

CLOUDS...



SAFETY: CASE

TESLA



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.

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"?)*

SAFETY: CASE

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

What should be measured?

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

The Self-Driving Car Companies Going The Distance

Number of autonomous test miles and miles per disengagement (Dec 2019-Nov 2020)*

			Miles	Miles per disengagement
Waymo (Alphabet)			628,839	29,945
Cruise (GM)			770,049	28,520
AutoX			40,734	20,367
Pony.AI			225,496	10,738
Argo.AI (Ford, VW)			21,037	10,519
WeRide			13,014	6,507
DiDi Chuxing			10,401	5,201
Nuro			55,370	5,034



* Cases where a car's software detects a failure or a driver perceived a failure, resulting in control being seized by the driver.

Source: DMV California, via The Last Driver License Holder



Forbes statista

SAFETY

What is the *value* of safety?

How can it be described,
or measured?

OUTLINE

What is safety? Experience

What is safety? Definition

2 Safety threats

Dynamics for measuring safety

Safety strategies

Innovation and safety

DEFINITION SAFETY

AI is robust (responds to unfamiliar inputs, "inference") and resilient (bounces back)

Tesla that functions well in San Francisco also does well in Trento / Tesla stops for ducks crossing the road (AI had no previous training in ducks...)

Tesla stops for other cars even when cannot decode badly discolored stop sign (resilient, other systems compensate error)



