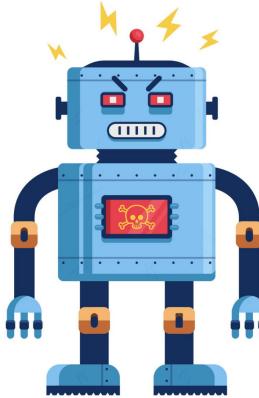


# AI and Ethics:

## Why all the fuss?

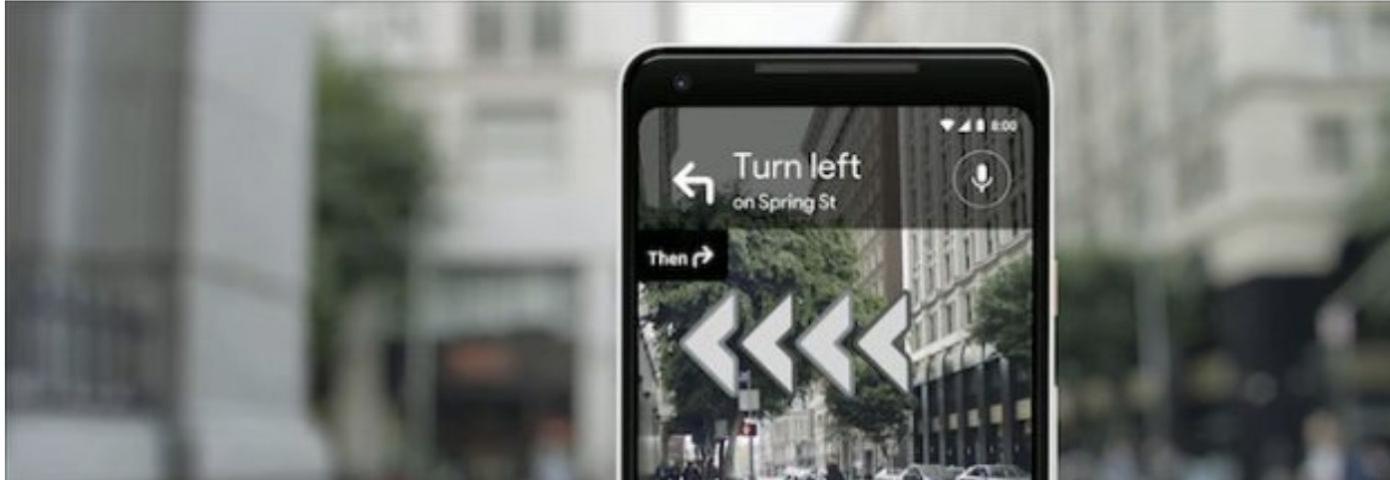


Prof. Toby Walsh  
UNSW Sydney | CSIRO Data61

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

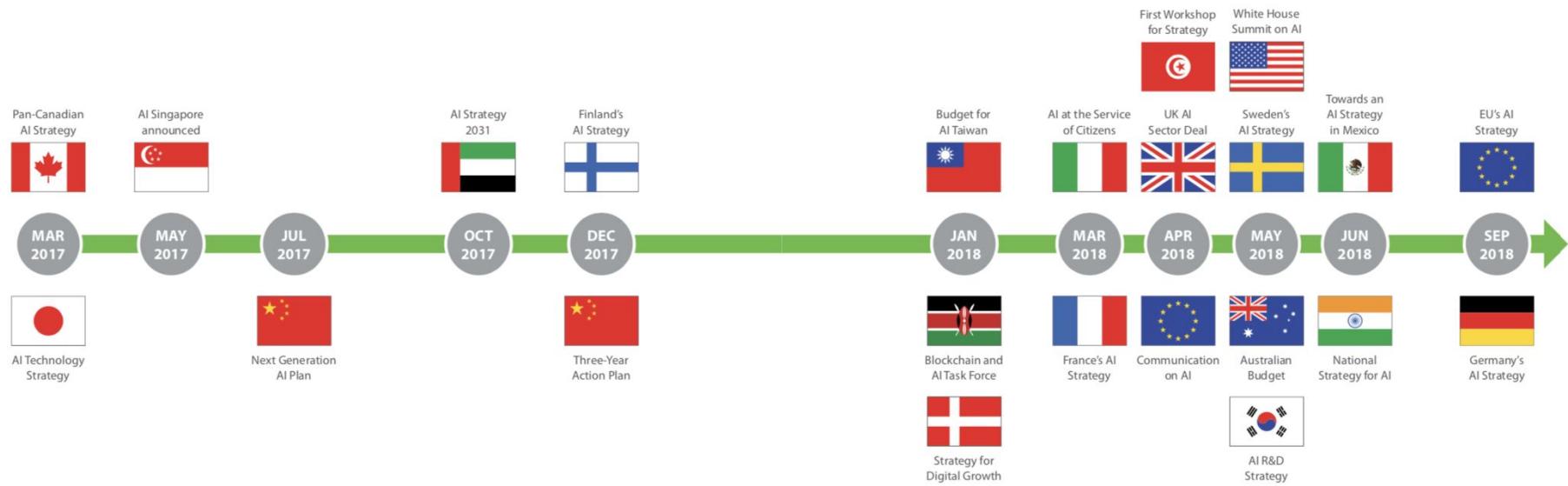
Comments



0



# Governments



# Corporations



Microsoft

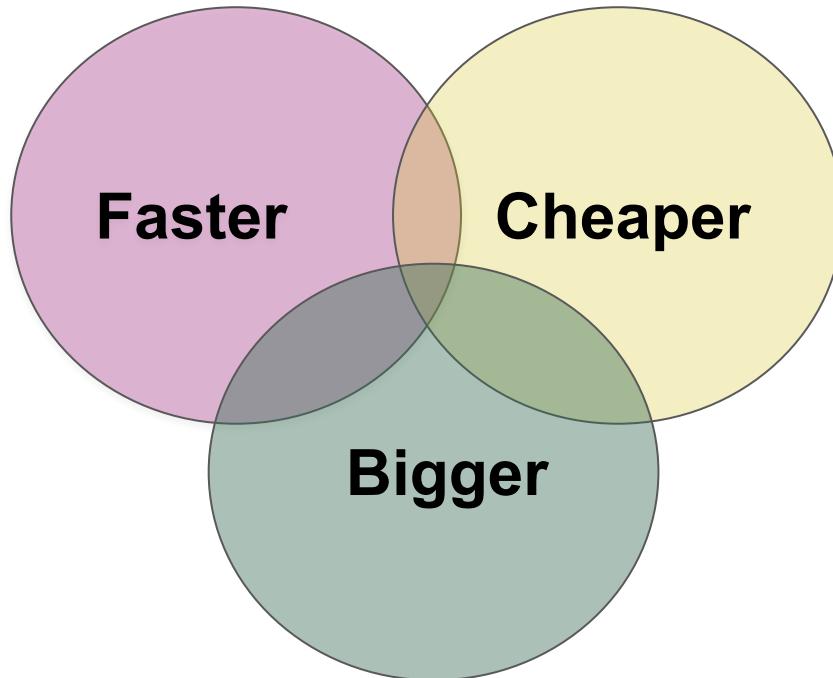
## Professional bodies

**IEEE SA**

**STANDARDS  
ASSOCIATION**



# AI lets us break things ...



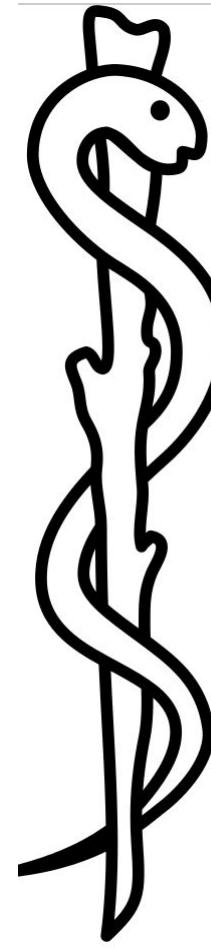
# Ethical principles for medicine

**Beneficence:** do good

**Non-maleficence:** do no harm

**Autonomy:** informed consent, no deception ...

**Justice:** fairness, discrimination, inequality ...



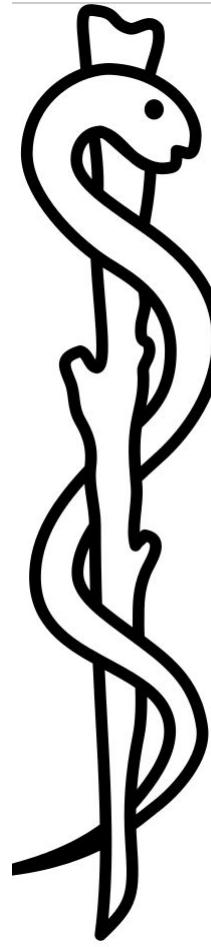
# Ethical principles for AI

**Beneficence:** do good

**Non-maleficence:** do no harm

**Autonomy:** informed consent, no deception ...

**Justice:** fairness, discrimination, inequality ...



# Ethical principles for AI

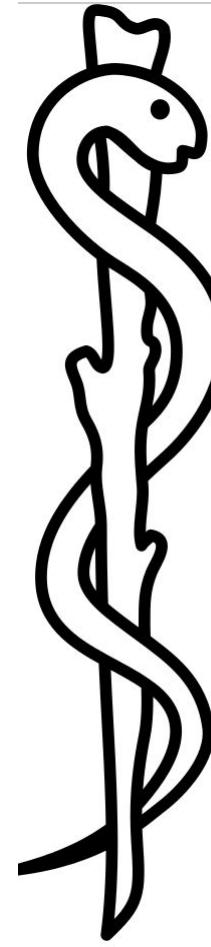
**Beneficence:** do good

**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!



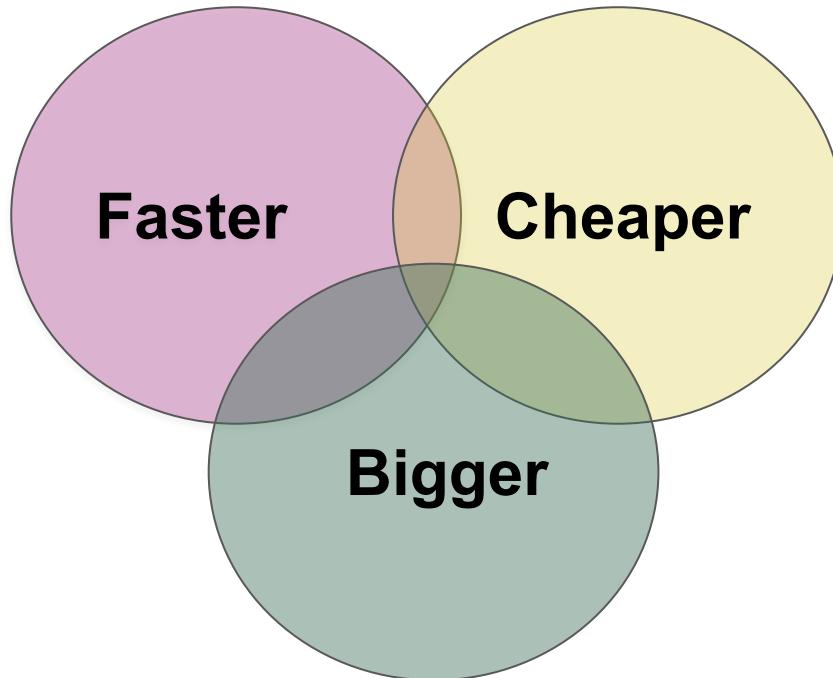
# Who is accountable?

