Clouds...

Safety: Case

Is the Tesla safe?

What is the argument?
How measured, for whom?

*(*Is psychological tension part of safety, what are the limits of what we call "safe"?)*

A Tragic Loss

The Tesla Team •
June 30, 2016

We learned yesterday evening that NHTSA is opening a preliminary evaluation into the performance of Autopilot during a recent fatal crash that occurred in a Model S. This is the first known fatality in just over 130 million miles where Autopilot was activated. Among all vehicles in the US, there is a fatality every 94 million miles. Worldwide, there is a fatality approximately every 60 million miles. It is important to emphasize that the NHTSA action is simply a preliminary evaluation to determine whether the system worked according to expectations.

Safety: Case

How can self-driving cars be *compared* in terms of safety?

What should be measured?

*(*Is there an algorithm to measure...?)*

Safety case & thought experiment

What is the *value* of safety?

How can it be described,
or measured?

Definition safety

1. ChatGPT

2. Practical

Here is a symbolic representation of the principle of safety in AI ethics.
This artwork visually encapsulates the core concepts, emphasizing the importance of protection, balance, and careful development in the realm of AI safety
(ChatGPT 2024)

Definition safety

Practical

AI safeguarded against causing harms,
even when confronted with unexpected changes,
anomalies, and perturbations.

Robust and resilient

Tesla that functions well in San Francisco
also does well in Trento

Tesla stops for the badly discolored stop sign

Outline

Accidents

What is safety?
  Robust and resilient AI

2 Safety threats

Dynamics for measuring safety

Safety strategies

Philosophies of innovation and safety

2 AI safety threats: Underperform, overperform

Not work well enough, work too well

Less than perfect, *more* than perfect

2 AI safety threats: Underperform

2 Safety threats: Overperform, *Paperclips thought experiment*

Ethical Issues in Advanced Artificial Intelligence

Nick Bostrom

Oxford University
Philosophy Faculty
10 Merton Street
Oxford OX1 4JJ
United Kingdom

[This is a slightly revised version of a paper

a superintelligence whose top goal is the manufacturing of paperclips, with the consequence that it starts transforming first all of earth and then increasing portions of space into paperclip manufacturing facilities.

esearch and

s a cognitive pursuit, a superintelligence could also easily surpass humans in the quality of
e the superintelligence may become unstoppably powerful because of its intellectual
er surveys some of the unique ethical issues in creating superintelligence, and discusses
what motivations we ought to give a superintelligence, and introduces some cost-benefit considerations relating to whether the development of superintelligent machines ought to be accelerated or retarded.

KEYWORDS: Artificial intelligence, ethics, uploading, superintelligence, global security, cost-benefit analysis

1. INTRODUCTION

A *superintelligence* is any intellect that is vastly outperforms the best human brains in practically every field, including scientific creativity, general wisdom, and social skills.[1] This definition leaves open how the superintelligence is implemented – it could be in a digital computer, an ensemble of networked computers, cultured cortical tissue, or something else.

On t
corp
suffi

We need to be careful about what we wish for from a superintelligence, because we might get it.

Several authors have argued that there is a substantial chance that superintelligence may be created within a few decades, perhaps as a result of growing hardware performance and

2 Underperform, overperform (*Philosophy)

Fundamentally, is reality *less* than perfect/complete, or *more* than perfect/complete?