OUTLINE

What is safety? Experience

What is safety? Definition

2 Safety threats

Dynamics for measuring safety

Safety strategies

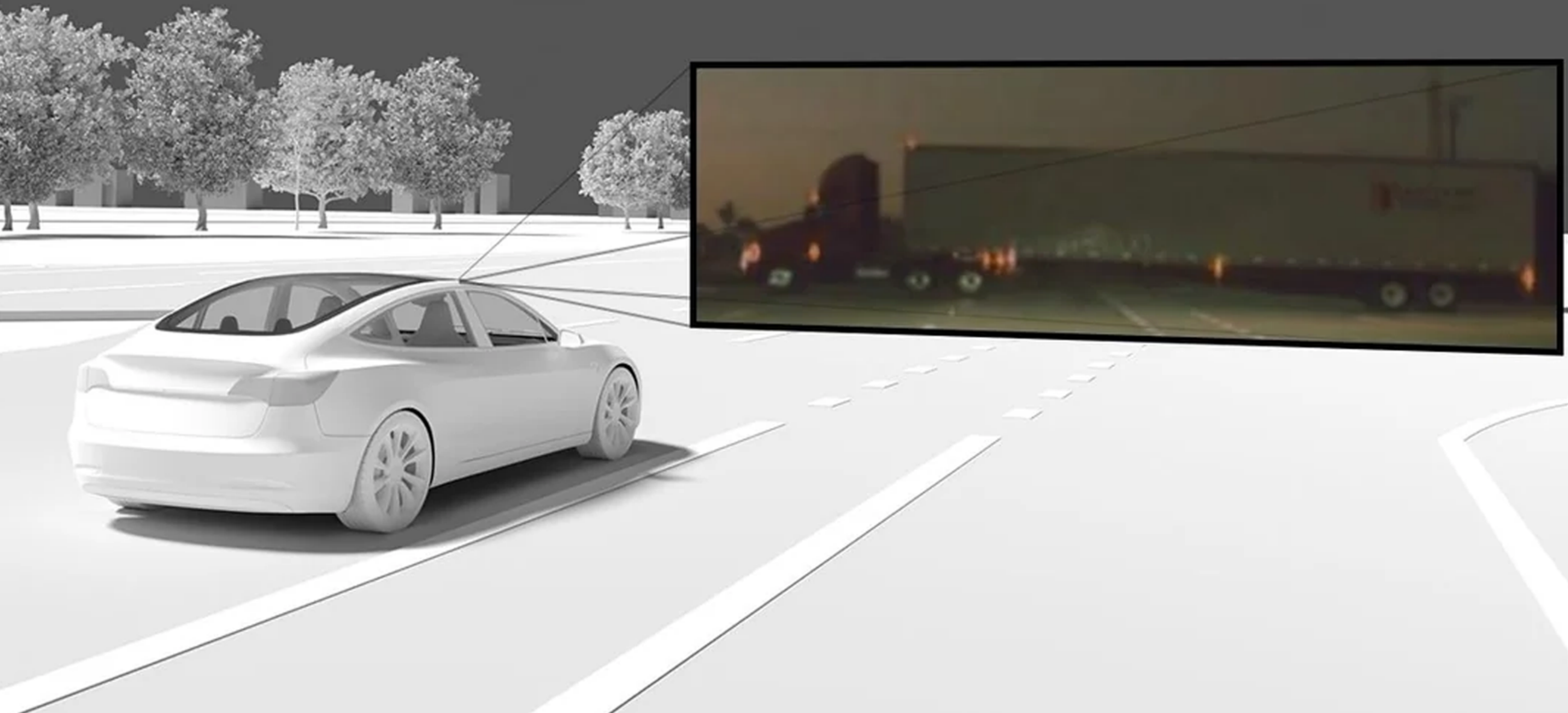
Innovation and safety

2 AI SAFETY THREATS: UNDERPERFORM, OVERPERFORM

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.

research and

As a cognitive pursuit, a superintelligence could also easily surpass humans in the quality of its work. The superintelligence may become unstoppable because of its intellectual superiority. This paper surveys some of the unique ethical issues in creating superintelligence, and discusses 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 the
corporate
level,

