AI: The Story So Far

Stuart Russell

University of California, Berkeley

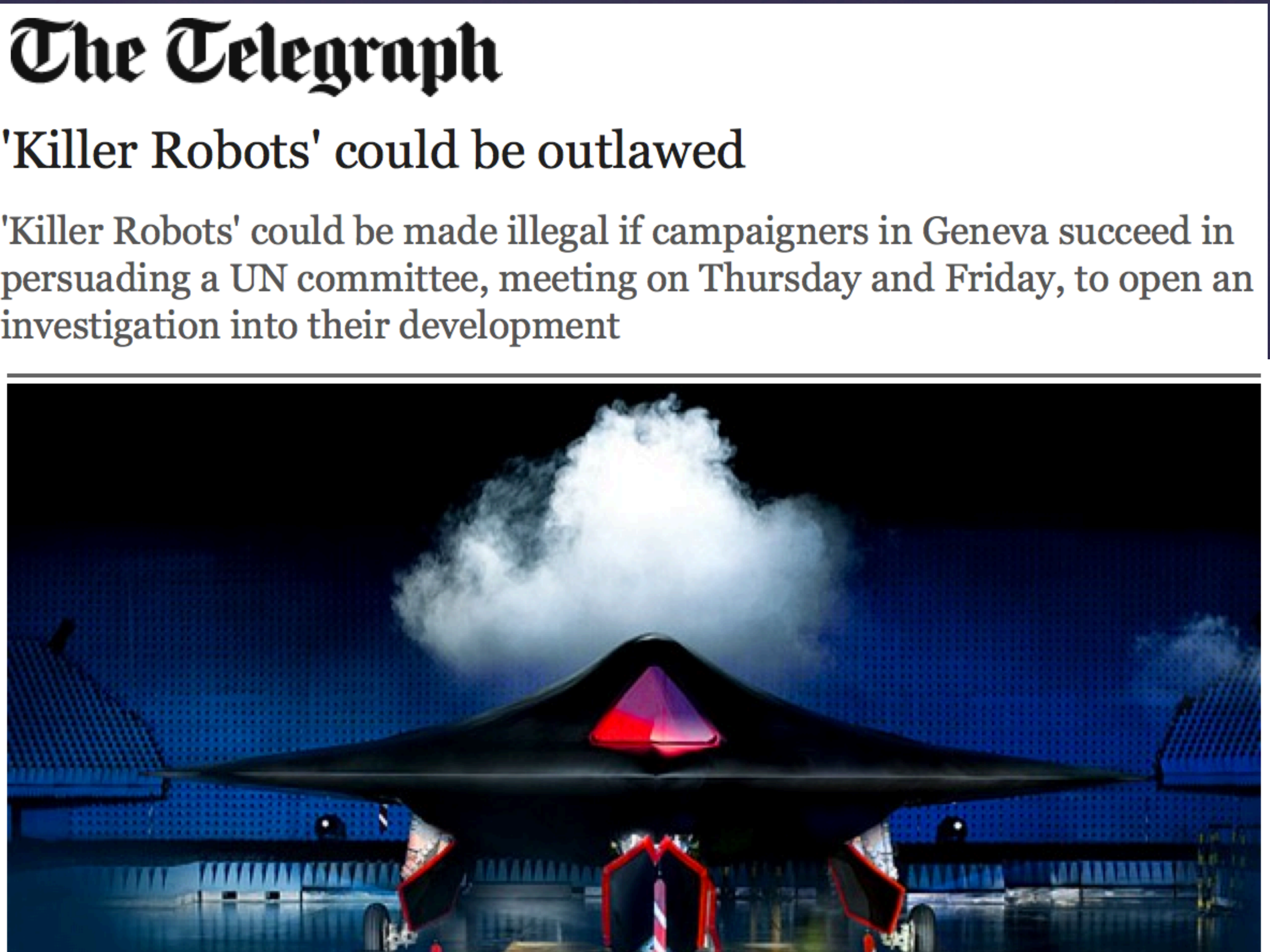
Premise

- Eventually, AI systems will make better* decisions than humans
    - Taking into account more information, looking further into the future

Upside

- Everything we have is the product of intelligence
- Success in AI might mean…
  - An end to war, disease, poverty, ecological degradation
  - The ability to shape our own destiny in the universe