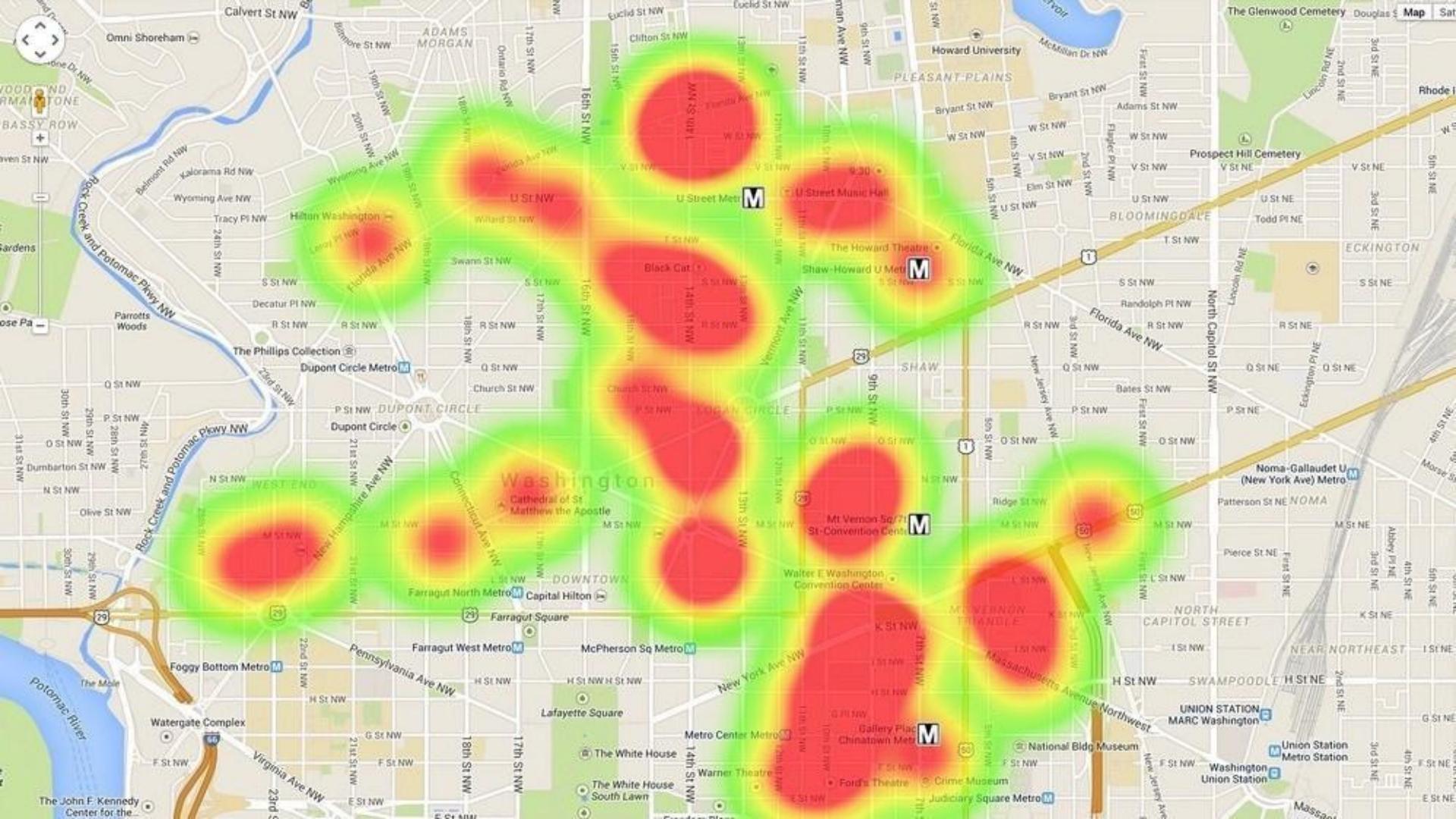


# Only one new ethical challenge!



# AI lets us break things ...

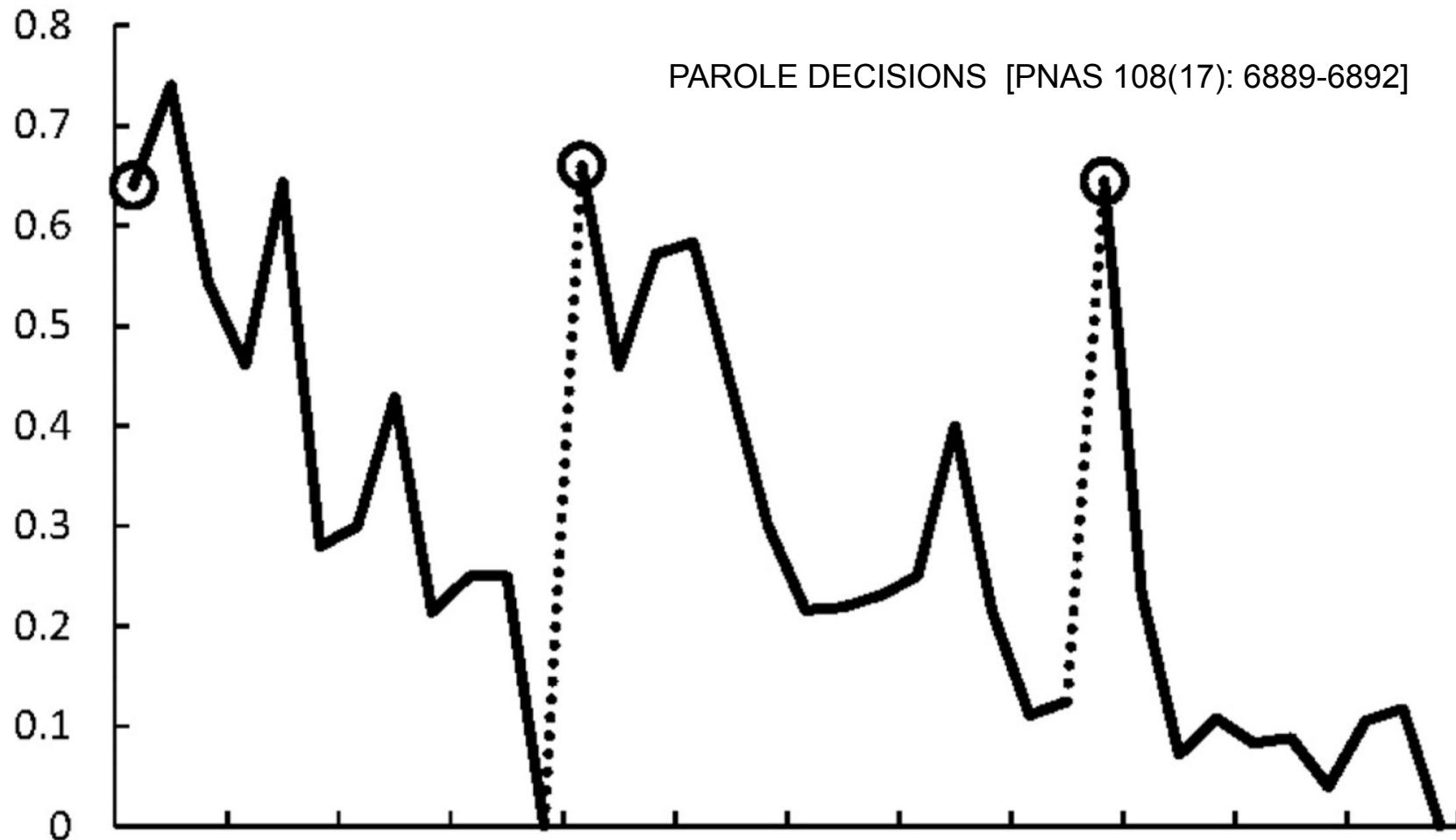


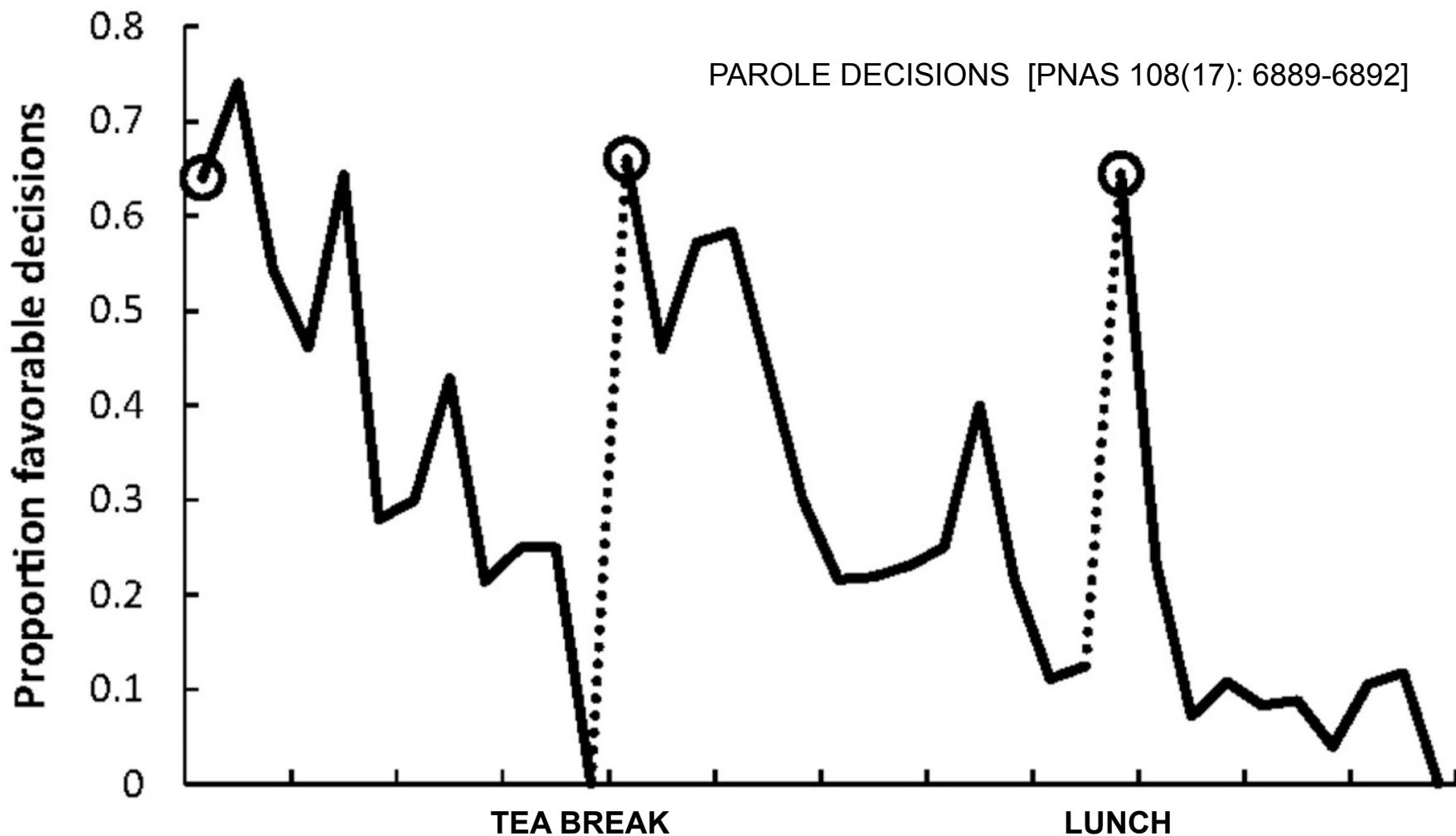


A wooden gavel with a gold band resting on a laptop keyboard.

**COMPAS tool**

Proportion favorable decisions





A wooden gavel with a gold band resting on a laptop keyboard.

