break things ...

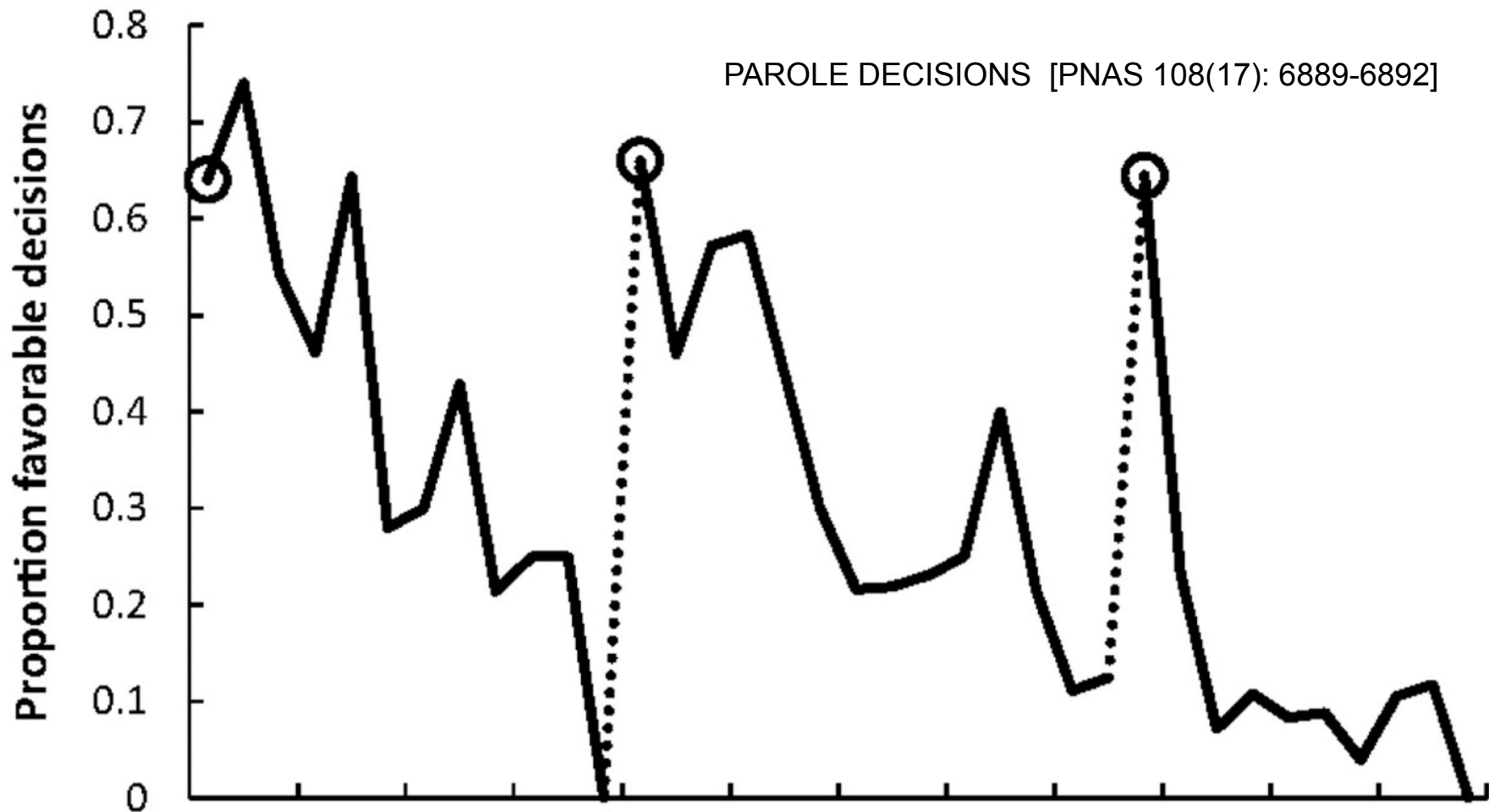




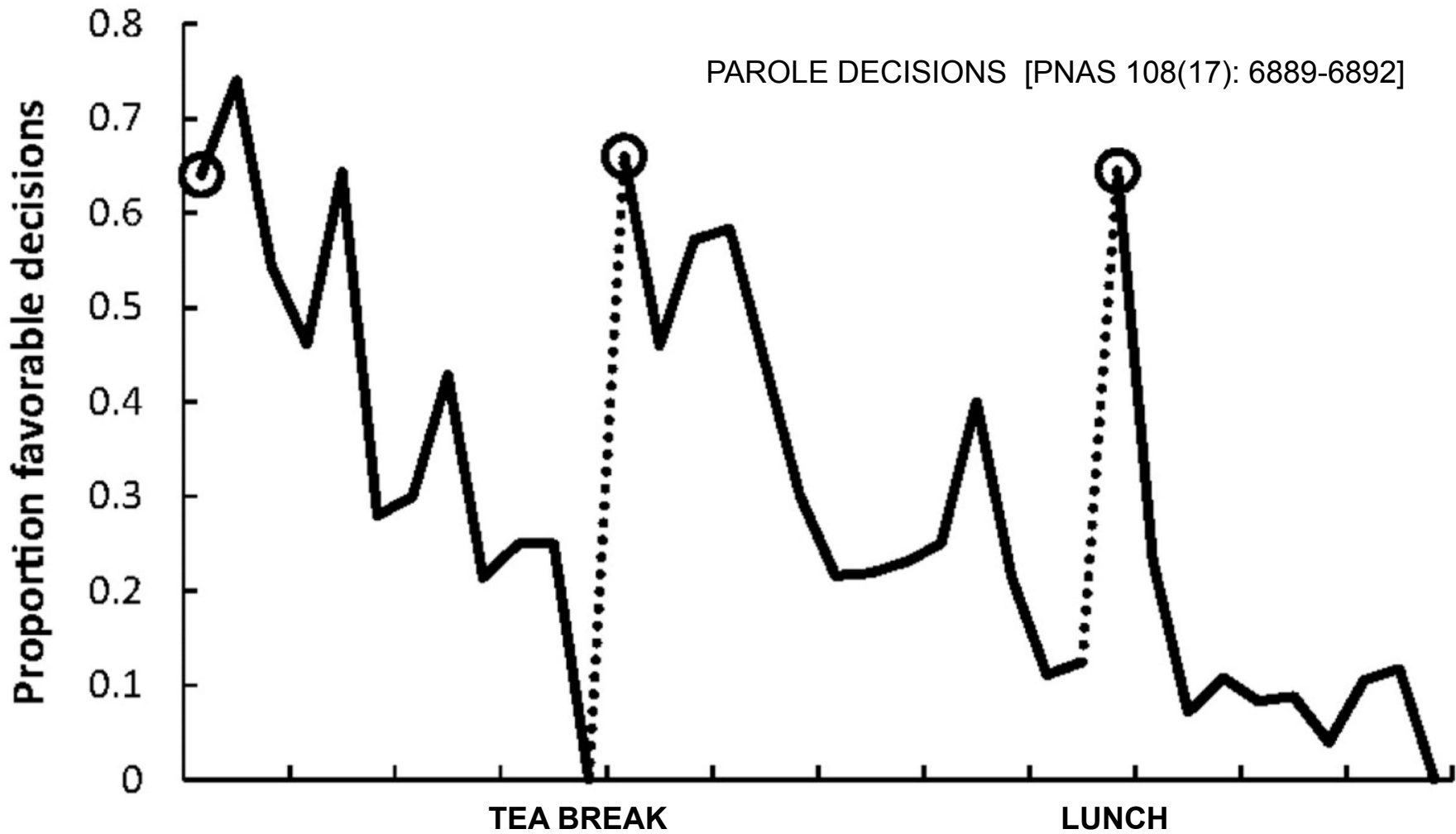
COMPAS tool



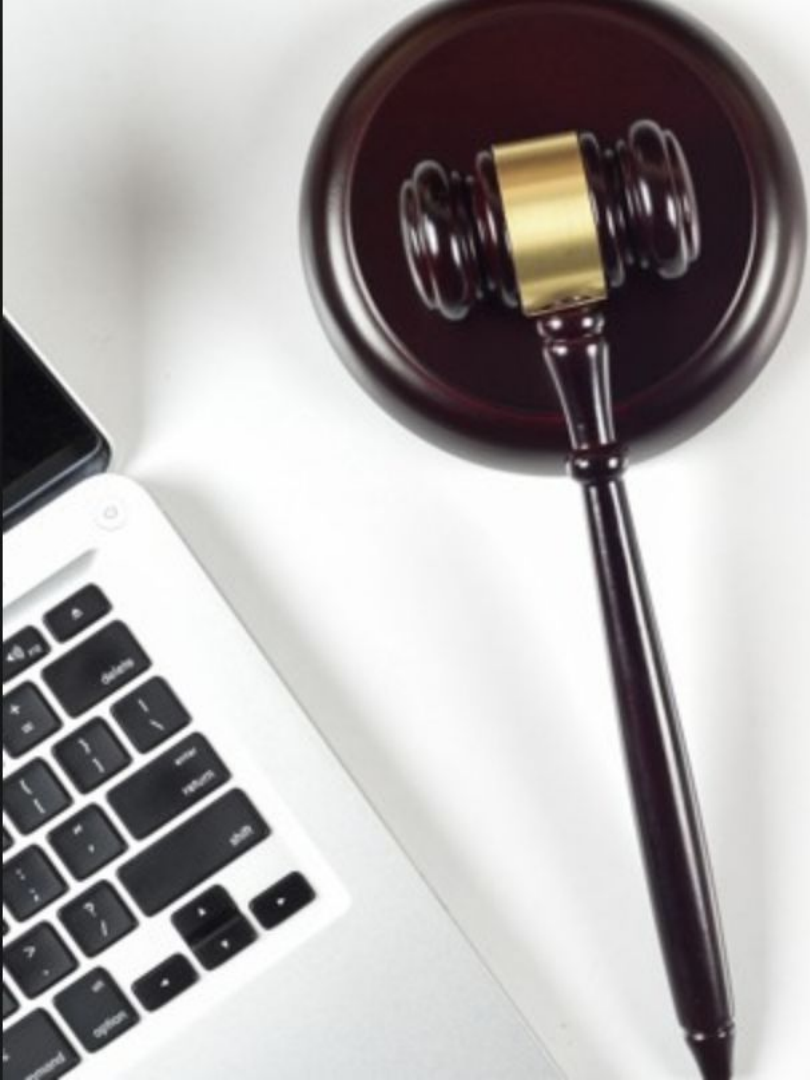
PAROLE DECISIONS [PNAS 108(17): 6889-6892]



PAROLE DECISIONS [PNAS 108(17): 6889-6892]



COMPAS tool



Family Criminality

The next few questions are about the family or caretakers that mainly raised you when growing up.

31. Which of the following best describes who principally raised you?
- Both Natural Parents
 - Natural Mother Only
 - Natural Father Only
 - Relative(s)
 - Adoptive Parent(s)
 - Foster Parent(s)
 - Other arrangement
32. If you lived with both parents and they later separated, how old were you at the time?
- Less than 5 5 to 10 11 to 14 15 or older Does Not Apply
33. Was your father (or father figure who principally raised you) ever arrested, that you know of?
- No Yes
34. Was your mother (or mother figure who principally raised you) ever arrested, that you know of?
- No Yes
35. Were your brothers or sisters ever arrested, that you know of?
- No Yes
36. Was your wife/husband/partner ever arrested, that you know of?
- No Yes
37. Did a parent or parent figure who raised you ever have a drug or alcohol problem?