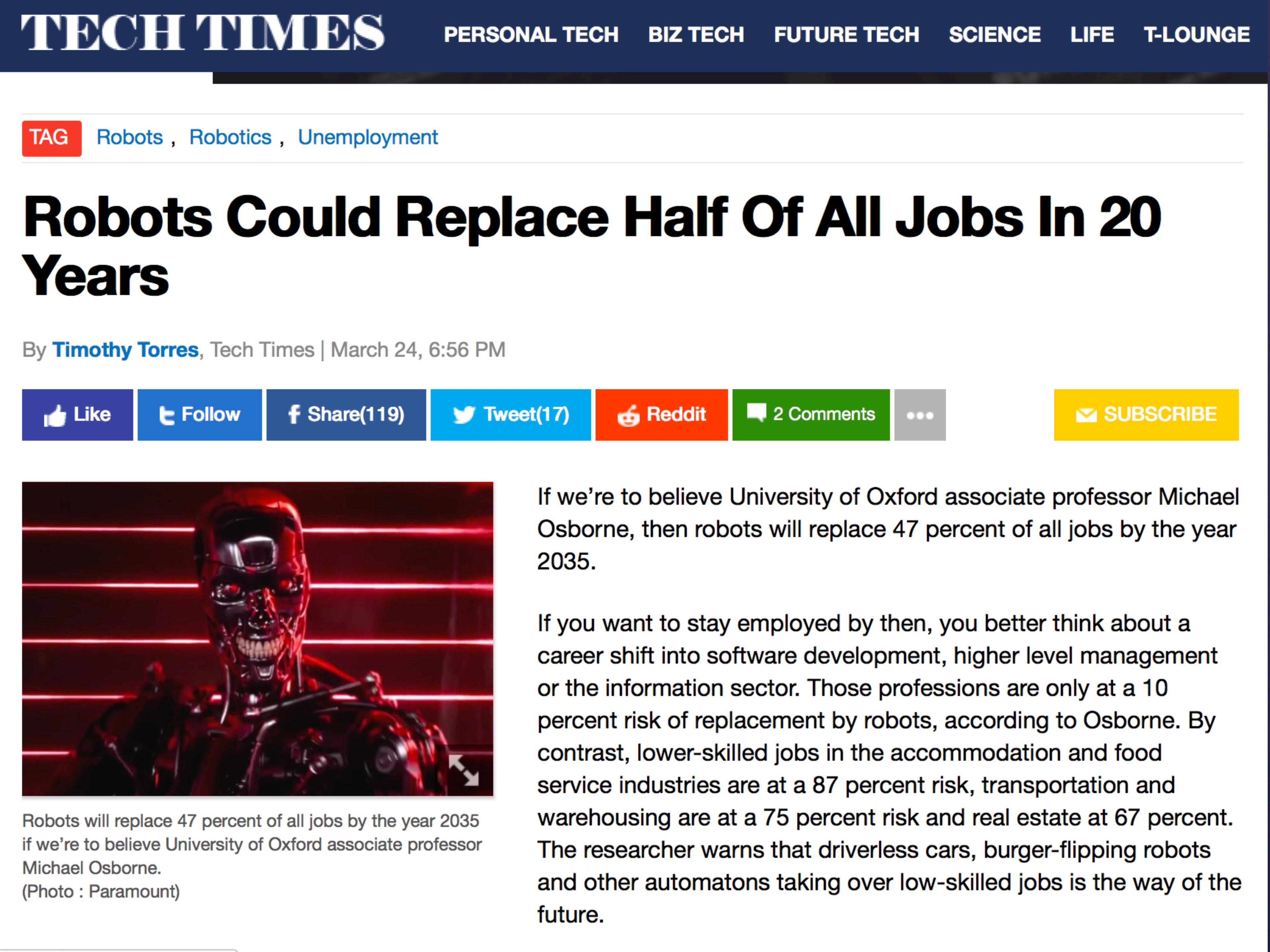
Downside



The Telegraph

'Killer Robots' could be outlawed

'Killer Robots' could be made illegal if campaigners in Geneva succeed in persuading a UN committee, meeting on Thursday and Friday, to open an investigation into their development