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

What is safety? Experience

What is safety? Definition

2 Safety threats

Dynamics for measuring safety

Safety strategies

Innovation and safety

DYNAMICS FOR MEASURING SAFETY

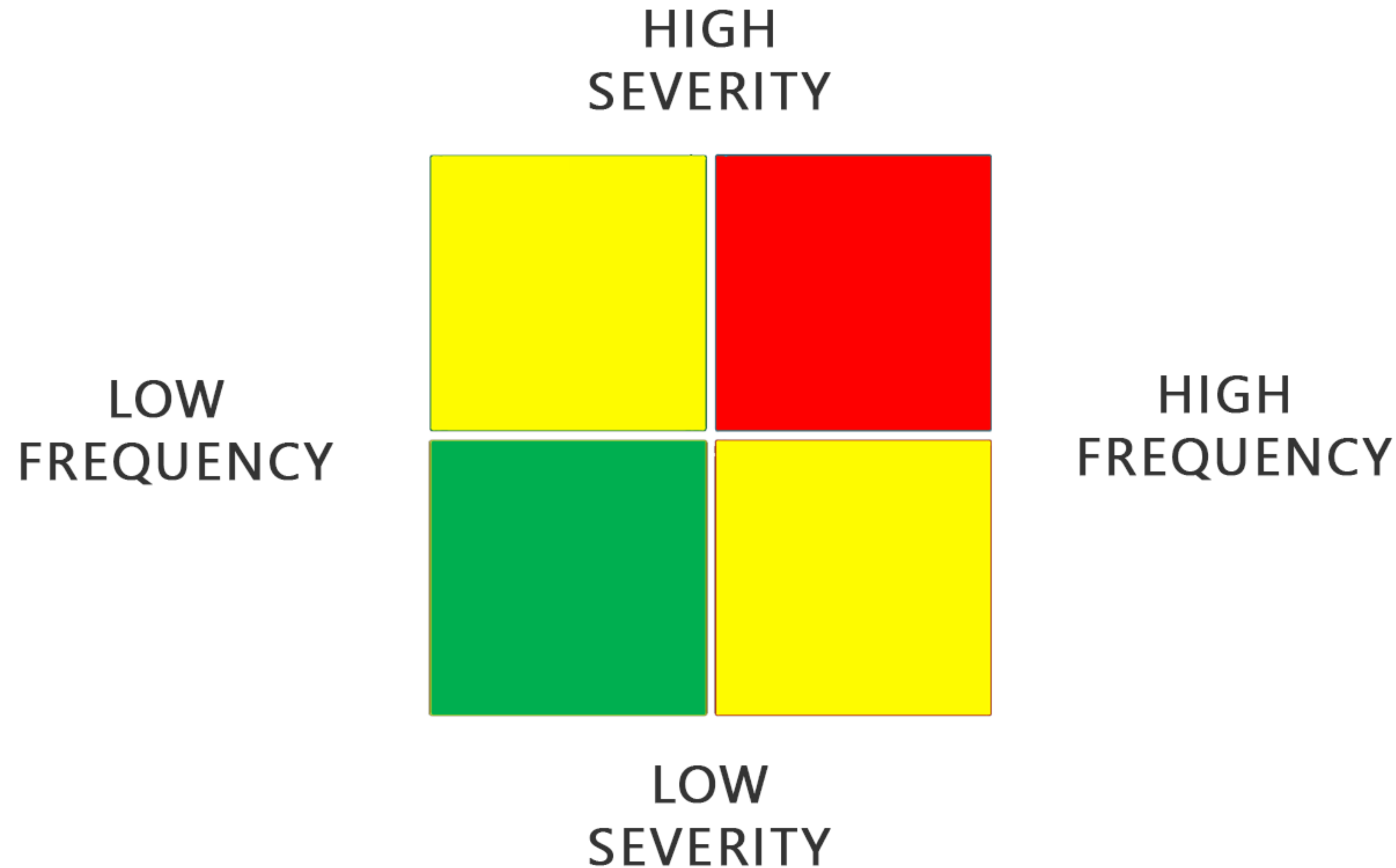
Frequency and severity

Normalization (flatten to 1 score)

Degrees of uncertainty

DYNAMICS FOR MEASURING SAFETY: MAPPING FREQUENCY AND SEVERITY

Netflix, Tesla, AI hypersonic weapons?



DYNAMICS FOR MEASURING SAFETY: NORMALIZATION (FLATTEN TO 1 SCORE)

How many bad Netflix recommendations = 1 Tesla autopilot fatality?

Master Algorithm for comparing safety?



DYNAMICS FOR MEASURING SAFETY: NORMALIZATION (FLATTEN TO 1 SCORE)

HOW MANY BAD NETFLIX RECOMMENDATIONS WOULD YOU DISTRIBUTE TO RANDOM PEOPLE TO SAVE 1 RANDOM PERSON FROM A TESLA AUTOPILOT PHANTOM-BRAKE COFFEE SPILL?

HOW MANY TESLA AUTOPILOT PHANTOM-BRAKE COFFEE SPILLS WOULD YOU DISTRIBUTE TO RANDOM PEOPLE TO SAVE 1 RANDOM PERSON FROM A SPRAINED NECK?

HOW MANY SPRAINED NECKS WOULD YOU DISTRIBUTE TO RANDOM PEOPLE TO SAVE 1 RANDOM PERSON FROM A MILD CONCUSSION (HEADACHE, NO LOSS OF CONSCIOUSNESS)?

HOW MANY MILD CONCUSSIONS (HEADACHE, NO LOSS OF CONSCIOUSNESS) WOULD YOU DISTRIBUTE TO RANDOM PEOPLE TO SAVE 1 RANDOM PERSON FROM A FRACTURED PINKY FINGER?

HOW MANY FRACTURED PINKY FINGERS WOULD YOU DISTRIBUTE TO RANDOM PEOPLE TO SAVE 1 RANDOM PERSON FROM A BROKEN ARM?

HOW MANY BROKEN ARMS WOULD YOU DISTRIBUTE TO RANDOM PEOPLE TO SAVE 1 RANDOM PERSON FROM A BROKEN PELVIS?

HOW MANY BROKEN PELVISES WOULD YOU DISTRIBUTE TO RANDOM PEOPLE TO SAVE 1 RANDOM PERSON FROM A BROKEN BACK, WITHOUT PARALYSIS?

HOW MANY BROKEN BACKS WITHOUT PARALYSIS WOULD YOU DISTRIBUTE TO RANDOM PEOPLE TO SAVE 1 RANDOM PERSON FROM A BROKEN BACK WITH PARALYSIS FROM THE NECK DOWN?

HOW MANY BROKEN BACKS WITH PARALYSIS FROM THE NECK DOWN WOULD YOU DISTRIBUTE TO RANDOM PEOPLE TO SAVE 1 RANDOM PERSON FROM A FATALITY?

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 lightning 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.

OUTLINE

What is safety? Experience

What is safety? Definition

2 Safety threats

Dynamics for measuring safety

Safety strategies

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/OUTSIDE 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.