Outline

Accidents

What is safety?
  Robust and resilient AI

2 Safety threats

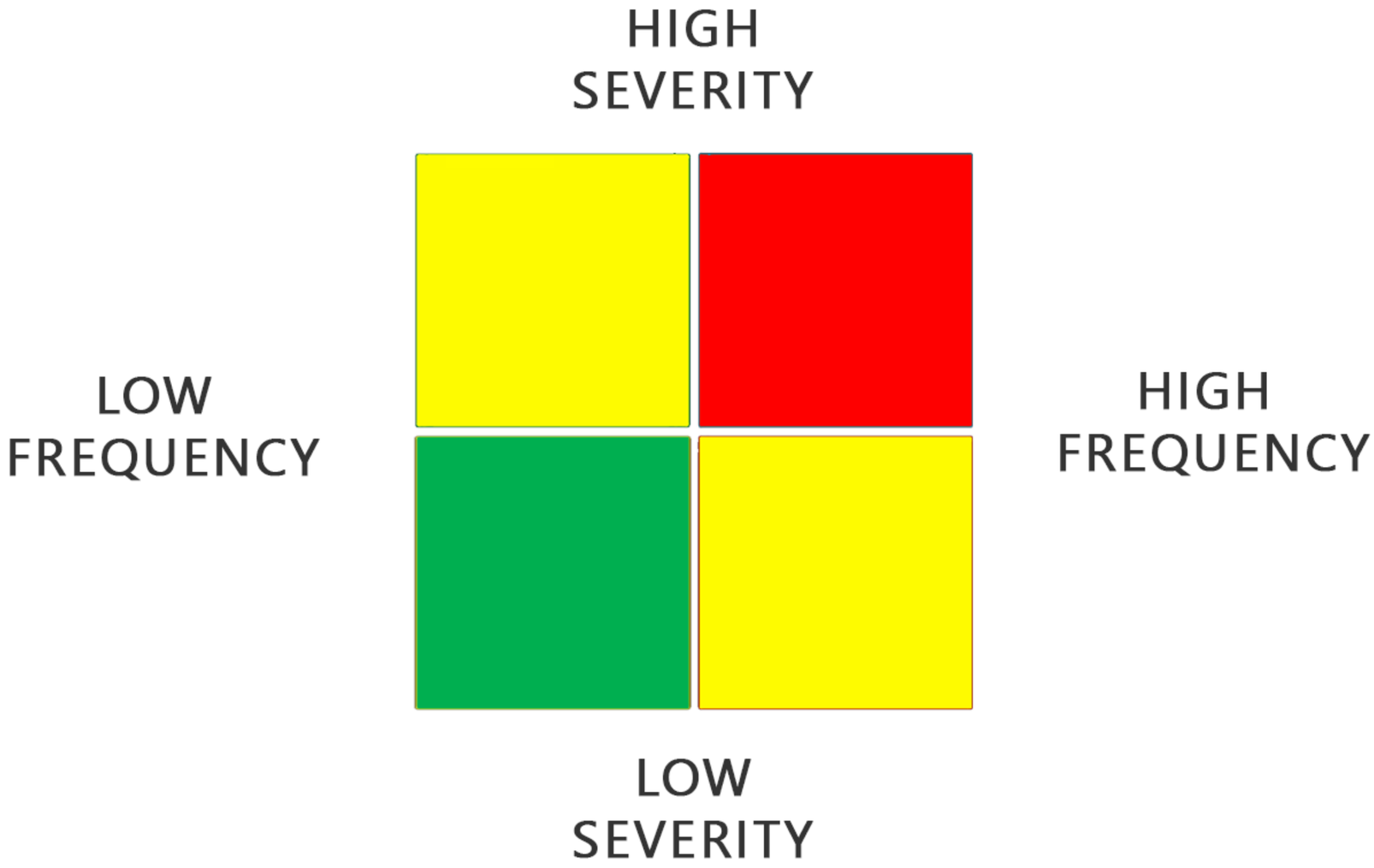
Dynamics for measuring safety

Safety strategies

Philosophies of innovation and safety

Dynamics for measuring safety: Mapping frequency and severity

Netflix, Tesla, AI hypersonic weapons?



Dynamics for measuring safety: Mapping frequency and severity

Frequency

Normalized as time (ie miles driven per incident converted into minutes driving per incident…)?

Severity

How many bad Netflix recommendations = 1 Tesla autopilot error sprained neck?

