

4 AI revolutions in
AI happened in the
last 4 years -
should we be
alarmed?

Shai Shalev-Shwartz

Joint work with A. Shashua





Prologue: 6 AI Revolutions

Machine Learning

Deep Learning

Self-supervised Foundation models

Consolidation

Sim2Real

Reasoning



Pre- Revolution

2000: I was working at a startup,
designing classical AI Computer
Vision for Detecting ellipses in
images



1. ML revolution

- 2010: I've Joined Mobileye
- Shifting from expert systems to learning algorithms
- SVM and AdaBoost on hand-tuned features

2. Deep Learning

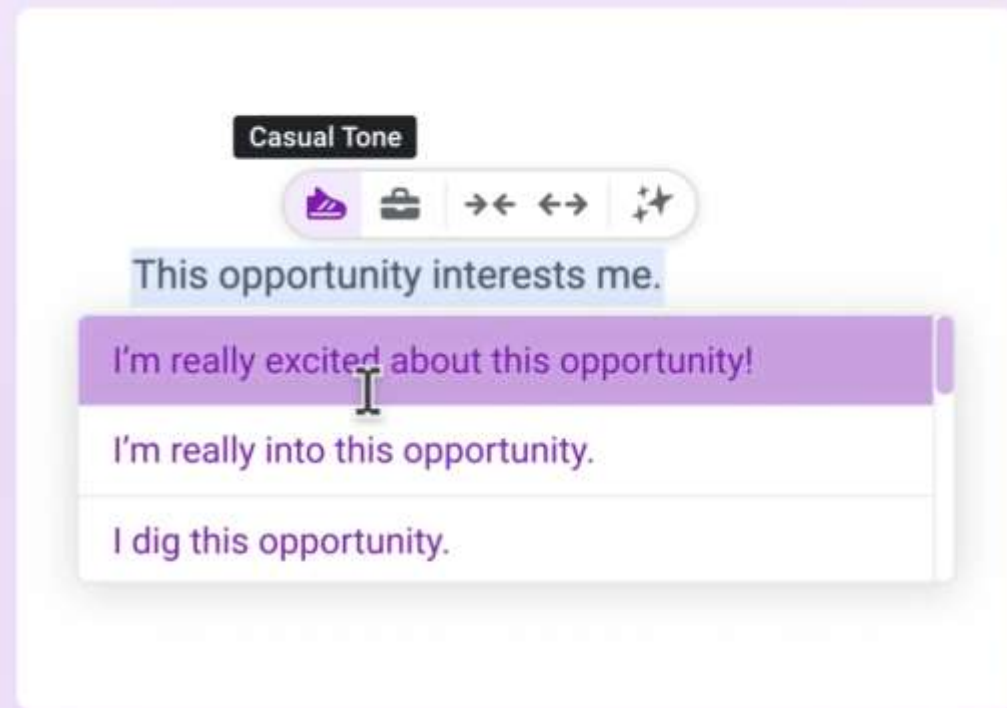
- Instead of hand-tuned features, let the machine learn the features as well
- Inspired by Krizhevsky-Sustkever-Hinton 2012 paper- Orienting Mobileye to deep learning models
- Deployment of deep learning models into an embedded system: EyeQ3 Chip
- Mobileye was first worldwide to do so (2014, on Tesla 1st gen Autopilot)



Your thoughts in words

AI21's premier product is a writing companion tool that helps you rephrase your writing to say exactly what you mean.

Start Writing



3. Self-supervised Foundation models

- During the 2010's, supervised deep learning became very successful, whenever:
 - a company could invest in generating high quality labeled data
 - for specific problems
- 2019; Joining AI21Labs, working on “rephrasing”
 - Supervised learning is not practical (hard to collect many good examples)
- Self-supervised foundation models:
 - No labels. Use parts of the data as “fake labels”
 - E.g.: instance is context, fake label is next word
 - In context learning: explain the “task” to the model in natural language

- **Pre-training:** given t_1, \dots, t_n , learn a network that minimizes:
 - $\sum \log P(t_i | t_{i-1}, \dots, t_1)$
 - Self-supervise, task-agnostic, can train over trillion of tokens
- **In context learning :** the model is trained to be a good model of the internet. Prompt-engineering helps the model to learn what task do we want it to solve.