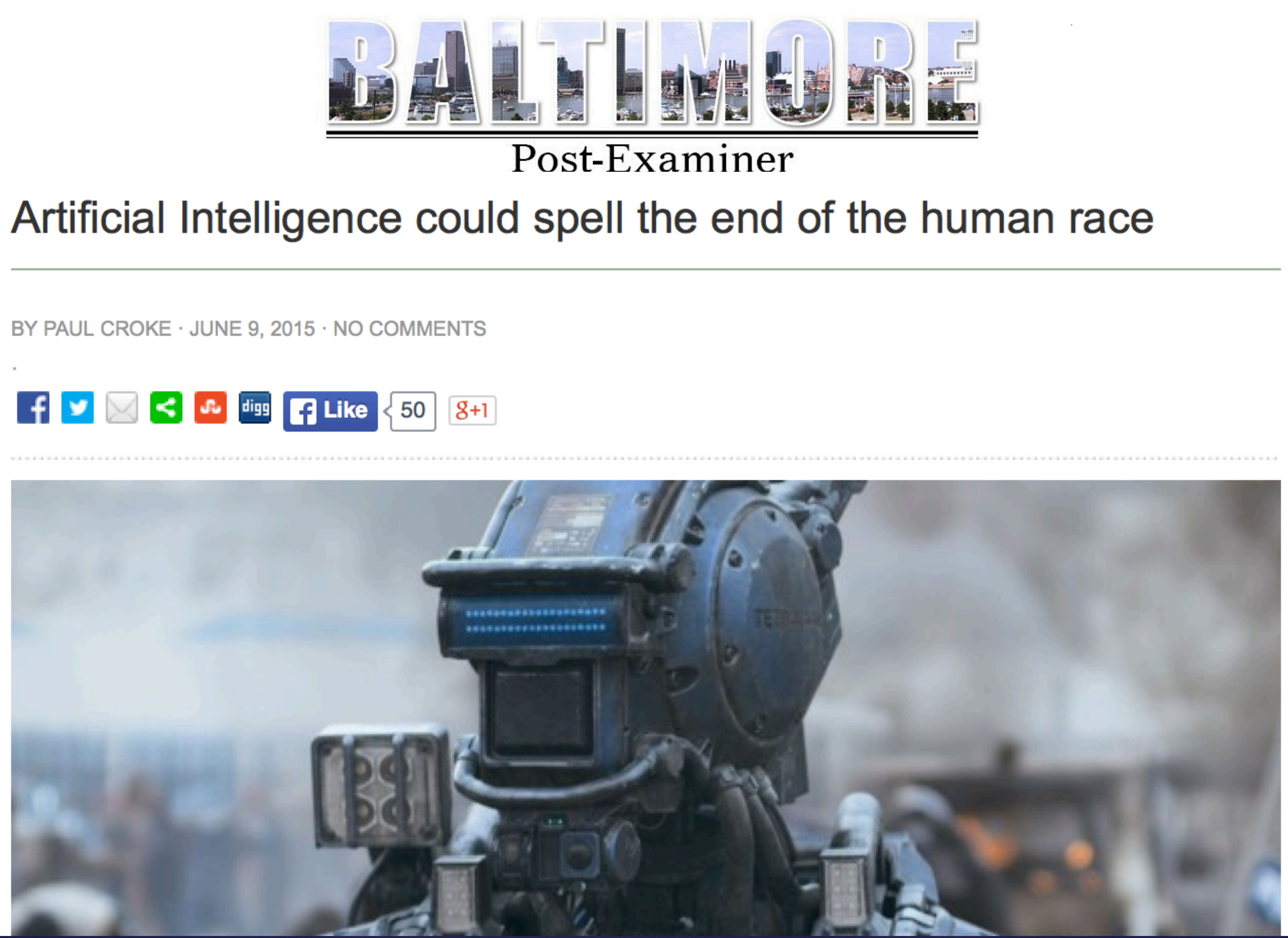
Robots Could Replace Half Of All Jobs In 20 Years

By **Timothy Torres**, Tech Times | March 24, 6:56 PM

Robots will replace 47 percent of all jobs by the year 2035 if we're to believe University of Oxford associate professor Michael Osborne.
(Photo : Paramount)

If we're to believe University of Oxford associate professor Michael Osborne, then robots will replace 47 percent of all jobs by the year 2035.