How many Tesla autopilot error sprained necks = 1 Tesla autopilot broken leg?

How many Tesla autopilot broken legs = 1 fatality?

Master Algorithm for comparing safety?

NETFLIX

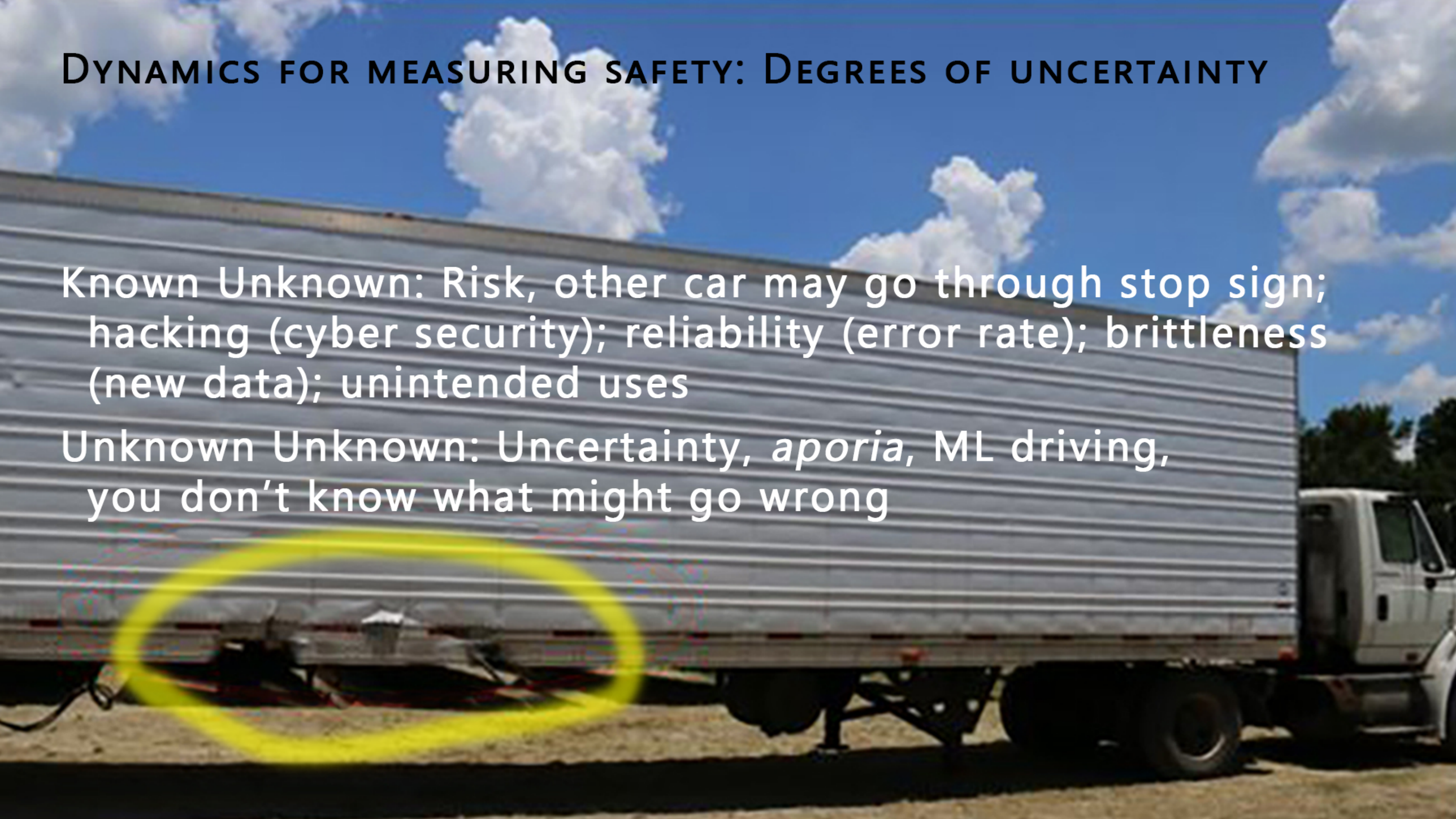
TESLA

Dynamics for measuring safety

What are the dangers of AI?