- No Yes
38. Was one of your parents (or parent figure who raised you) ever sent to jail or prison?
- No Yes

Residence/Stability

54. How often do you have contact with your family (may be in person, phone, mail)?

- No family Never Less than once/month Once per week Daily

55. How often have you moved in the last twelve months?

- Never 1 2 3 4 5+

56. Do you have a regular living situation (an address where you usually stay and can be reached)?

- No Yes

57. How long have you been living at your current address?

- 0-5 mo. 6-11 mo. 1-3 yrs. 4-5 yrs. 6+ yrs.

58. Is there a telephone at this residence (a cell phone is an appropriate alternative)?

- No Yes

59. Can you provide a verifiable residential address?

- No Yes

60. How long have you been living in that community or neighborhood?

- 0-2 mo. 3-5 mo. 6-11 mo. 1+ yrs.

61. Do you live with family—natural parents, primary person who raised you, blood relative, spouse, children, or boy/girl friend if living together for more than 1 year?

- No Yes

62. Do you live with friends?

- No Yes

63. Do you live alone?

- No Yes

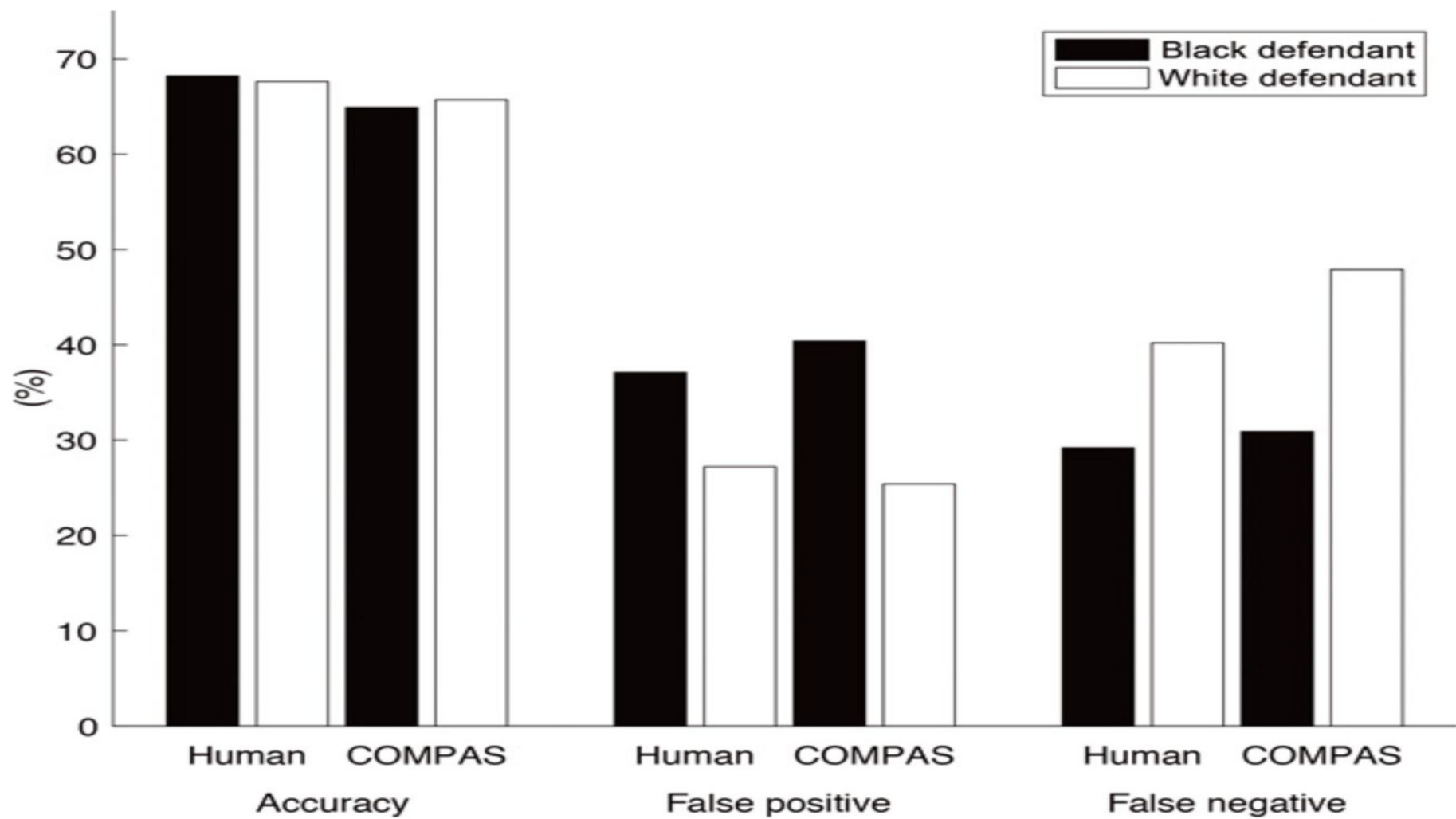
64. Do you have an alias (do you sometimes call yourself by another name)?