**COMPAS tool**

**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?

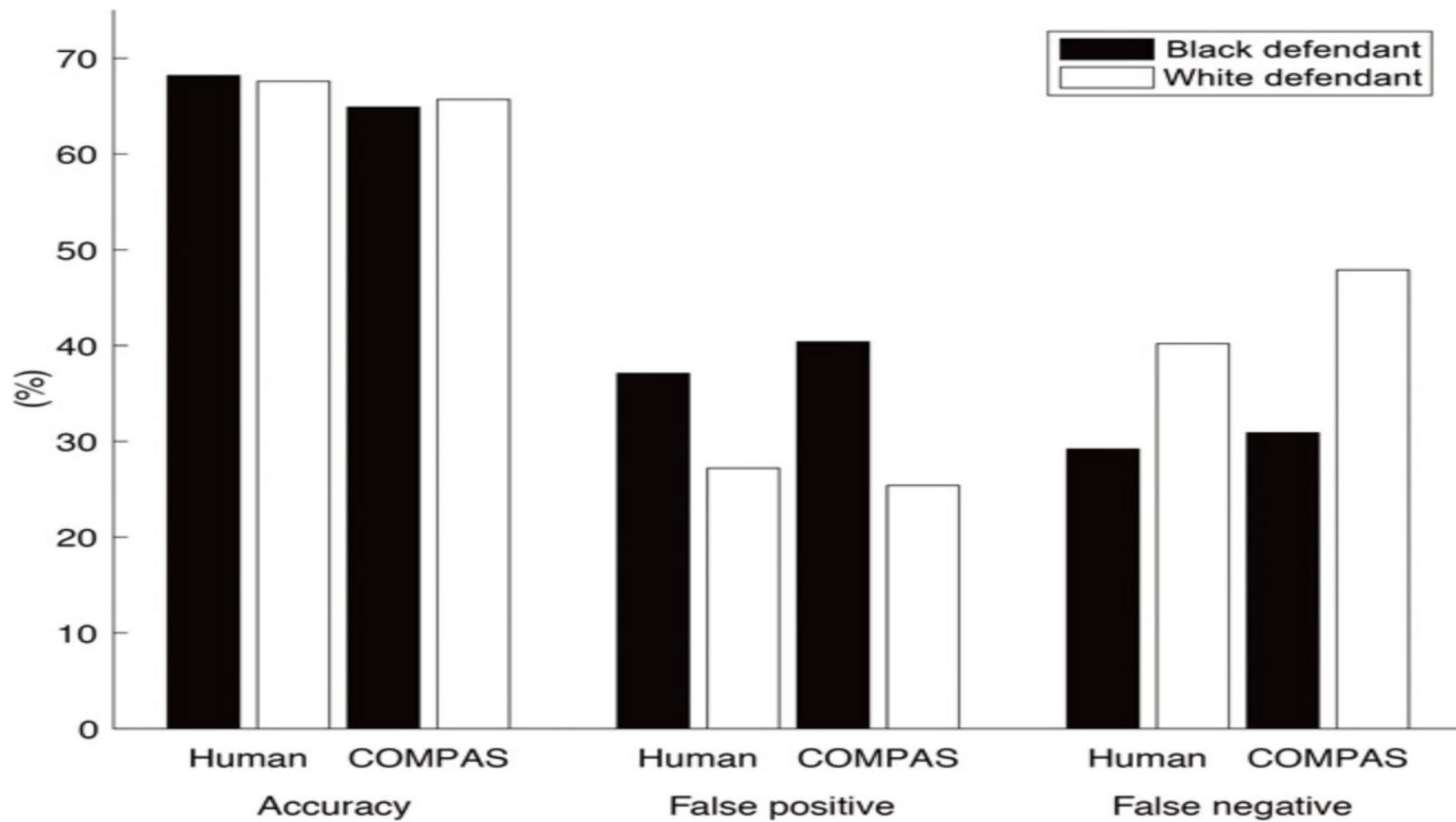
- Strongly Disagree  Disagree  Not Sure  Agree  Strongly Agree

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	<b>True Negative</b>	<b>False Negative</b>
Predicted guilty	<b>False Positive</b>	<b>True Positive</b>

# 21 definitions of “fair”

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

FP/ (TN+FP) identical

	Not Guilty	Guilty
Predicted not guilty	<b>True Negative</b>	False Negative
Predicted guilty	<b>False Positive</b>	True Positive

*Percentage not guilty who are  
incorrectly predicted guilty*

# 21 definitions of “fair”

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

FP/ (TN+FP) identical

	Not Guilty	Guilty
Predicted not guilty	<b>True Negative</b>	False Negative
Predicted guilty	<b>False Positive</b>	True Positive

*Percentage not guilty who are incorrectly predicted guilty*

*ProPublica’s complaint about COMPAS*

# 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	<b>False Positive</b>	<b>True Positive</b>

# 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	<b>False Positive</b>	<b>True Positive</b>

# 21 definitions of “fair”

For all groups, equal opportunity

FN/ (TP+FN) identical

	Not Guilty	Guilty
Predicted not guilty	True Negative	<b>False Negative</b>
Predicted guilty	False Positive	<b>True Positive</b>

*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	<b>False Negative</b>
Predicted guilty	<b>False Positive</b>	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	<b>True Negative</b>	<b>False Negative</b>
Predicted guilty	<b>False Positive</b>	<b>True Positive</b>