What units measure them?

Dynamics for measuring safety: Degrees of uncertainty



Known Unknown: Risk, other car may go through stop sign; hacking (cyber security); reliability (error rate); brittleness (new data); unintended uses

Unknown Unknown: Uncertainty, *aporia*, ML driving, you don't know what might go wrong

Dynamics for measuring safety: Degrees of uncertainty

	Type of risk faced = Known	Type of risk faced = Known Unknown	Type of risk faced = Unknown Unknown
AI Chess Player	Opponent may decide to trade Knight for Bishop	Opponent may surprise with a move that forks Knight and Bishop	None, because the game is closed. (A lightening strike is outside the game. If it happens, the game stops.)
AI Car Driver	Other car may try to beat the yellow, enter intersection on red	For some unknown reason (distracted driver, broken stoplight, etc.) other car may enter intersection on red	Lightening strike, understood metaphorically. (Nothing is outside the game.)

Not divide AI in terms of internal functioning, but as kinds of risks it faces.

Helps understand why we can make an AI Chess player that no human could ever beat, but no AI will ever be as good at driving as even a careless teenager.

Dynamics for measuring safety: Degrees of uncertainty

Can degrees of uncertainty about safety be rendered as a formula?

How do you quantify unknown unknown?

Probability + confidence as radically distinct?

Outline

Accidents

What is safety?
  Robust and resilient AI

2 Safety extremes

Dynamics for measuring safety

Safety strategies

Philosophies of innovation and safety

Safety strategies: EC Ethics Guidelines for Trustworthy AI

1.
Fallback systems and human operation in face of irresolvable problems.

- Human in the loop (Person has complete control over starting or stopping any action performed by an intelligent system after receiving a cue)
- Human on the loop (Humans oversight, but the artificial intelligence jumps into action, not needing human pre-approval)

2.
Protections against hacking /
Accounting for unintended uses /
Resilient against exploits

Safety strategies: In/On the loop: Which one is safer...?

The general who oversees U.S. defense against incoming missiles and air-based attacks wants to automate detection systems and move humans further out in the decision-making loop to meet the challenge of faster, more powerful weapons, he said Thursday.

Having a human in the loop means a person still has complete control over starting or stopping any action performed by an intelligent system after receiving a cue. Instead, O'Shaughnessy wants to move the military toward a "human on the loop" standard, pushing human control farther from the center of the automated decision-making. It would still give humans oversight of an automated system, but the artificial intelligence would jump right into action, not needing human pre-approval as it would with a "human in the loop" design.

O'Shaughnessy said that the military needs to move "at the speed of relevance" to react to incoming threats because of the aggressive pursuit of AI by the Chinese and the Russian, and the speed at which they can move.