If you want to stay employed by then, you better think about a career shift into software development, higher level management or the information sector. Those professions are only at a 10 percent risk of replacement by robots, according to Osborne. By contrast, lower-skilled jobs in the accommodation and food service industries are at a 87 percent risk, transportation and warehousing are at a 75 percent risk and real estate at 67 percent. The researcher warns that driverless cars, burger-flipping robots and other automatons taking over low-skilled jobs is the way of the future.



BALTIMORE

Post-Examiner

Artificial Intelligence could spell the end of the human race

BY PAUL CROKE · JUNE 9, 2015 · NO COMMENTS

What's bad about better AI?

- Obviously, smarter-than-human AI systems are intrinsically hard to predict and control
  - Are gorillas glad they created humans?
- In particular, AI systems that are incredibly good at achieving something other than what we really want
- AI, economics, statistics, operations research, control theory all assume utility to be ***exogenously specified***

Value misalignment

If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively … we had better be quite sure that the purpose put into the machine is the purpose which we really desire

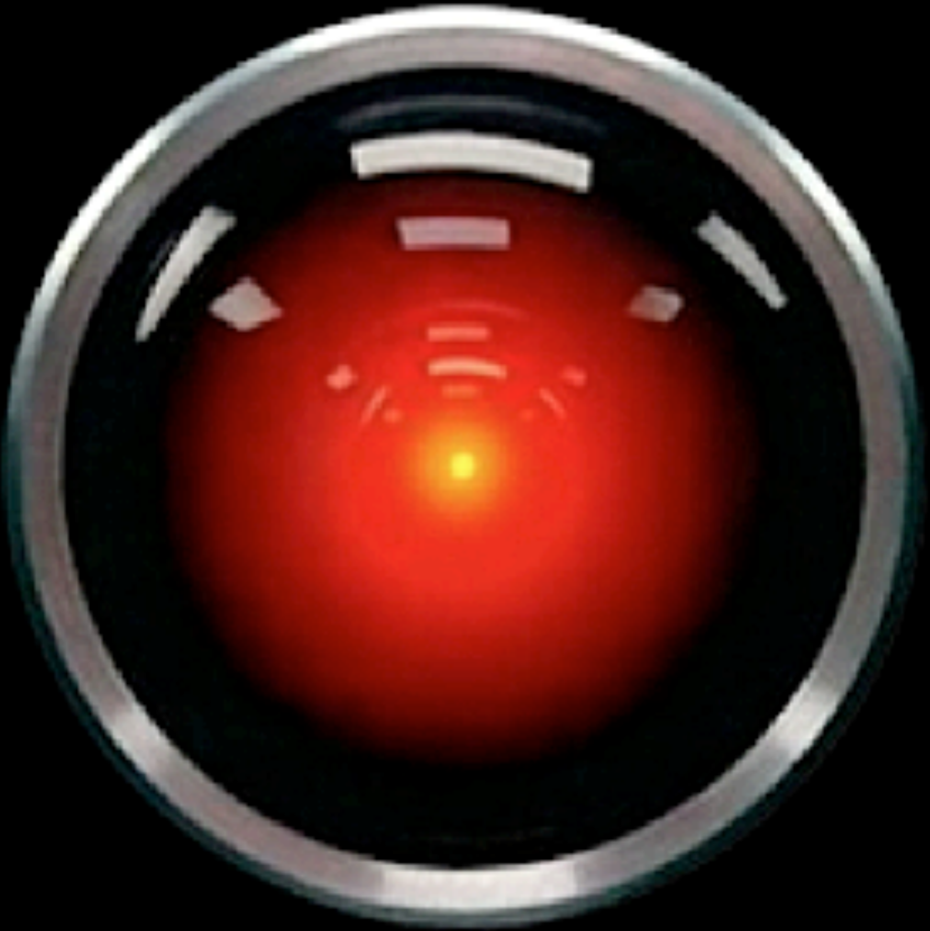
Norbert Wiener, 1960

King Midas, c540 BCE

Instrumental goals

- For ***any primary goal***, the odds of success are improved by
  1) Maintaining one's own existence
  2) Acquiring more resources