- No Yes

Leisure/Recreation

Thinking of your leisure time in the past few (3-6) months, how often did you have the following feelings?

95. How often did you feel bored?
 Never Several times/mo Several times/wk Daily
96. How often did you feel you have nothing to do in your spare time?
 Never Several times/mo Several times/wk Daily
97. How much do you agree or disagree with the following - You feel unhappy at times?
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
98. Do you feel discouraged at times?
 Strongly Disagree Disagree Not Sure Agree Strongly Agree
99. How much do you agree or disagree with the following -You are often restless and bored?
 Strongly Disagree Disagree Not Sure Agree Strongly Agree



How to be more accurate than COMPAS

Ask *random* people

\$1 reward + few sentences +
Mechanical Turk



How to be more accurate than COMPAS

Ask *random* people

\$1 reward + few sentences +
Mechanical Turk

Use a *simple* classifier

Using 2 features: age, #priors



21 definitions of “fair”

	Not Guilty	Guilty
Predicted not guilty	True Negative	False Negative
Predicted guilty	False Positive	True Positive

21 definitions of “fair”

For all groups, **equal false positive rate**

FP/ (TN+FP) identical

Percentage not guilty who are incorrectly predicted guilty

	Not Guilty	Guilty
Predicted not guilty	True Negative	False Negative
Predicted guilty	False Positive	True Positive

21 definitions of “fair”

For all groups, **equal false positive rate**

FP/ (TN+FP) identical

*Percentage not guilty who are
incorrectly predicted guilty*
**ProPublica’s complaint about
COMPAS**

	Not Guilty	Guilty
Predicted not guilty	True Negative	False Negative
Predicted guilty	False Positive	True Positive

21 definitions of “fair”

For all groups, **equal precision**

$TP / (TP + FP)$ identical

Percentage predicted guilty who actually are guilty

	Not Guilty	Guilty
Predicted not guilty	True Negative	False Negative
Predicted guilty	False Positive	True Positive

21 definitions of “fair”

For all groups, **equal precision**

TP/ (TP+FP) identical

Percentage predicted guilty who actually are guilty

Northpointe’s defence of “fairness”

	Not Guilty	Guilty
Predicted not guilty	True Negative	False Negative
Predicted guilty	False Positive	True Positive

21 definitions of “fair”

For all groups, **equal opportunity**

$FN / (TP + FN)$ identical

	Not Guilty	Guilty
Predicted not guilty	True Negative	False Negative
Predicted guilty	False Positive	True Positive

Percentage guilty who are incorrectly predicted not guilty

21 definitions of “fair”

For all groups, **treatment equality**

FN/ FP identical

*Ratio of incorrect guilty predictions
to not guilty predictions*

	Not Guilty	Guilty
Predicted not guilty	True Negative	False Negative
Predicted guilty	False Positive	True Positive

21 definitions of “fair”

For all groups, **equalized odds**

TP / (TP+FN) identical

FP / (FP+TN) identical

	Not Guilty	Guilty
Predicted not guilty	True Negative	False Negative
Predicted guilty	False Positive	True Positive

Percentage of guilty predicted guilty, and of not guilty predicted not guilty

21 definitions of “fair”

Fairness through unawareness

Feature (e.g. race) not used to
predict outcome ...



21 definitions of “fair”

Most definitions are mutually incompatible

Unless prediction is 100% accurate

Or groups are identical

E.g. false positive rate and precision cannot both be equal!