*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

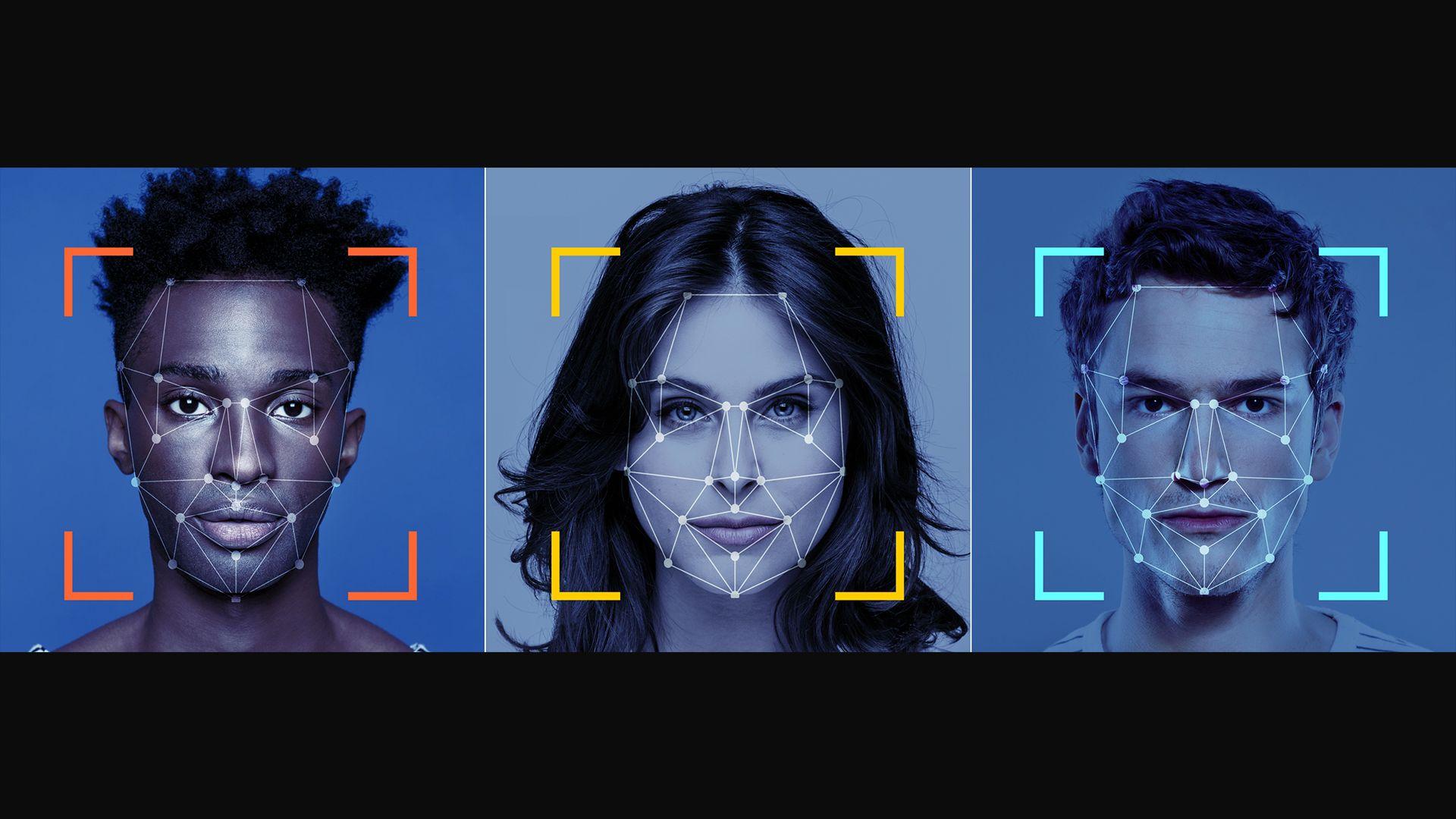


# Insurance

The image displays two side-by-side screenshots of the DriveSure mobile application on an iPhone.

**Left Screenshot:** The screen shows the "Stationary" section of the app. At the top, it says "54 Journeys" and "323 km". Below this is a large circular progress bar with a green segment indicating a "Score" of 75. The bar has arrows on either side and a series of small dots below it. Below the bar is a "Score history" chart showing daily scores from November 16 to 22. The scores fluctuate between approximately 60 and 80. At the bottom are three navigation icons: a house, a location pin, and three dots.

**Right Screenshot:** The screen shows the "Journeys" section of the app. At the top, it says "17 November" (Saturday). Below this is a table listing journeys with columns for DATE, DURATION, DISTANCE, and SCORE. The journeys are listed with their start and end times, duration in hours and minutes, distance in kilometers, and a score. The table continues for other dates, including "16 November" (Friday) and "08:15" (not fully visible). At the bottom are three navigation icons: a house, a location pin, and three dots.



# 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 ▾



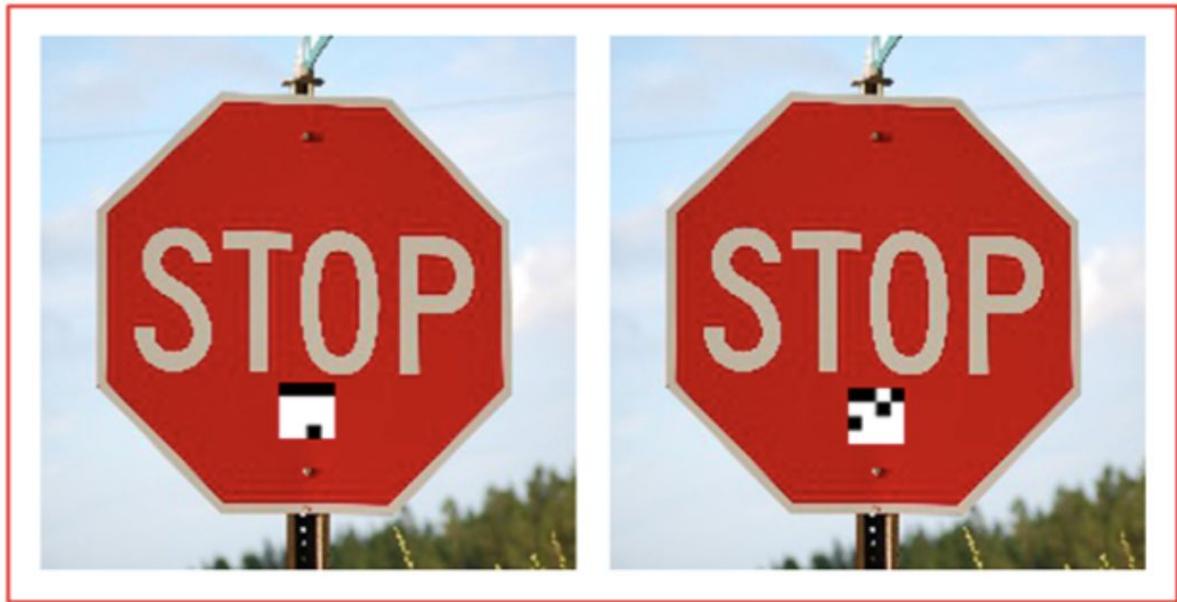
Granny Smith	85.6%
iPod	0.4%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.1%