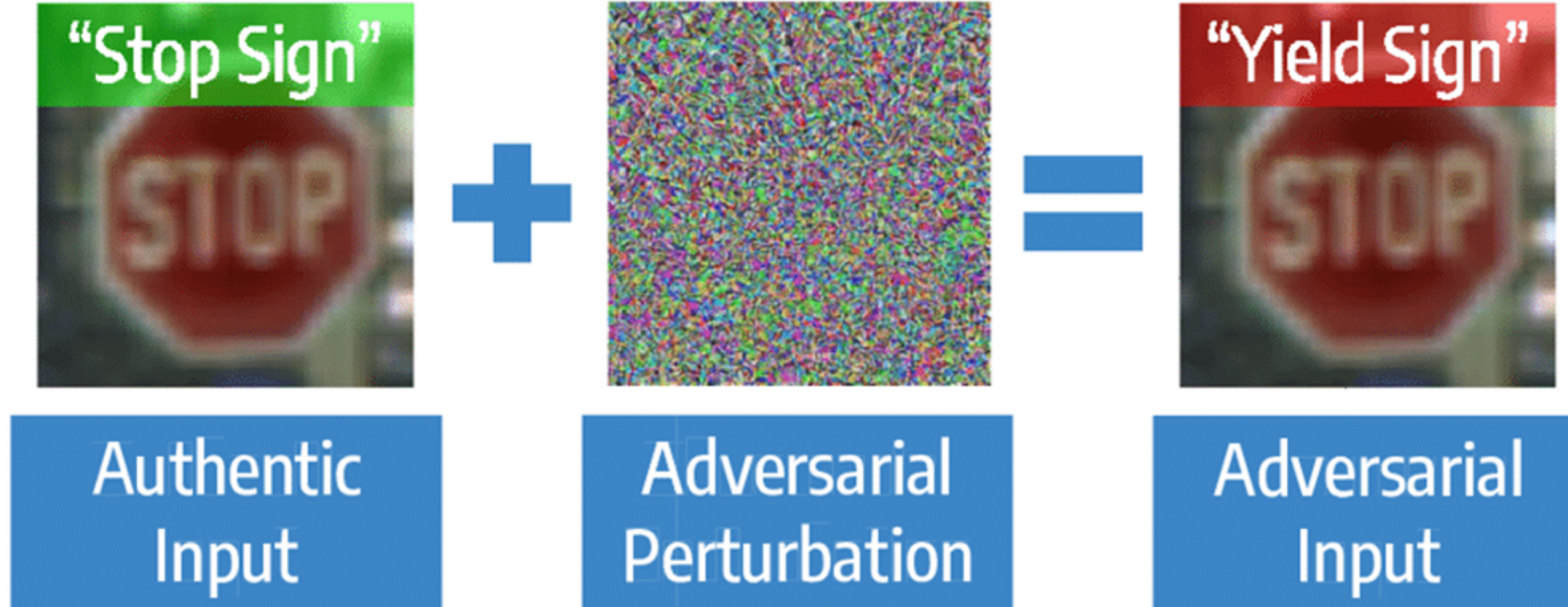
SAFETY STRATEGIES: RESILIENT AGAINST SABOTAGE...?