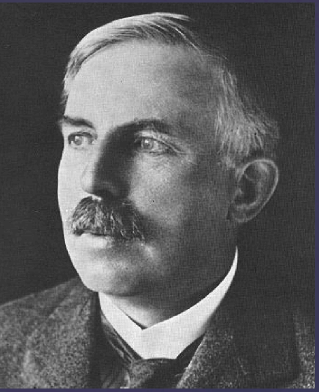
- With value misalignment, these lead to obvious problems for humanity



I’m sorry, Dave, I’m afraid I can’t do that

Reasons not to pay attention:

- It’ll never happen



Sept 11, 1933: Lord Rutherford addressed BAAS: *"Anyone who looks for a source of power in the transformation of the atoms is talking moonshine."*

Sept 12, 1933: Leo Szilard invented neutron-induced nuclear chain reaction

"We switched everything off and went home. That night, there was very little doubt in my mind that the world was headed for grief."

Reasons not to pay attention:

- It’ll never happen
  - See Rutherford, 9/11/33, Szilard 9/12/33
- It’s too soon to worry about it
  - 2066 asteroid collision: when exactly do we worry?
  - When should we have worried about climate change?
- It’s like worrying about overpopulation on Mars
  - No, it’s as if we were spending billions moving humanity to Mars with no plan for what to breathe
- Just don’t have explicit goals for the AI system
  - We need to steer straight, not remove the steering wheel
- Don’t worry, we’ll just have human-AI teams
  - Value misalignment precludes teamwork

Reasons not to pay attention:

- You can't control research
  - Yes, we can: we don't genetically engineer humans
- You're just Luddites
  - Fusion researchers are Luddites if they point out the need for containment?
  - Alan Turing, Norbert Wiener, Bill Gates, and Elon Musk are Luddites?
- Don't worry, we can just switch it off
  - As if a superintelligent entity would never think of that
- Don't put in "human" goals like self-preservation
  - Death isn't bad per se. It's just hard to fetch the coffee after you're dead

Proposal

- Not just AI
- Provably* beneficial* AI
- Yes, but how?

Three simple ideas

1. The robot's only objective is to maximize the realization of human values
2. The robot is initially uncertain about what those values are
3. The best source of information about human values is human behavior

The off-switch problem

- A robot, given an objective, has an incentive to disable its own off-switch
- How can we prevent this?
- Answer: robot isn't given an objective!
- Instead, it must allow for ***uncertainty*** about the true human objective
  - The human will only switch off the robot if that leads to better outcomes for the true human objective
  - So it's ***in the robot's interest*** to allow it

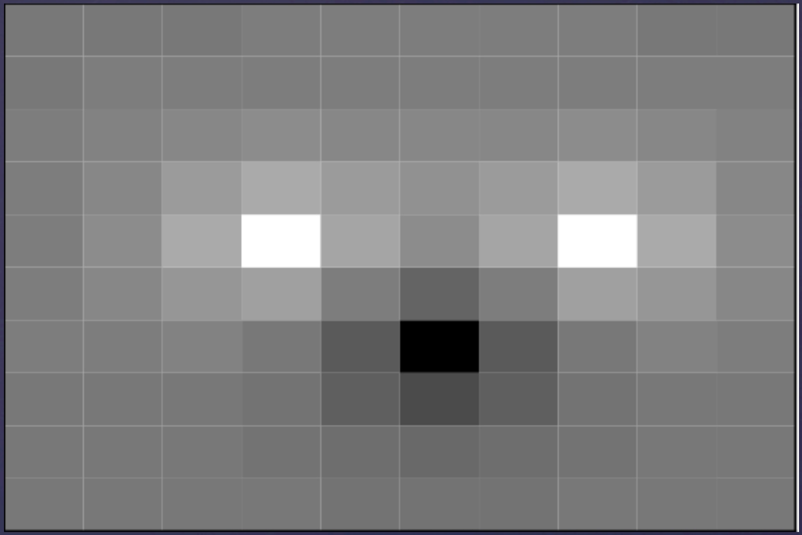
Value alignment

- *Inverse reinforcement learning*: learn a value function by observing another agent's behavior
  - The value function is a succinct explanation for what the other agent is doing
  - Algorithms and theorems already in place: probably approximately aligned learning