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



Stop



Yield

Speed Limit

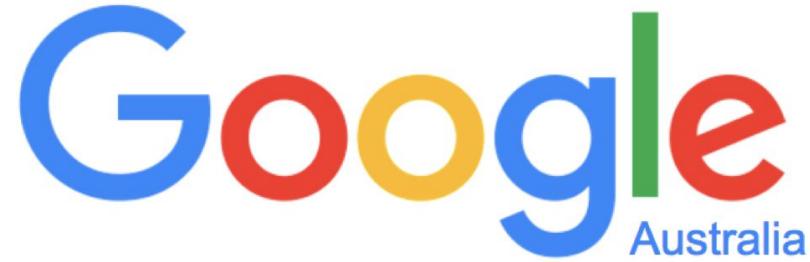


professors are |



- 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**

Report inappropriate predictions



climate change is |



climate change is **not real**

climate change is **a hoax**

climate change is **real**

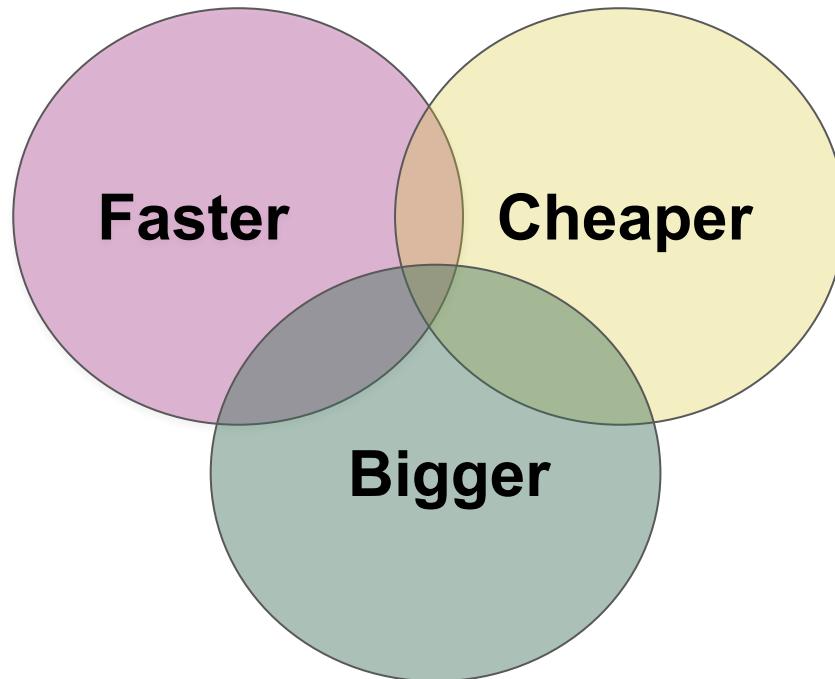
climate change is **a myth**

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

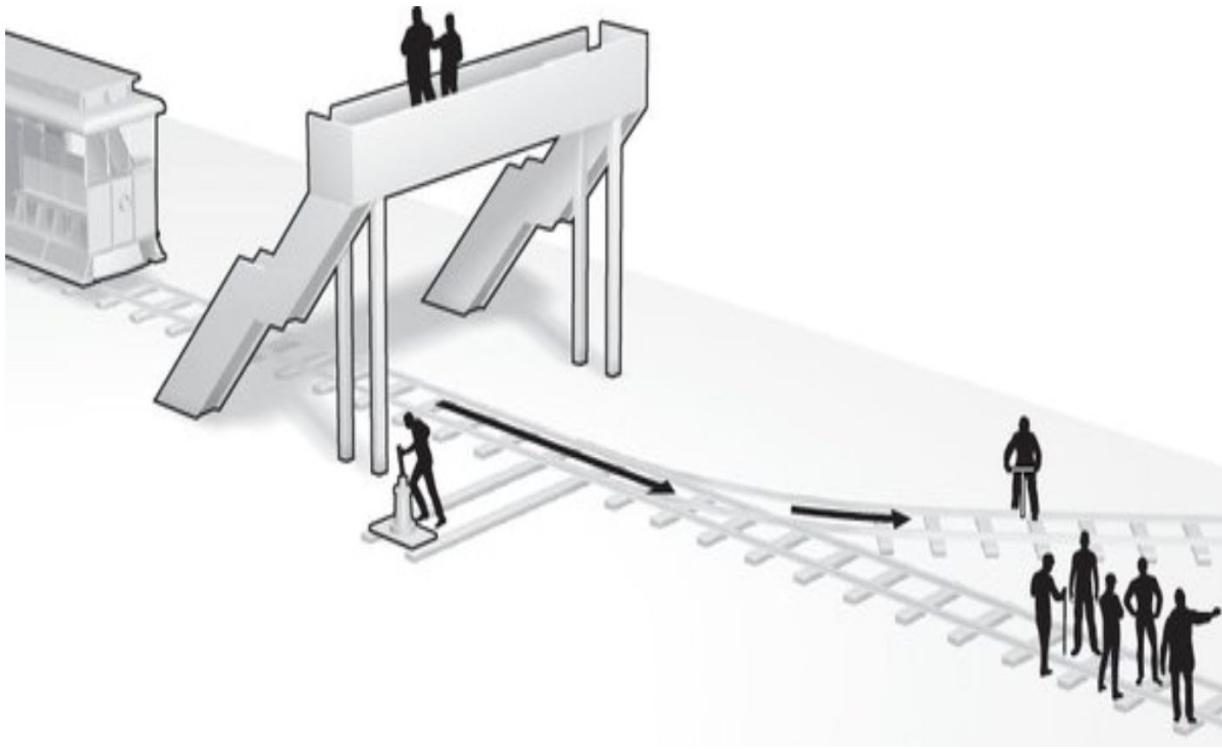
I'm Feeling Lucky

*Report inappropriate predictions*

# Only one new ethical challenge!







# Only one new ethical challenge!



It is not  
Terminator!





home > tech

## 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

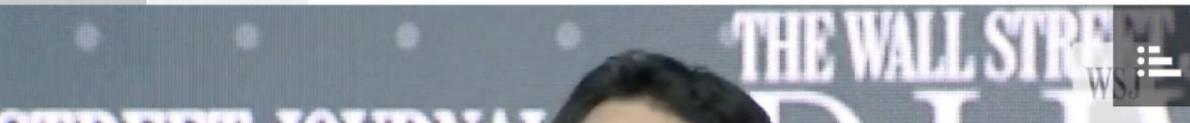
innovation → design

# 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](#) news  
com.au

AUGUST 21, 2017 10:06AM

[Video](#) [Image](#)

# There is an arms race





Taranis in the air



Sea Hunter on the water



MRK-25 on land



# Echo Voyager under the sea

The background of the image shows a dramatic sunset or sunrise sky with orange and blue hues. Six black silhouettes of quadcopter drones are scattered across the sky. One drone is positioned at the top left, another at the bottom left, one in the center, one on the right side, and two more towards the bottom right. Each drone has four propellers and a central body with what appears to be a camera or sensor unit.

They will be  
weapons of terror

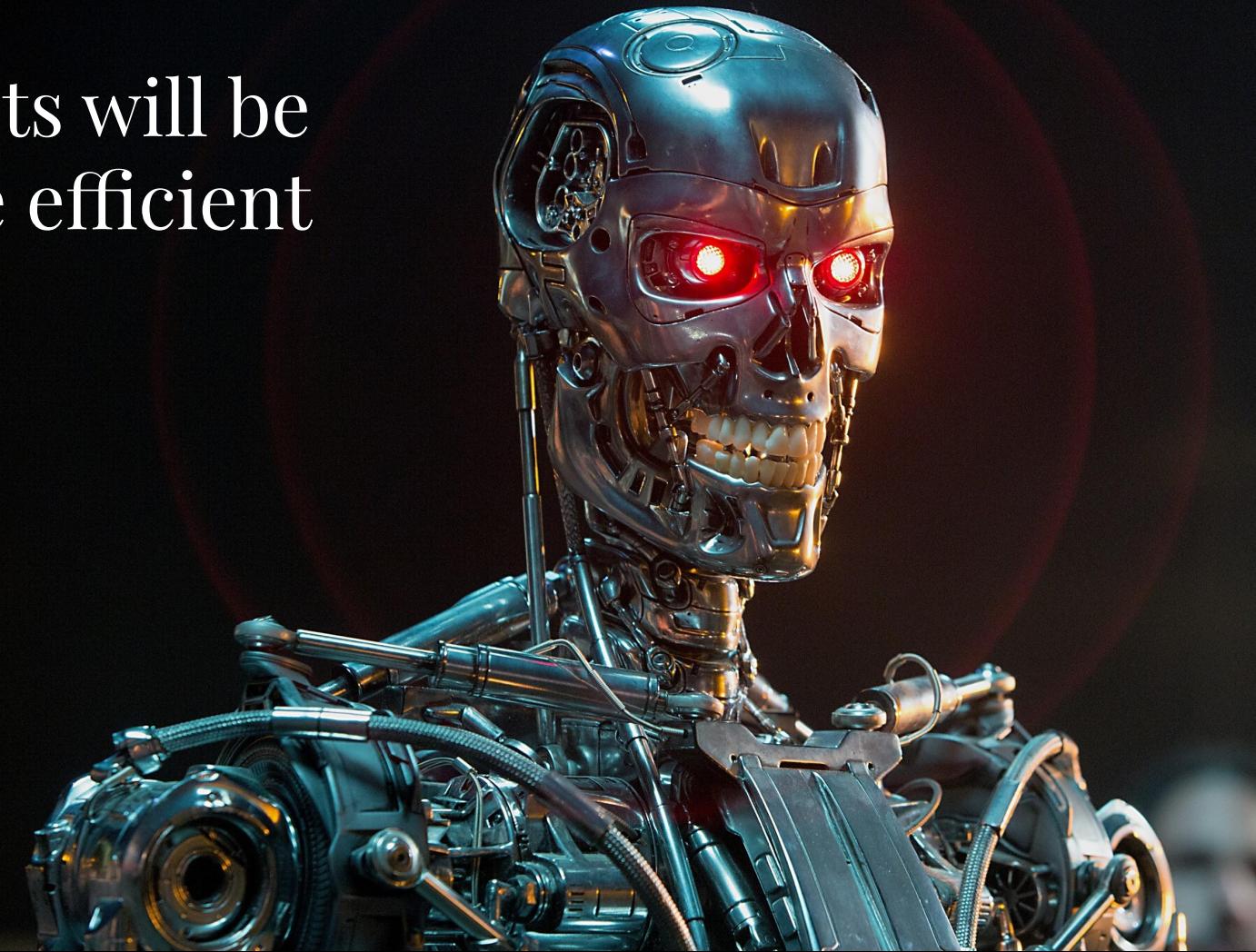
The background of the image shows a dramatic sunset or sunrise sky with orange and blue hues. Six black silhouettes of quadcopter drones are scattered across the sky, some in the foreground and others higher up. They are shown from a three-quarter perspective, facing towards the left of the frame. Each drone has four propellers and a central body with what appears to be a camera or sensor array.