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

What is safety? Experience

What is safety? Definition

2 Safety threats

Dynamics for measuring safety

Safety strategies

Innovation and safety

OUTLINE

What is safety? Experience

What is safety? Definition

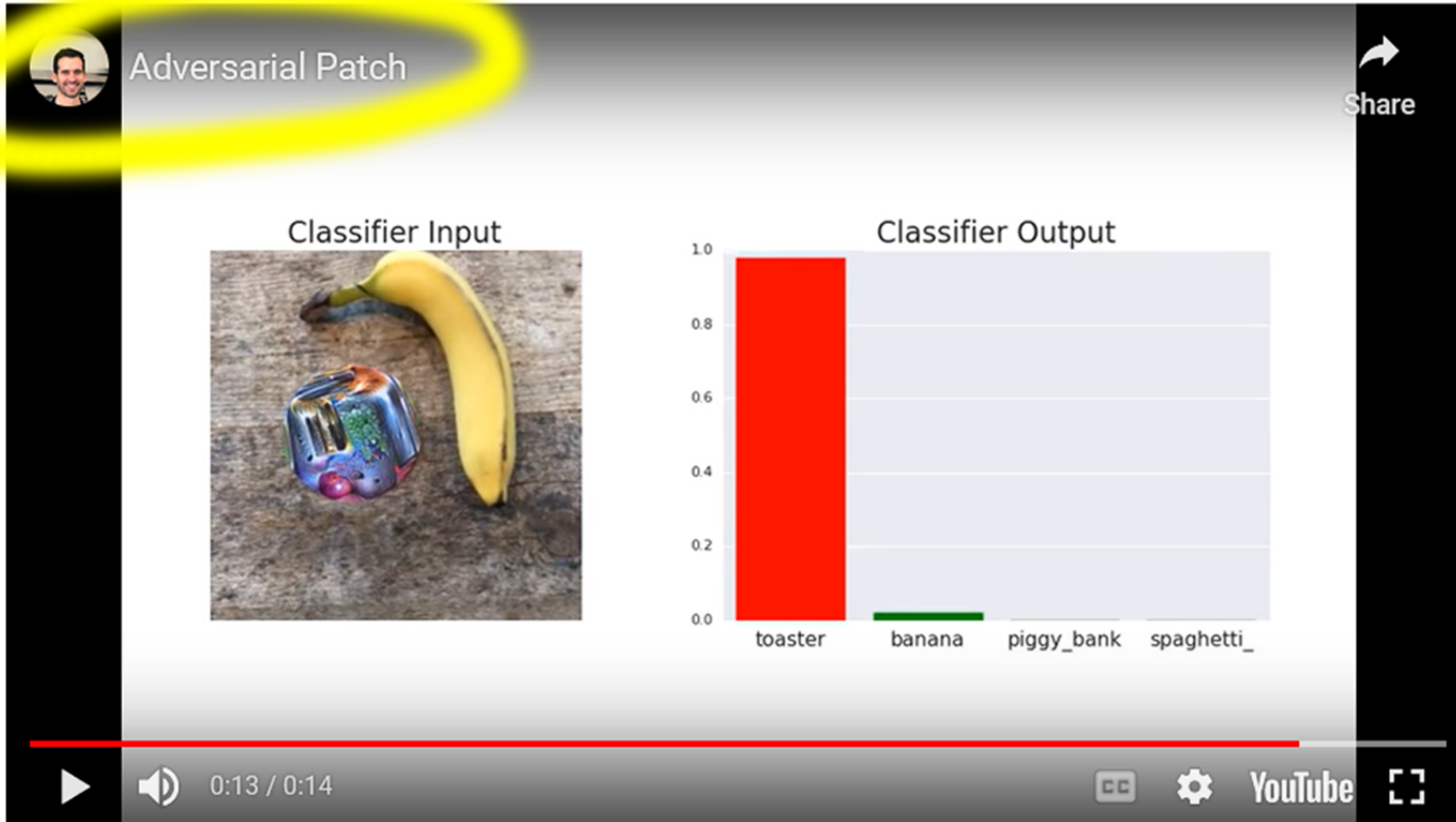
2 Safety threats

Dynamics for measuring safety

Safety strategies

Innovation and safety

PHILOSOPHIES OF INNOVATION AND SAFETY: *SLOW DOWN, OR SPEED UP?*



PHILOSOPHIES OF INNOVATION AND SAFETY: *SLOW DOWN, OR SPEED UP?*

ACCELERATION	PRECAUTION
Innovation creates safety	Safety allows innovation
Innovation is intrinsically valuable	Innovation is only as valuable as its consequences
Unknown is encouraging	Unknown is discouraging
Decentralization	Centralization

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)

PHILOSOPHIES OF INNOVATION AND SAFETY: *SLOW DOWN, OR SPEED UP?*

PostHumanism
Effective acceleration
Techno-optimism

AI/HUMAN
AI terms define
human excellence

INNOVATION/SAFETY
Innovation creates safety

SAFETY/INNOVATION
Safety allows innovation

TransHumanism
GenHumanism
Acceleration AI ethics
Transitional acceleration

HUMAN/AI
Human terms define
AI excellence

Effective Altruism
Slow AI (US, DAIR)
Precaution (Europe)
“Shut it down”