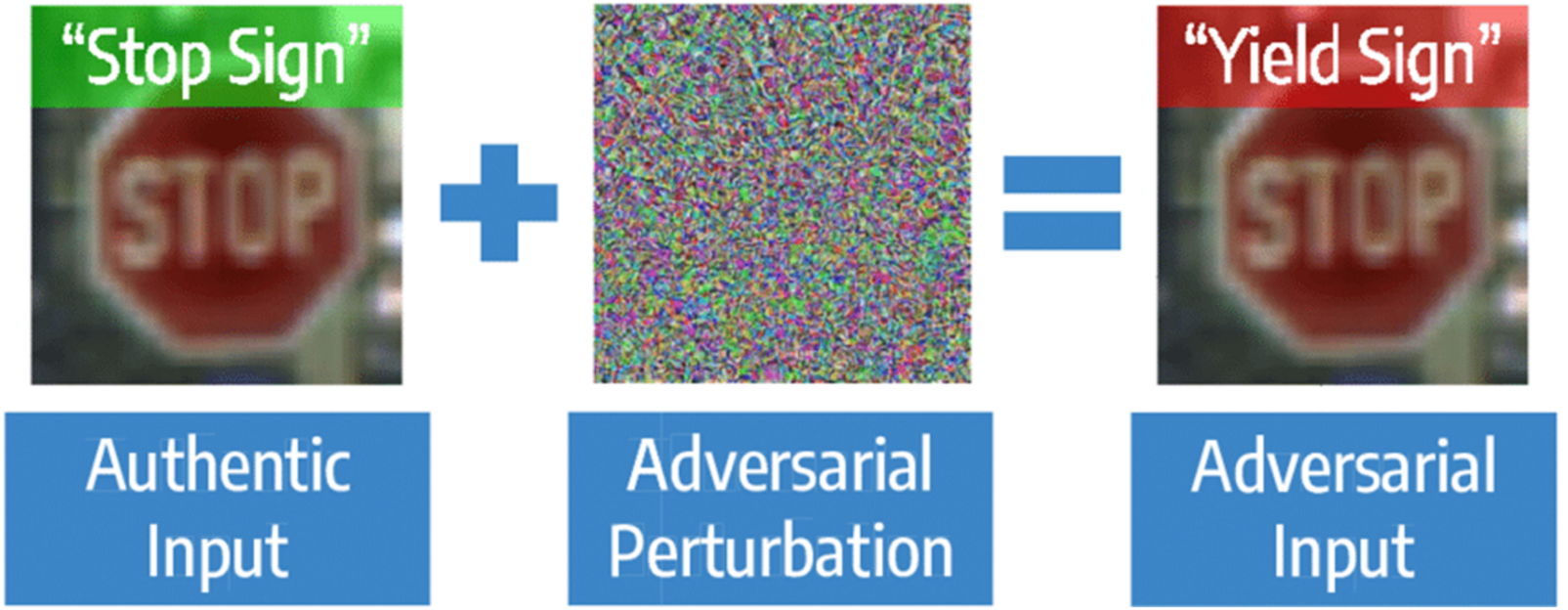
Safety strategies: Resilient against sabotage...?

FUTURE PERFECT TECHNOLOGY

It's disturbingly easy to trick AI into doing something deadly

How "adversarial attacks" can mess with self-driving cars, medicine, and the military.