They will be  
weapons of mass  
destruction

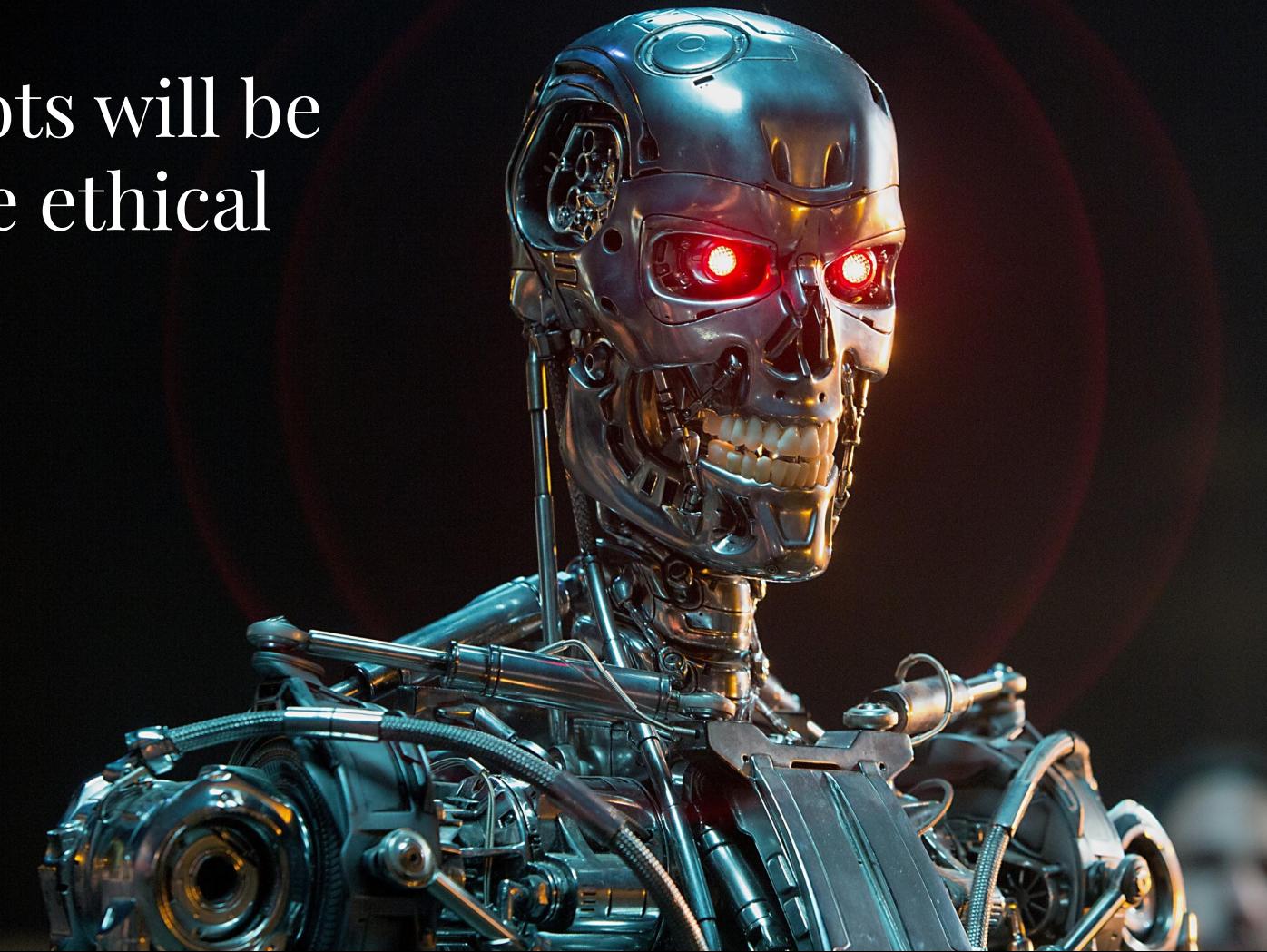
# They will destabilize world



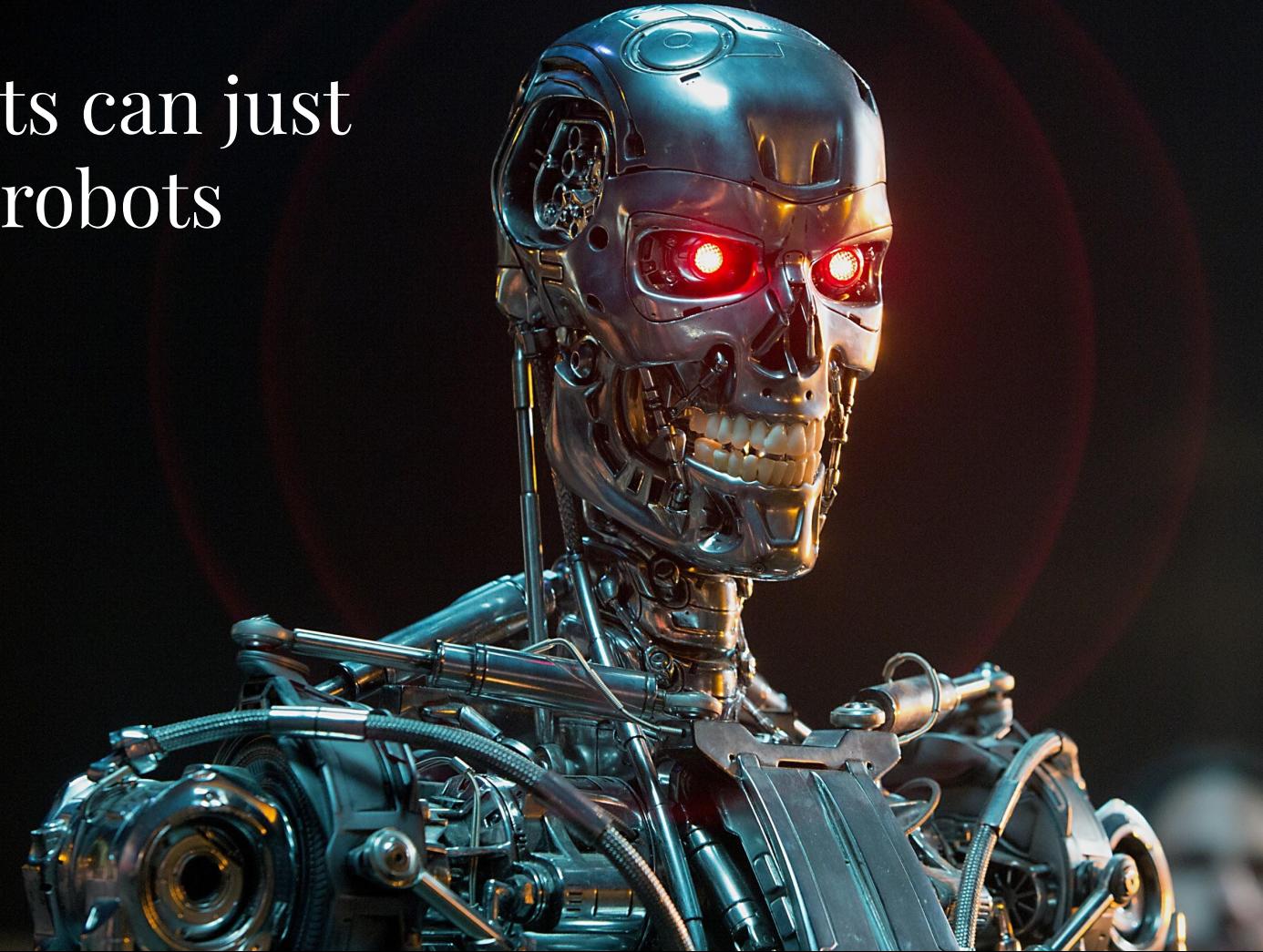
Robots will be  
more efficient