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 cheese => _____ ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← example
3 cheese => _____ ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

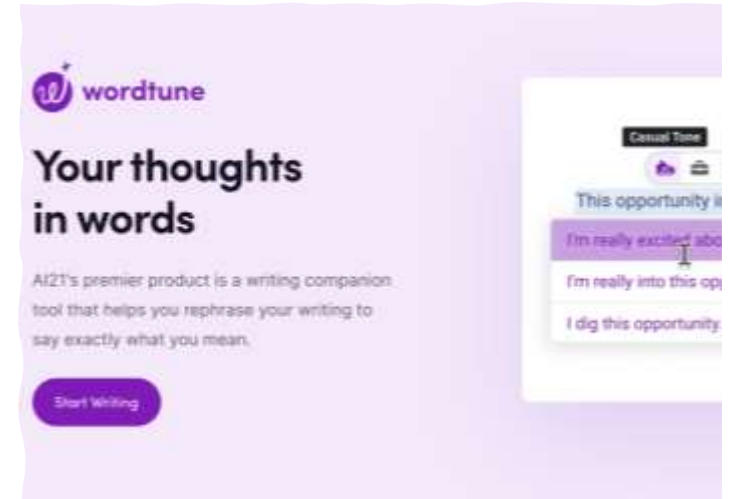
```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← examples
3 peppermint => menthe poivrée ←
4 plush girafe => girafe peluche ←
5 cheese => _____ ← prompt
```

4. Consolidation Revolution

There used to be AI experts in
Computer Vision, Speech Processing,
Natural Language Processing, etc.

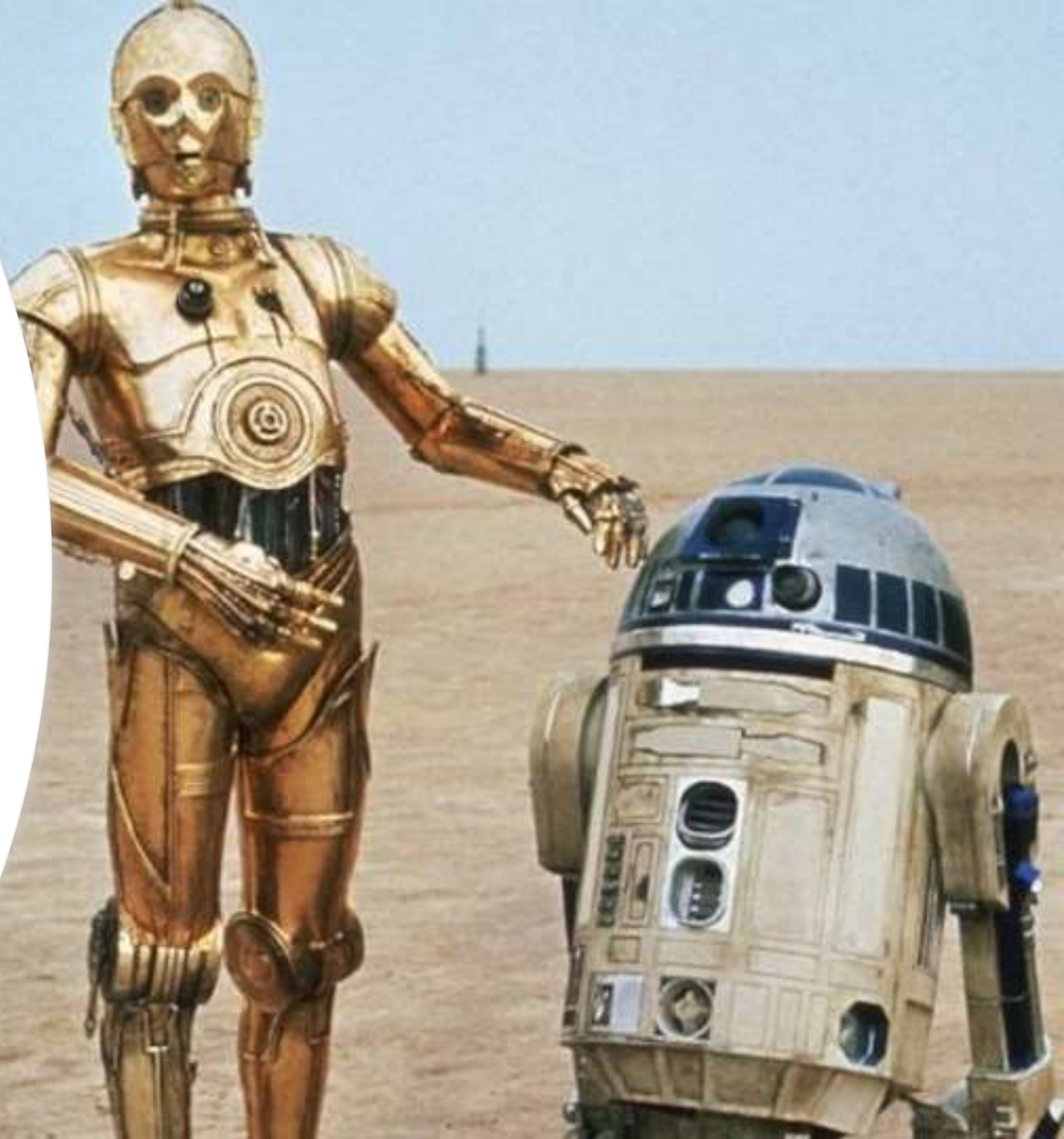
Today: deep learning dominates all
these fields