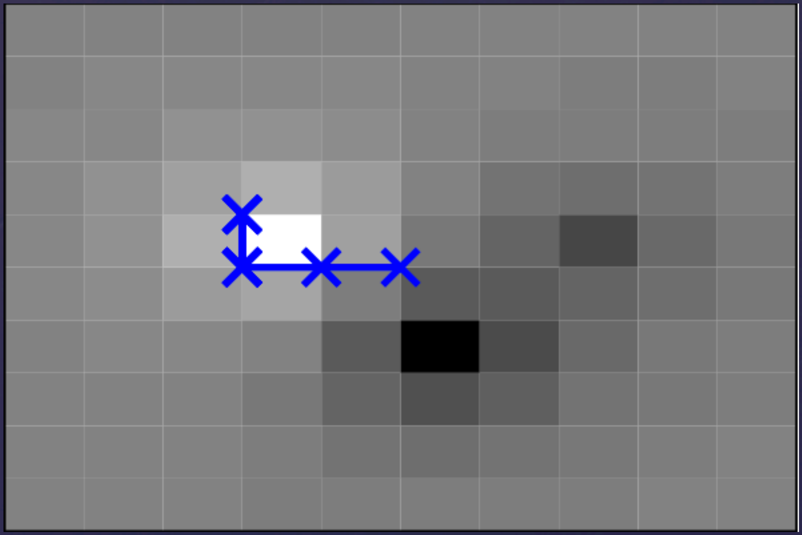
Cooperative inverse reinforcement learning

- A two-player game with “human” and “robot”
  - Human knows the value function
  - Robot doesn’t know it, but wants to maximize it
- Robot has an incentive to ask questions, explore cautiously
- Human has an incentive to teach the robot rather than just displaying “optimal” behavior

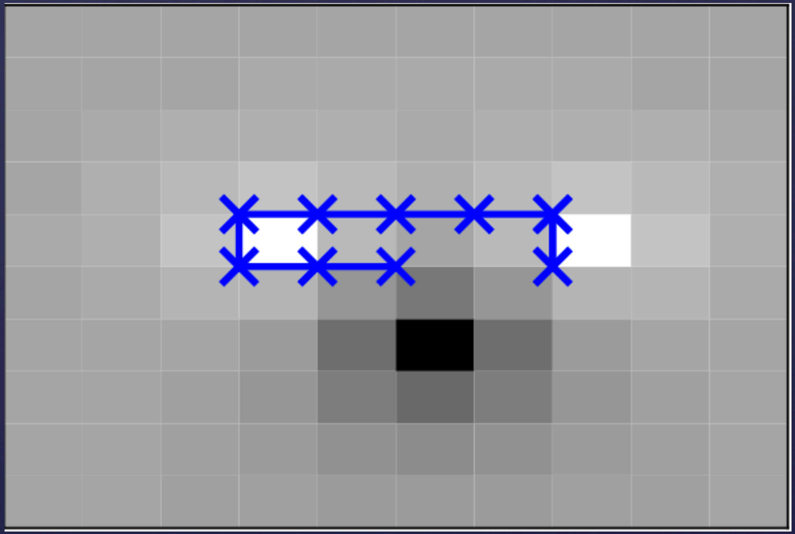
Example: IRL vs CIRL



TRUE REWARD



“OPTIMAL” BEHAVIOR,
INFERRED REWARD

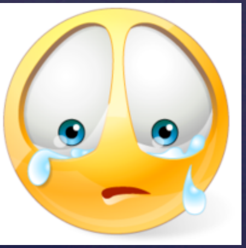


CIRL SOLUTION,
INFERRED REWARD

Value alignment contd.

- Vast amounts of evidence for human behavior and *human attitudes towards that behavior*

- We need value alignment even for *subintelligent* systems in human environments; strong economic incentives!



- Humans are nasty, irrational, inconsistent, weak-willed, computationally limited, and heterogeneous

Questions

- Can we change the way AI defines itself?
- How will the process draw from and contribute to our understanding of ethical issues?
- Will it make us better people?
- What exactly does Values 'R Us sell?

Wiener, contd.

> [this work] requires an imaginative forward glance at history which is difficult, exacting, and only partially achievable. …
>
> **We must always exert the full strength of our imagination to examine where the full use of our new modalities may lead us**