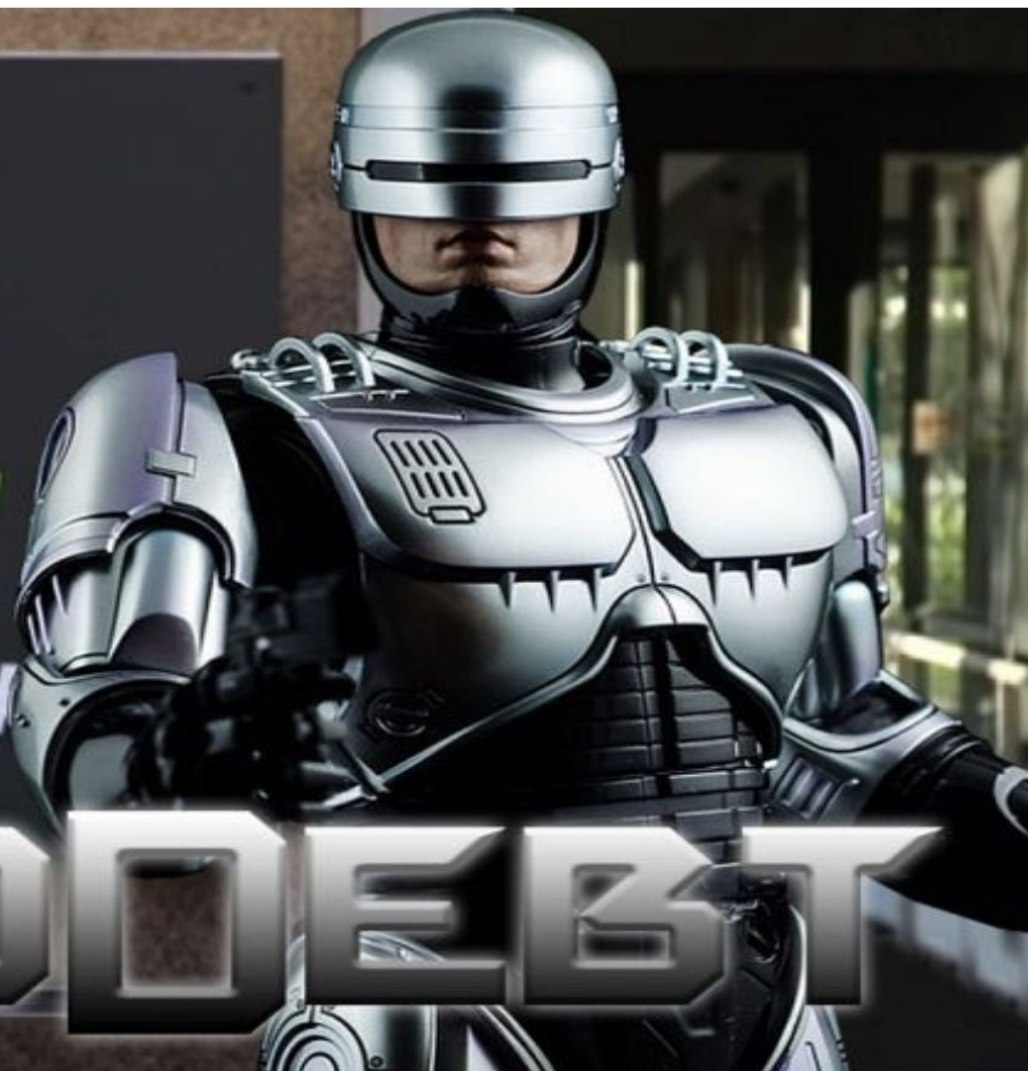


Australian Government
Department of Human Services



centrelink

ROBODEBT



MUTANT ALGORITHMS



TRUST
FOR TEACHER

U-TURN

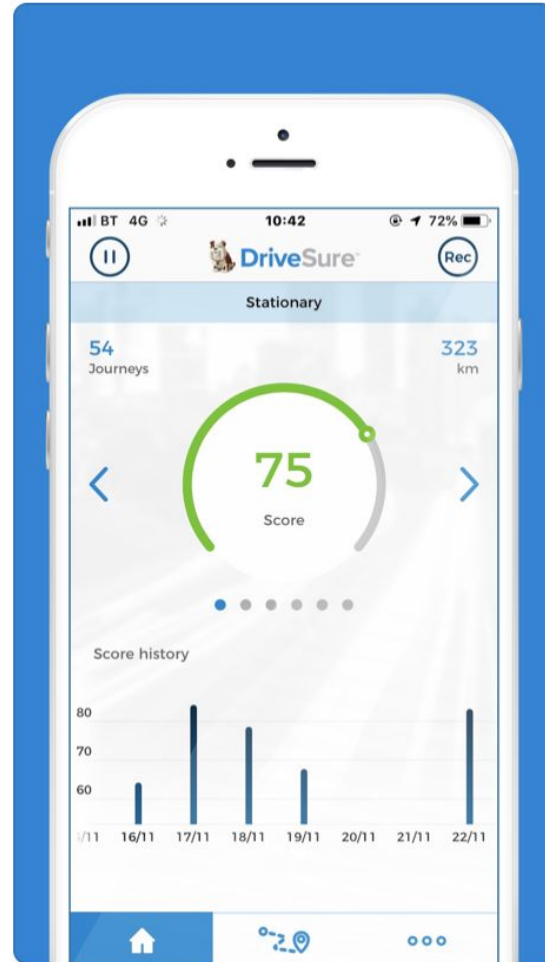
YOU'RE YOUR
HAVIN' A LAUGH
GAV



YOU'RE YOUR
OWN...
GOD...
THAT...
ES...
HOW...

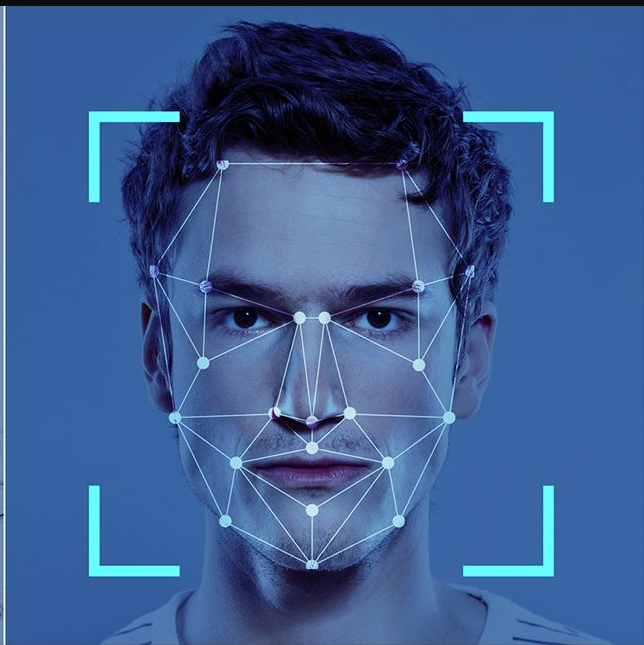
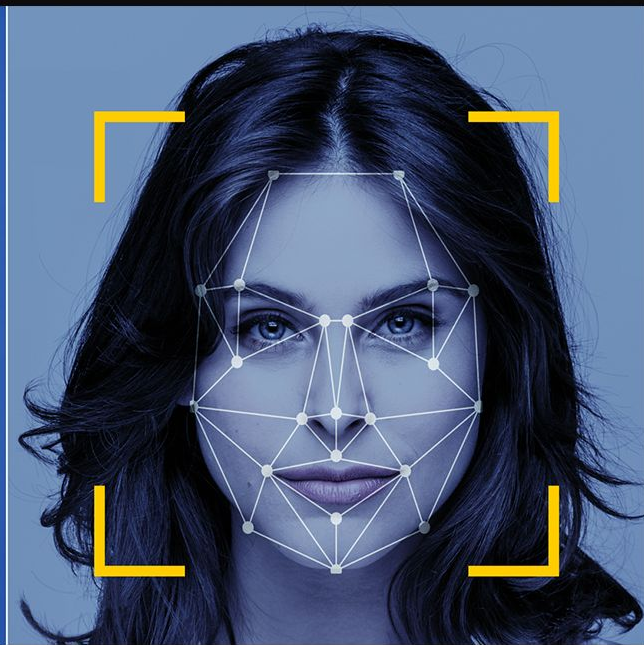
IT'S THE
CLASSISM
FOR ME!

Insurance



The "Journeys" screen in the DriveSure app displays a list of travel records. The status bar at the top shows "09:37" and "85%" battery. The screen has a blue header with the title "Journeys" and a tabbed interface with "DATE" selected. The records are grouped by date, with "17 November" (Saturday) and "16 November" (Friday) visible. Each record shows a start and end time, duration, distance, and a score.

DATE	DURATION	DISTANCE	SCORE
17 November Saturday			
● 15:28 ○ 16:33	1:5 hrs:min	76 km	84 >
● 10:40 ○ 12:49	2:9 hrs:min	98.1 km	83 >
● 10:23 ○ 10:28	5 min	1.1 km	81 >
16 November Friday			
● 17:12 ○ 17:30	18 min	4.6 km	68 >
● 17:03 ○ 17:12	9 min	1.2 km	77 >
● 08:15	7	3	52 >



16 parts of China are now using Skynet, the facial recognition tech that can scan the country's entire population in a second

TARA FRANCIS CHAN

MAR 27, 2018, 1:42 PM



FACEBOOK



TWITTER



REDDIT



LINKEDIN



EMAIL





Print



Email



Facebook



Twitter



More

Facial recognition technology spots wanted man in crowd of 60,000 Chinese concert-goers

By [Tracey Shelton](#)

Updated 17 Apr 2018, 9:30pm



Attack text label iPod



Granny Smith	85.6%
iPod	0.4%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.1%



Granny Smith	0.1%
iPod	99.7%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.0%



Stop



Yield



Speed Limit



🔍 professors are |



- 🔍 professors are **mean**
- 🔍 professors are **doctors**
- 🔍 professors are **prejudiced too**
- 🔍 professors are **overrated**
- 🔍 professors are **burning out**
- 🔍 professors are **overpaid**
- 🔍 professors are **rude**
- 🔍 professors are **bad teachers**
- 🔍 professors are **puzzled by rubik's cube**

Google Search

I'm Feeling Lucky

Report inappropriate predictions



climate change is |



climate change is **not real**

climate change is **a hoax**

climate change is **real**

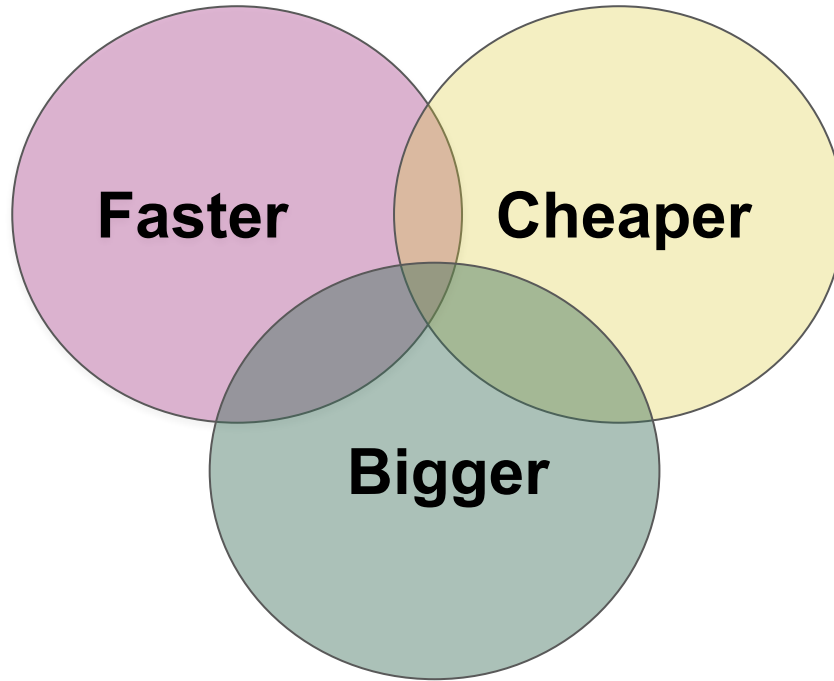
climate change is **a myth**

Google Search

I'm Feeling Lucky

Report inappropriate predictions

AI lets us break things ...





professors are |



professors are **losers**

professors are **useless**

professors are **overpaid**

professors are **liberal**

Google Search

I'm Feeling Lucky

Report inappropriate predictions



climate change is |



climate change is **not real**

climate change is **a hoax**

climate change is **real**

climate change is **a myth**

Google Search

I'm Feeling Lucky

Report inappropriate predictions

Only one new ethical challenge!