Even the architectures are being
consolidated (“attention is all you
need”)



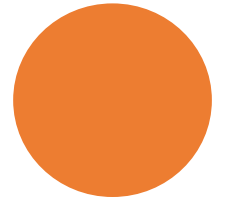
5. The Sim2Real Revolution

- Two learning paradigms:
 - “Learning from data”
 - “Learning from experience”
- Real world experience- slow and dangerous
- Simulation can be safe and fast, but doesn't reflect reality
- Sim2Real: learn in simulation, apply in real world



6. Reasoning

- **Reasoning** The holly grail of AI
- Classical ML: generalize on unseen data, if train and test are i.i.d. from the same distribution
- Reasoning:
“instruction” + “few examples” =
“out of distribution generalization”



How come suddenly Reasoning emerges from Large Language Models (LLMs) ?

- LLMs were not trained to Reason... what happened?
- Our best guess, so far, is that the key is **training on code**



Chain-of-thought prompting elicits reasoning in LLMs (Wei-et-al, Jan. 2022)

Standard prompting

Input: Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?

A: The answer is 9.

...

Q: John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. How many hours a week does he spend taking care of dogs?

A:

Model output: The answer is 50. ❌

Chain of thought prompting

Input: Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?

A: Shawn started with 5 toys. If he got 2 toys each from his mom and dad, then that is 4 more toys. $5 + 4 = 9$. The answer is 9.

...

Q: John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. How many hours a week does he spend taking care of dogs?

A:

Model output: John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. So that is $10 \times .5 = 5$ hours a day. $5 \text{ hours a day} \times 7 \text{ days a week} = 35 \text{ hours a week}$. The answer is 35 hours a week. ✅

“PAL: Program Aided LMs” (Gao-et-al, Jan. 2023)

Chain-of-Thought (Wei et al., 2022)

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 tennis balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The bakers at the Beverly Hills Bakery baked 200 loaves of bread on Monday morning. They sold 93 loaves in the morning and 39 loaves in the afternoon. A grocery store returned 6 unsold loaves. How many loaves of bread did they have left?

Model Output

A: The bakers started with 200 loaves. They sold 93 in the morning and 39 in the afternoon. So they sold $93 + 39 = 132$ loaves. The grocery store returned 6 loaves. So they had $200 - 132 - 6 = 62$ loaves left. The answer is 62.



Program-aided Language models (this work)

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 tennis balls.
`tennis_balls = 5`
2 cans of 3 tennis balls each is
`bought_balls = 2 * 3`
tennis balls. The answer is
`answer = tennis_balls + bought_balls`

Q: The bakers at the Beverly Hills Bakery baked 200 loaves of bread on Monday morning. They sold 93 loaves in the morning and 39 loaves in the afternoon. A grocery store returned 6 unsold loaves. How many loaves of bread did they have left?