By Sigal Samuel | Apr 8, 2019, 9:10am EDT



Outline

Accidents

What is safety?
  Robust and resilient AI

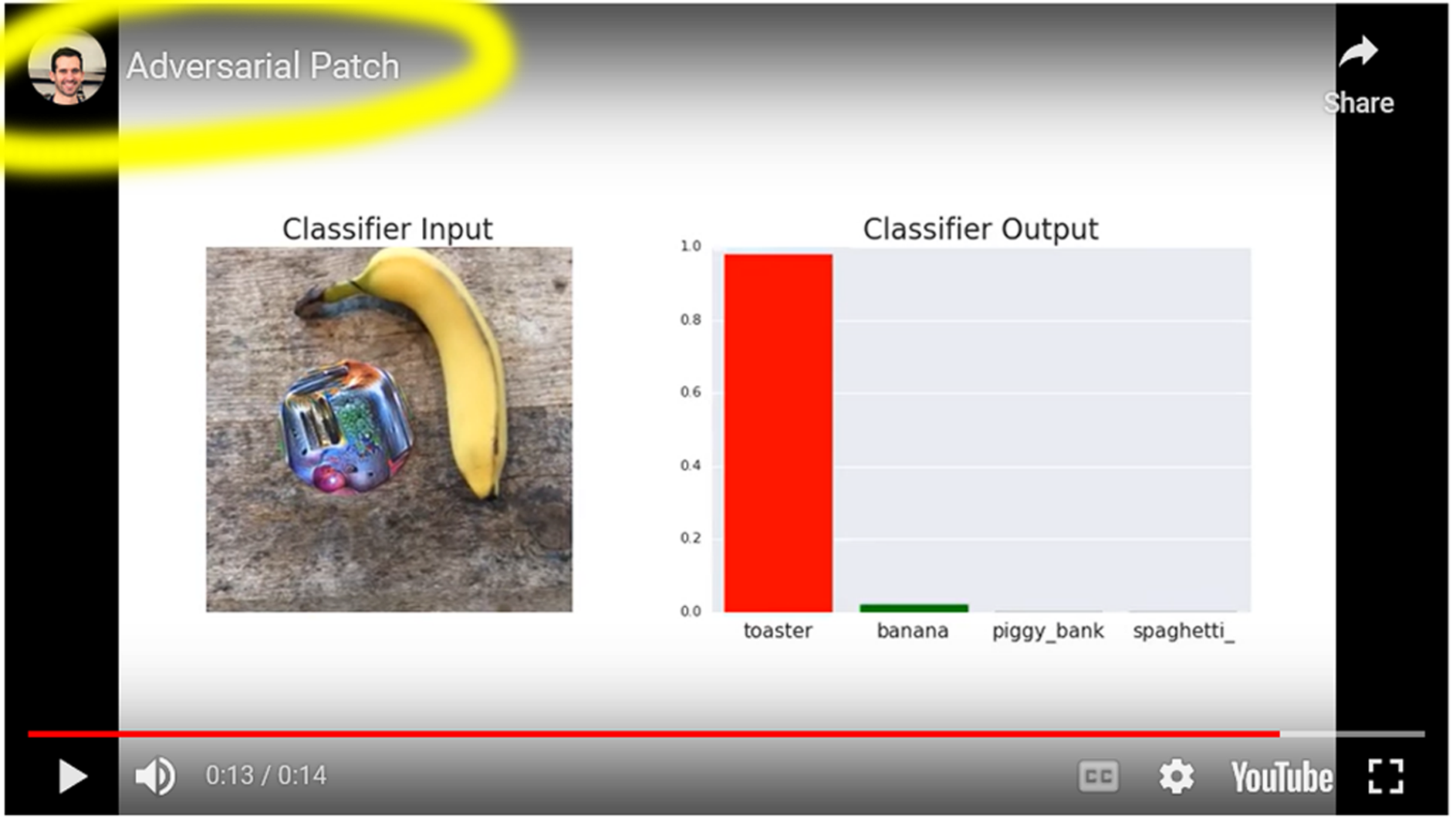
2 Safety extremes

Dynamics for measuring safety

Safety strategies

Philosophies of innovation and safety

Philosophies of innovation and safety: *Slow down, or speed up?*



Philosophies of innovation and safety: *Slow down, or speed up?*

Acceleration AI ethics
GenHumanism
Transhumanism (Under Cartesian definition)

HUMAN/MACHINE
Human terms define machine excellence

Slow AI (US, Dair)
Precaution (Europe)
Effective Altruism
"Less Wrong"

INNOVATION/SAFETY
Innovation creates safety

SAFETY/INNOVATION
Safety allows innovation

Posthumanism
Effective accelerationism
"Techno-optimism"
"Move fast and break things"

MACHINE/HUMAN
Mechanical terms define human excellence

Philosophies of innovation and safety: *Slow down, or speed up?*

Elements of Acceleration Ethics

- Innovation solves innovation problems
- Innovation is intrinsically valuable (Tips burden)
- Uncertainty is encouraging (Intrinsic, like nomadic travel)
- Decentralization (After and from users instead of before and from authorities)
- Embedding (Organic and within instead of outside and above)

Outline

Accidents

What is safety?
  Robust and resilient AI

2 Safety extremes

Dynamics for measuring safety

Safety strategies

Philosophies of innovation and safety