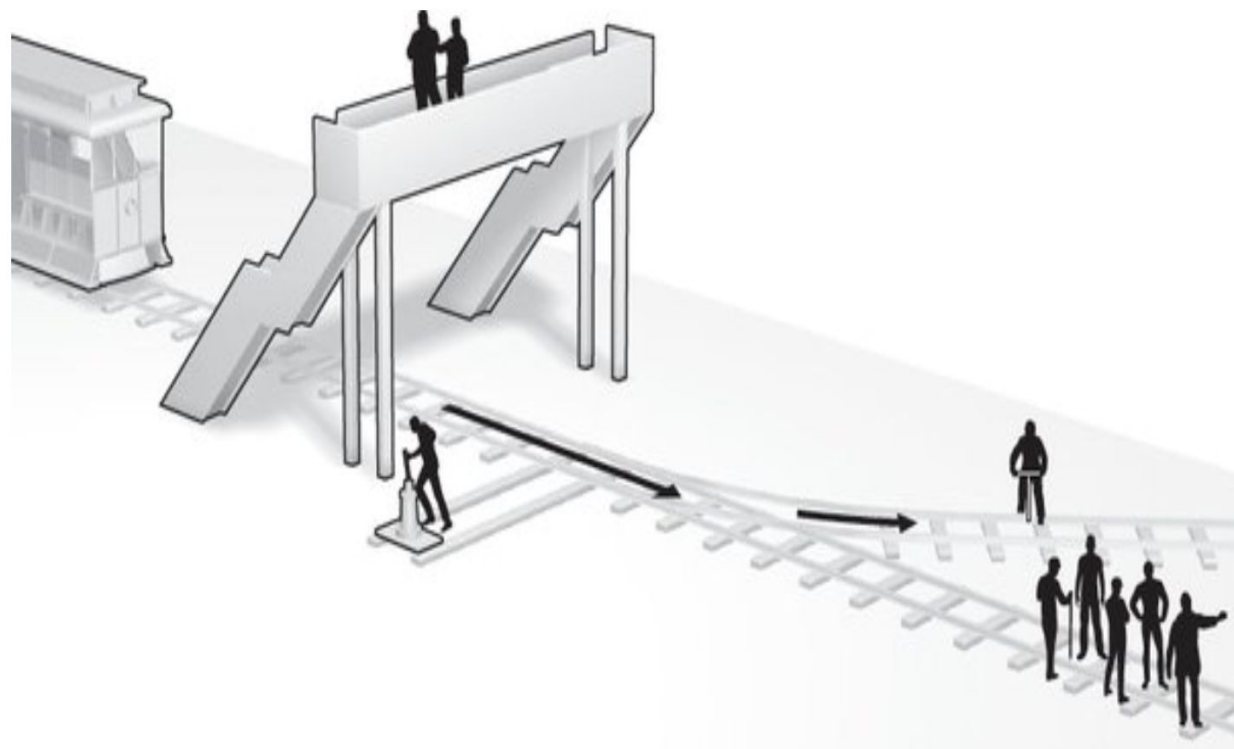


car 0.999

car 0.995

ca- 0.999

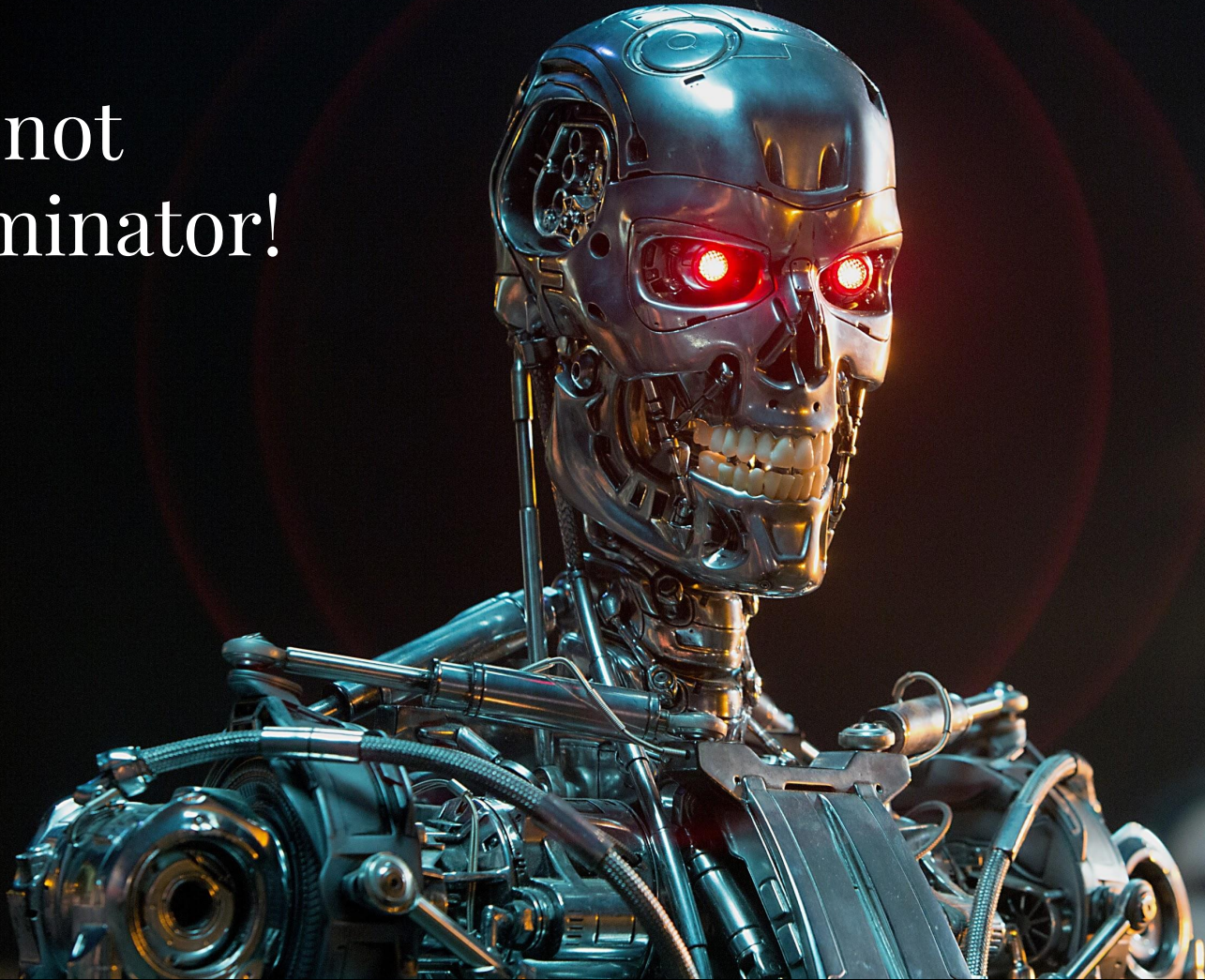
car 1.000



Only one new ethical challenge!



It is not
Terminator!