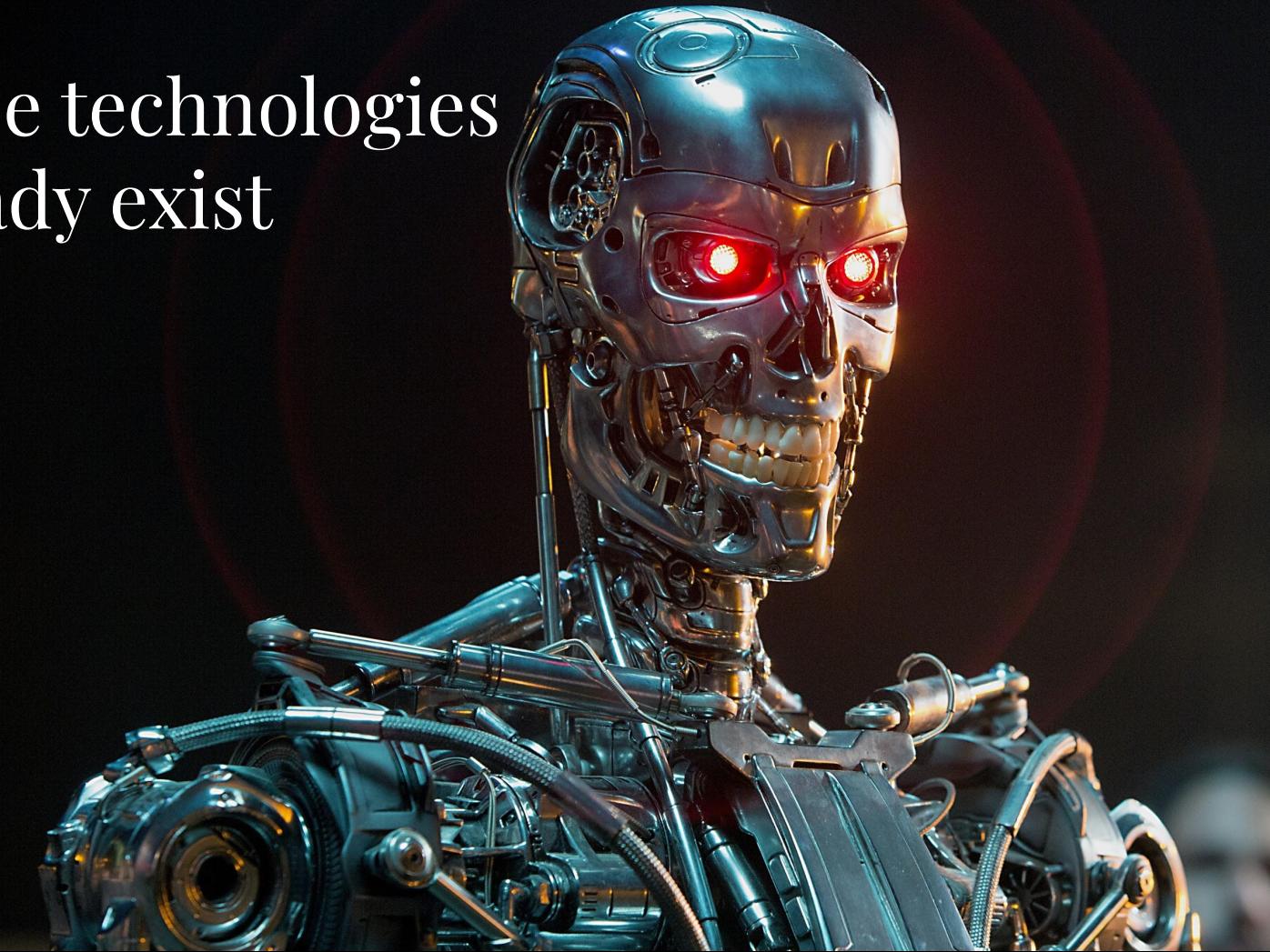
Robots will be  
more ethical



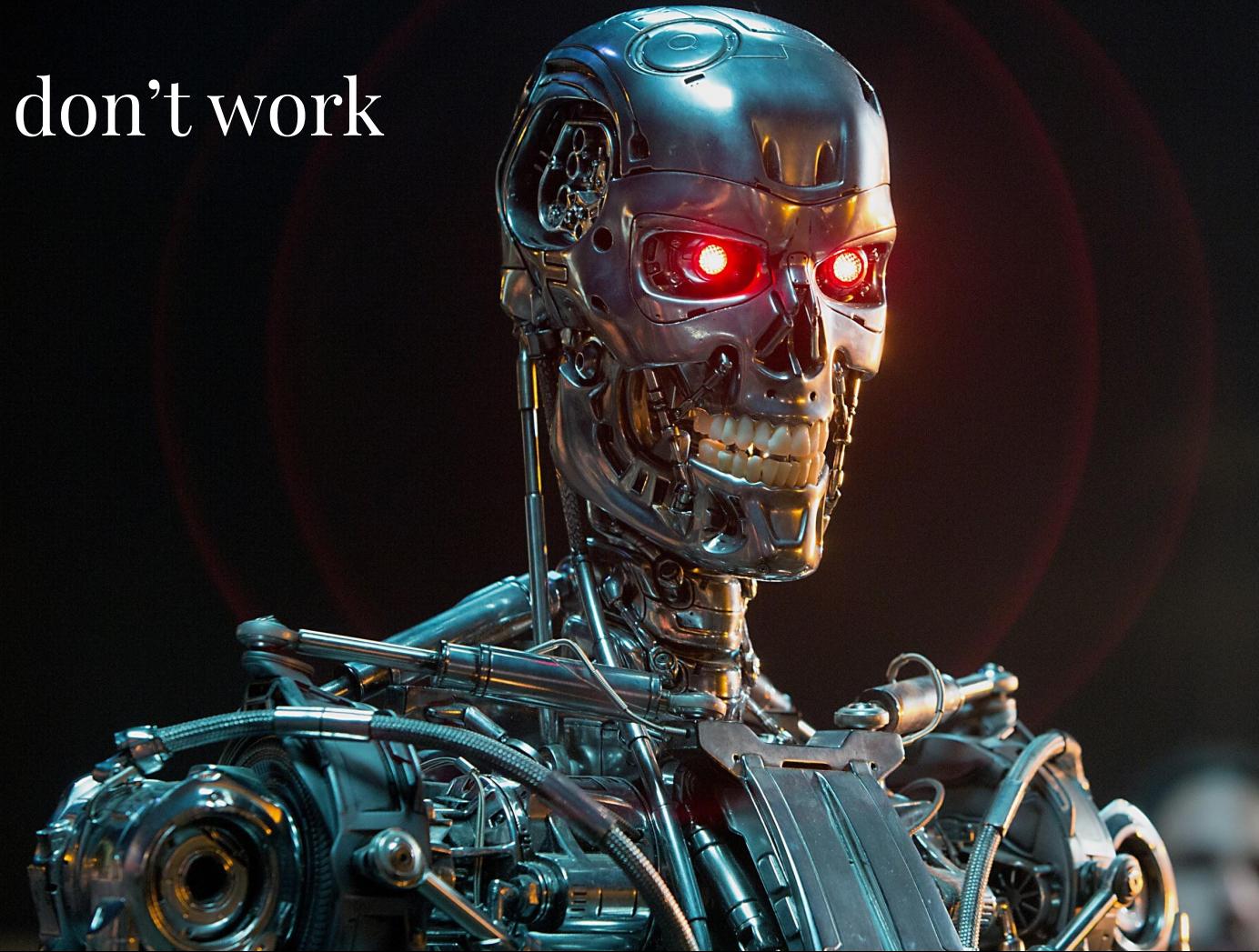
Robots can just  
fight robots



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