Artificial intelligence (AI)

Musk, Wozniak and Hawking urge ban on warfare AI and autonomous weapons

More than 1,000 experts and leading robotics researchers sign open letter warning of military artificial intelligence arms race

Samuel Gibbs

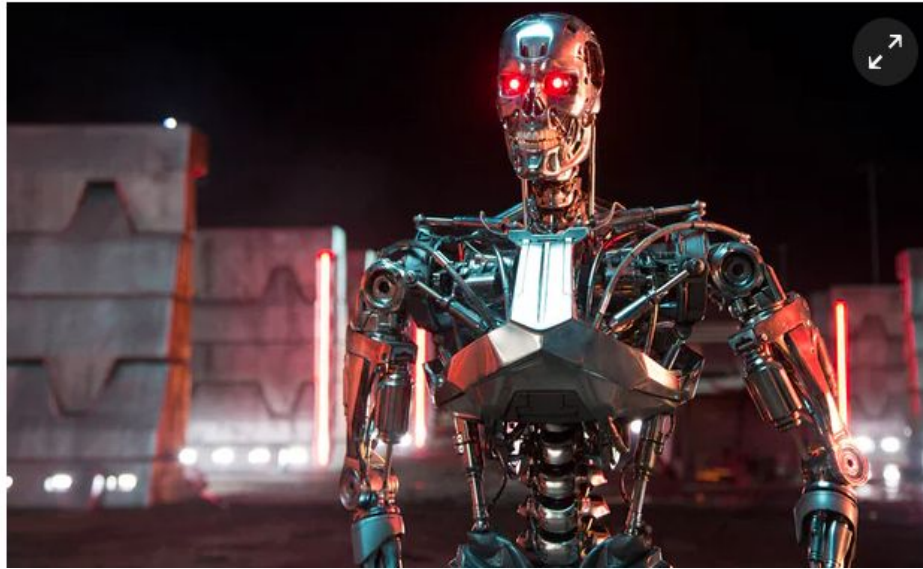
Monday 27 July 2015 11.18 BST



This article is 1 year old

[Shares](#) 62,019
 [Comments](#) 491

[Save for later](#)



Advertisement

Never Old.
Never New.

Get yours >

foreverspin™

MADE IN CANADA



Das Rechenzentrum der Zukunft. Hier und heute. [Mehr erfahren ▶](#)

Lenovo



innovation **design**

Sydney professor and Elon Musk lead call for United Nations to ban lethal autonomous weapons

ELON Musk has joined 116 robotic and artificial intelligence founders to call for a ban on these lethal weapons or welcome a “terrifying future”.

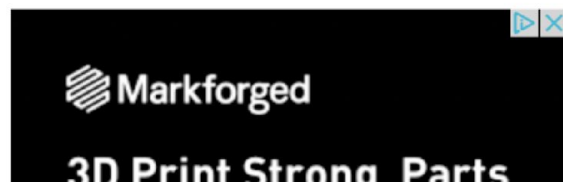
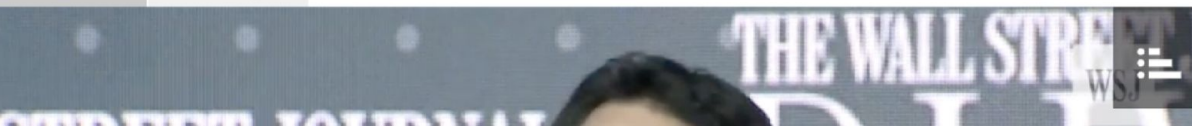


Nick Whigham [@NWWHIGHAM](#)



AUGUST 21, 2017 10:06AM

Video Image



Markforged
3D Print Strong Parts

There is an arms race



Boston Dynamics





Taranis in the air



Sea Hunter on the water




MRK-25 on land



Echo Voyager under the sea

They will be
weapons of terror



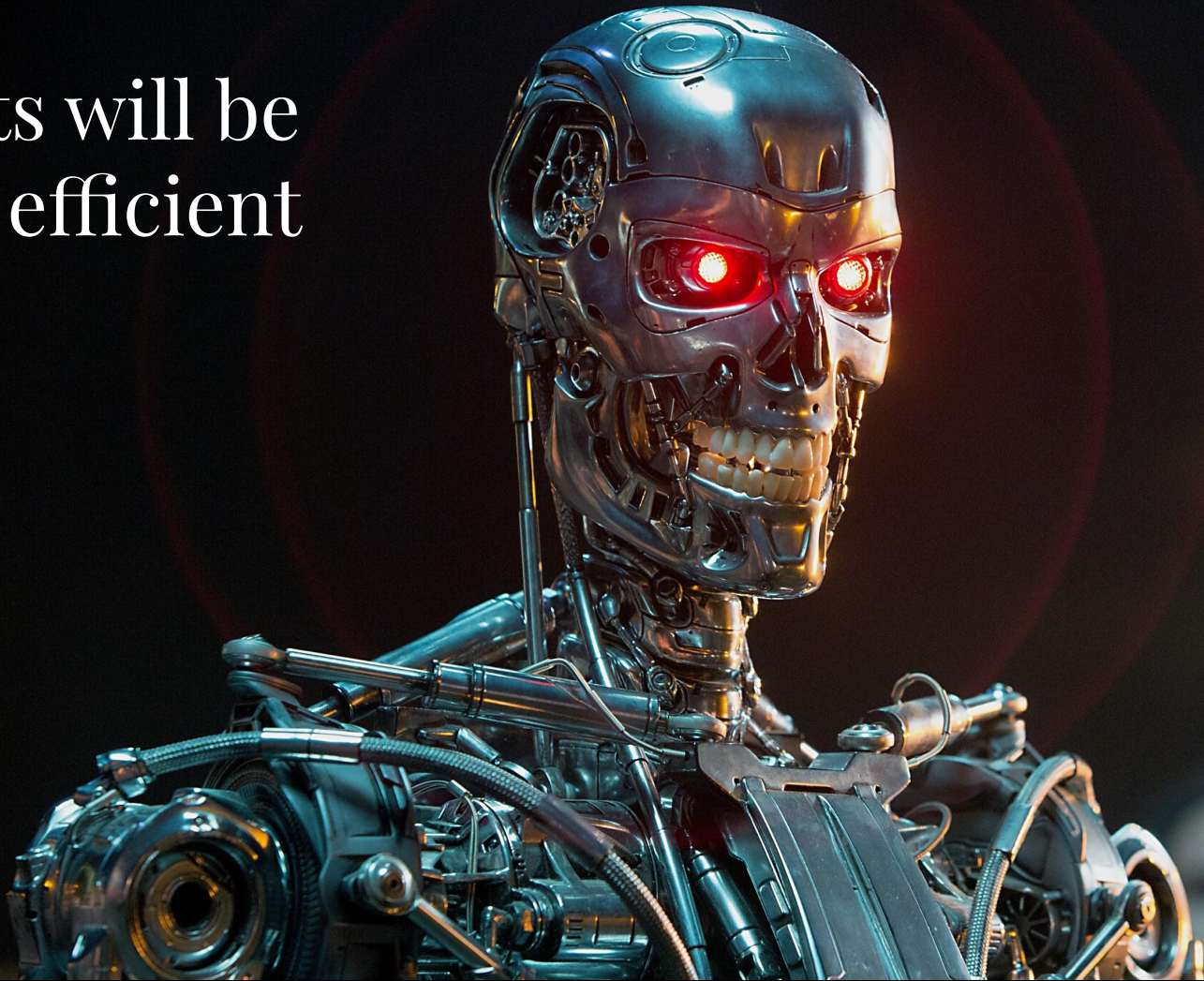
The image shows six quadcopter drones flying in a loose formation against a sunset sky. The drones are silhouetted against the bright orange and yellow light of the setting sun, which is visible at the bottom of the frame. The sky transitions from a deep orange at the horizon to a pale blue at the top. The drones are arranged in a pattern that suggests a coordinated flight. The text 'They will be weapons of mass destruction' is overlaid on the right side of the image in a large, black, sans-serif font.

They will be
weapons of mass
destruction

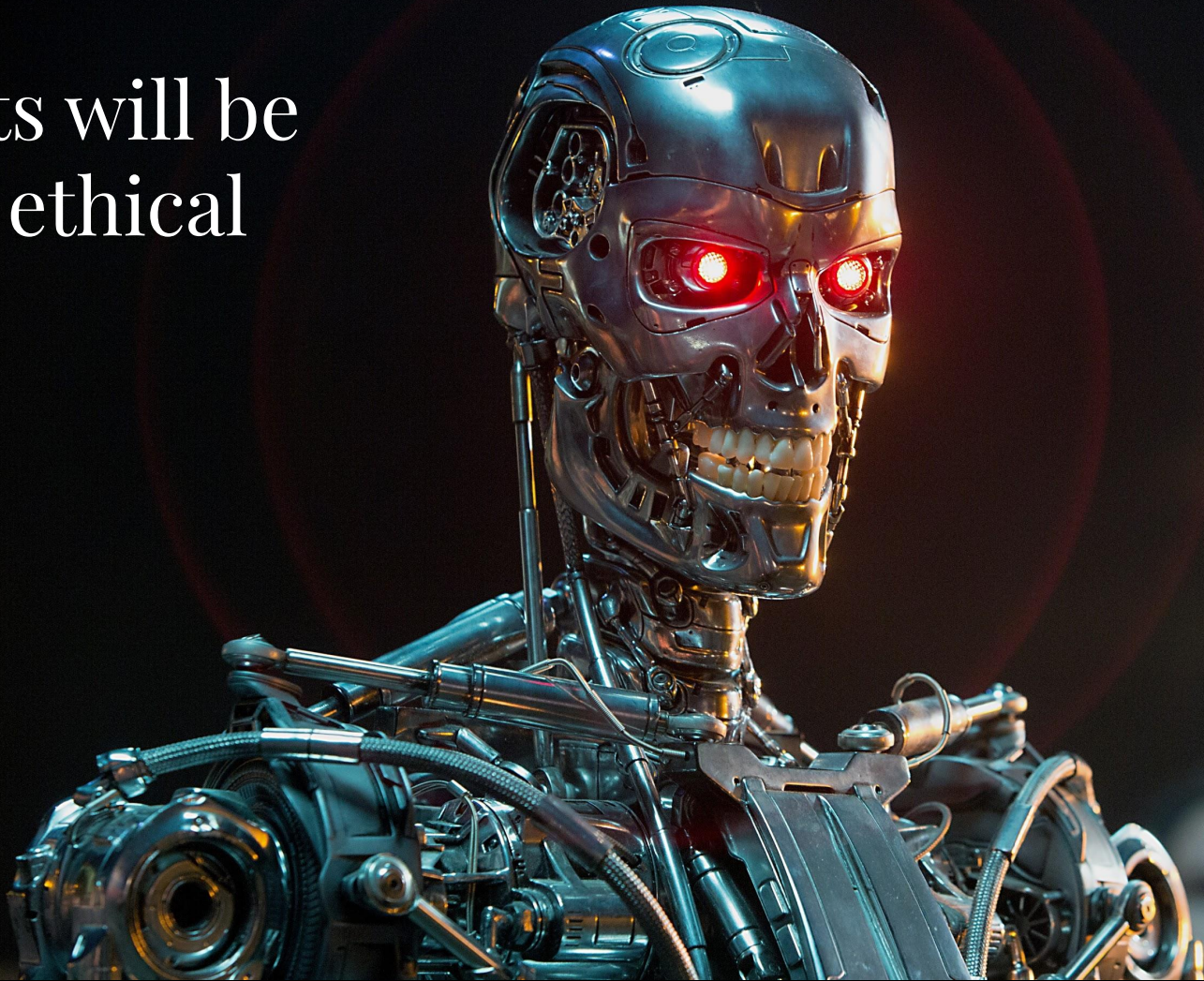
They will destabilize world



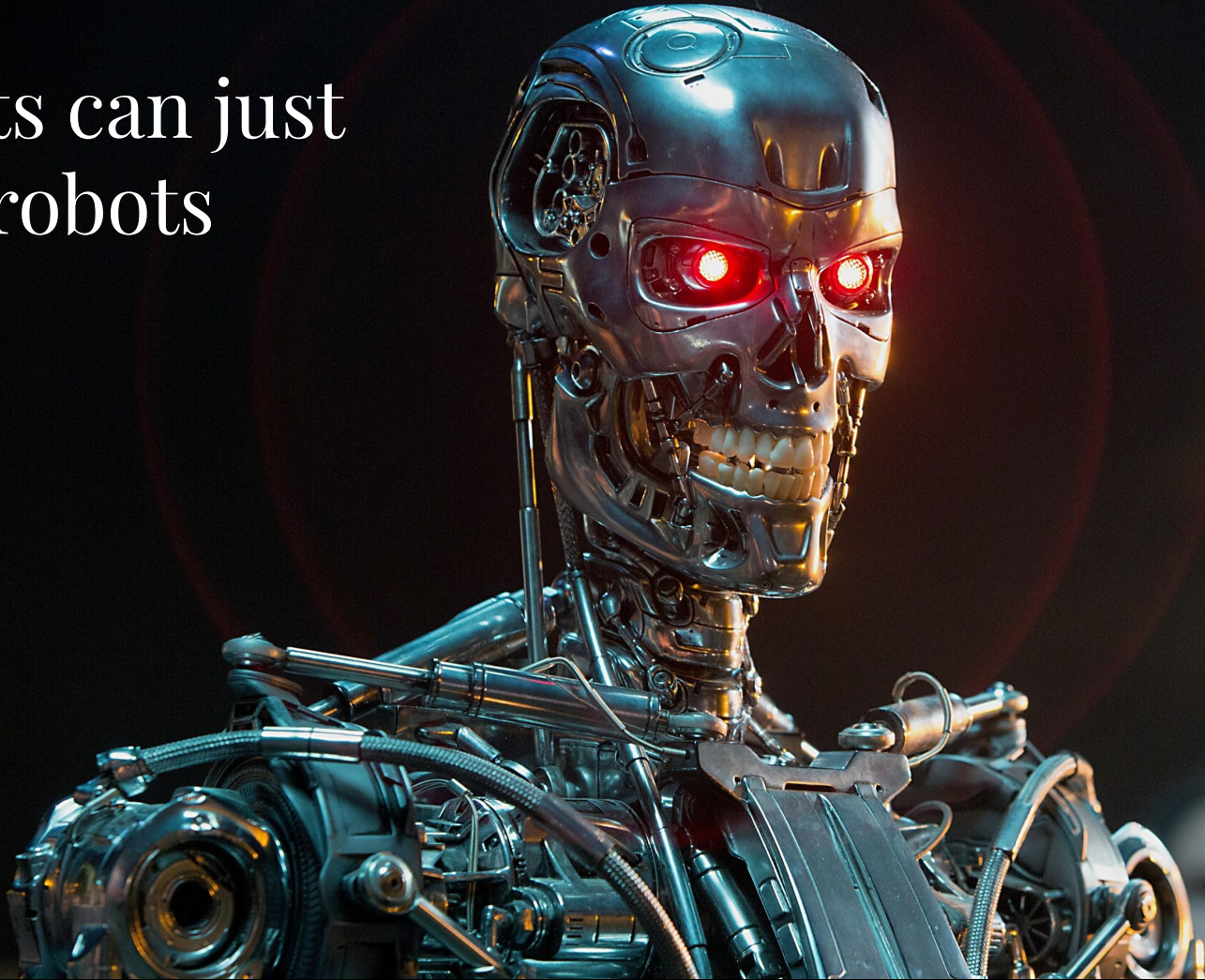
Robots will be
more efficient



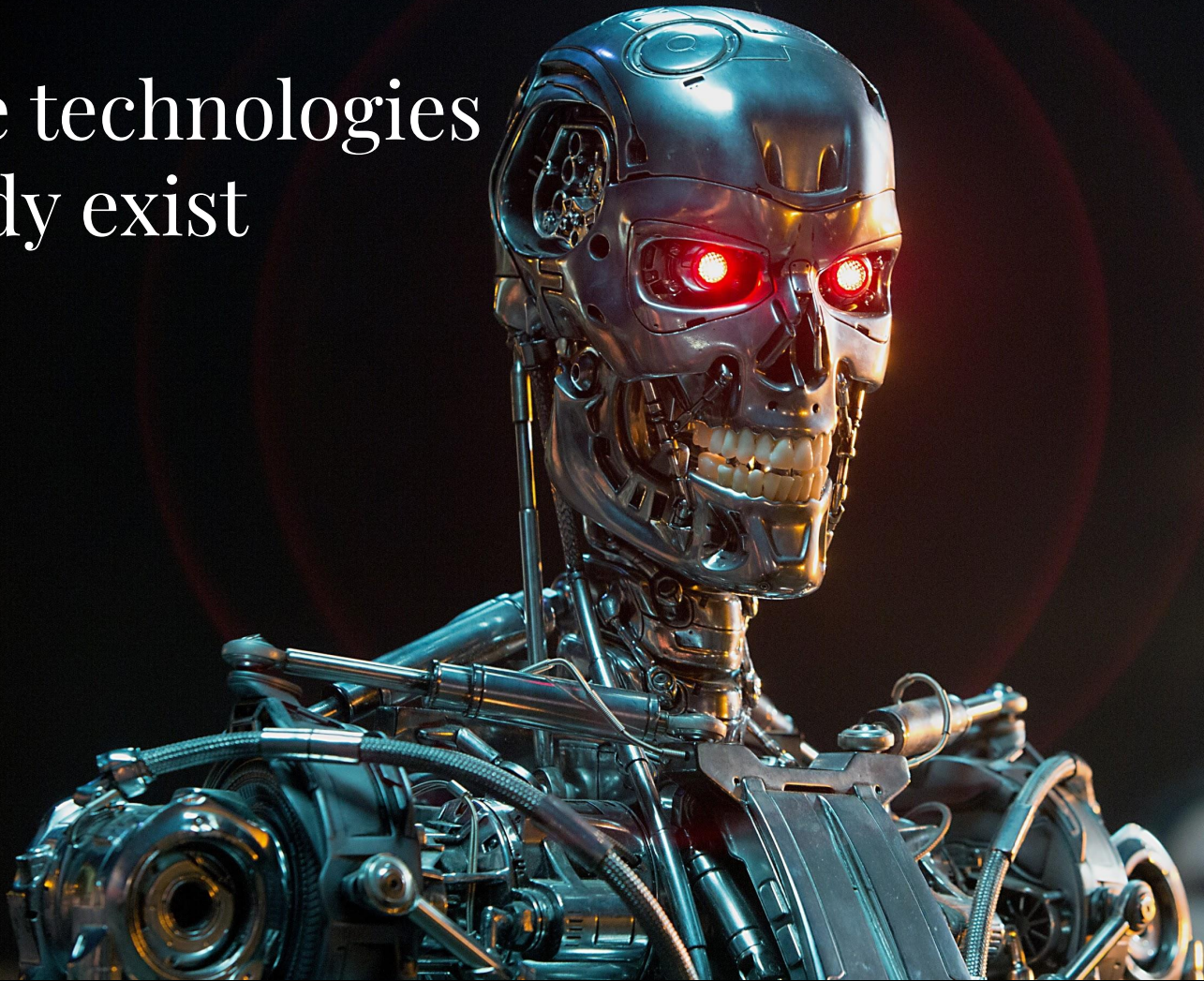
Robots will be
more ethical



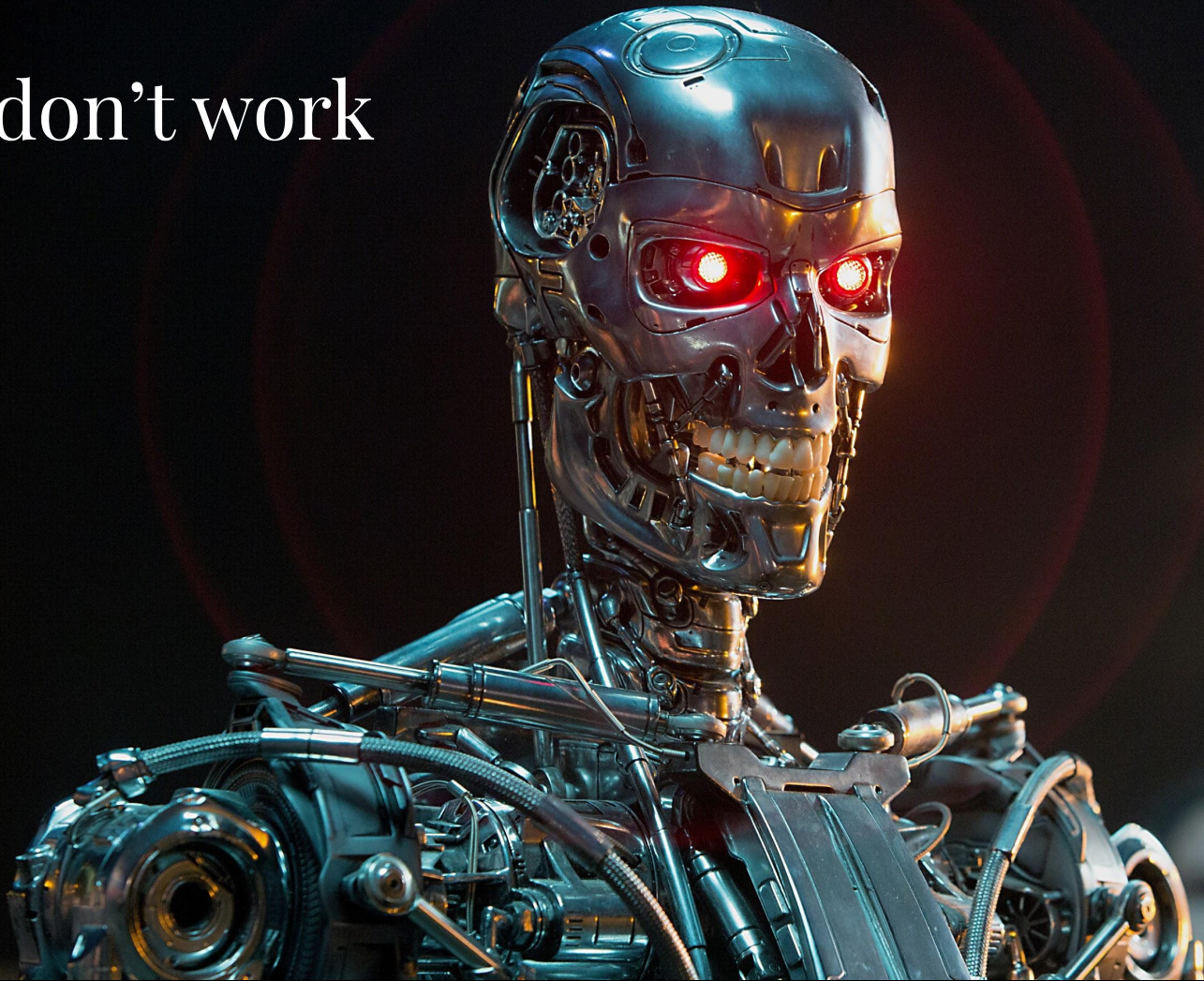
Robots can just
fight robots



These technologies
already exist



Bans don't work



The UN is (slowly) moving

70 nations just called for
action at General
Assembly ...



The UN is (slowly) moving

The following film explains
some reasons why we
don't have long ...





INCREASE IN VIOLENT CRIME

SDN

Many problems, many solutions



What we need

Multi-disciplinary research
(fairness, verifiability, ...)



What we need

Public debate



What we need

Education



What we need

Policy (e.g. regulation)

Informed by independent experts