Model Output

A: The bakers started with 200 loaves
`loaves_baked = 200`
They sold 93 in the morning and 39 in the afternoon
`loaves_sold_morning = 93`
`loaves_sold_afternoon = 39`
The grocery store returned 6 loaves.
`loaves_returned = 6`
The answer is
`answer = loaves_baked - loaves_sold_morning`
`- loaves_sold_afternoon + loaves_returned`

```
>>> print(answer)  
74
```



Neuro-Symbolic reasoning

- Statistical Learning – “from examples” (induction)
- Symbolic Learning – “from rules” (deduction)
 - E.g., learning to multiply any 2 numbers using examples doesn’t make sense. Better to learn the long multiplication algorithm
- General neuro-symbolic approach:
 - Translate problems into python code snippets, execute them, and translate the python output to an answer
- Karpas et al, “MRKL systems: A modular, neuro-symbolic architecture”, AI21-labs, 5/2022
- Shick et al, “Toolformer”, Meta, 2/2023

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

Figure 1: Exemplary predictions of Toolformer. The model autonomously decides to call different APIs (from top to bottom: a question answering system, a calculator, a machine translation system, and a Wikipedia search engine) to obtain information that is useful for completing a piece of text.

Where are we going from here?

- The rise of AI models that can
 - “learn in context” and “reason”
 - Operate in the real world (“sim2real”)

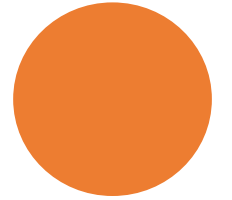
will create a new era of machines/robots that can be instructed by humans to perform various new tasks with out-of-distribution generalization

Should we be alarmed?
Is AI Dangerous?
Why aren't we scared?



Safety of AI

- Safety of Self-Driving cars
- The AI alignment problem



Autonomous
ACTIVE
0
km/h
Limit
35
-51°



Jerusalem



Once he gets back in to give us more space to pass, the AV is able to continue.

AV's main challenge

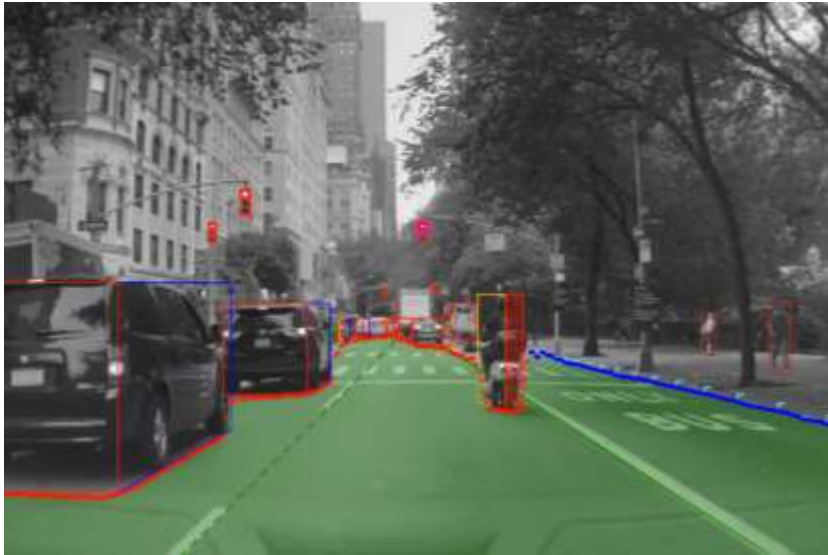
There exist sophisticated technologies with low accuracy requirements, or “simple” technologies with high accuracy requirements. AV is both ...



Sense / Plan / Act Robotic Methodology

Sense

Perception of the environment



Plan (Driving Policy)

Decision making

“What would happen if”
type of reasoning



Act (Control)

Execute the plan



Safety Elements --- what can go wrong?

- Software and hardware bugs (“heart attack, fell asleep”)
- Perception errors (“I didn’t see this car”)
- Bad decision making (“I thought I can pass before him”)
- Actuation error (“I hit the gas instead of the brake”)

This talk is about Safety of Driving Policy

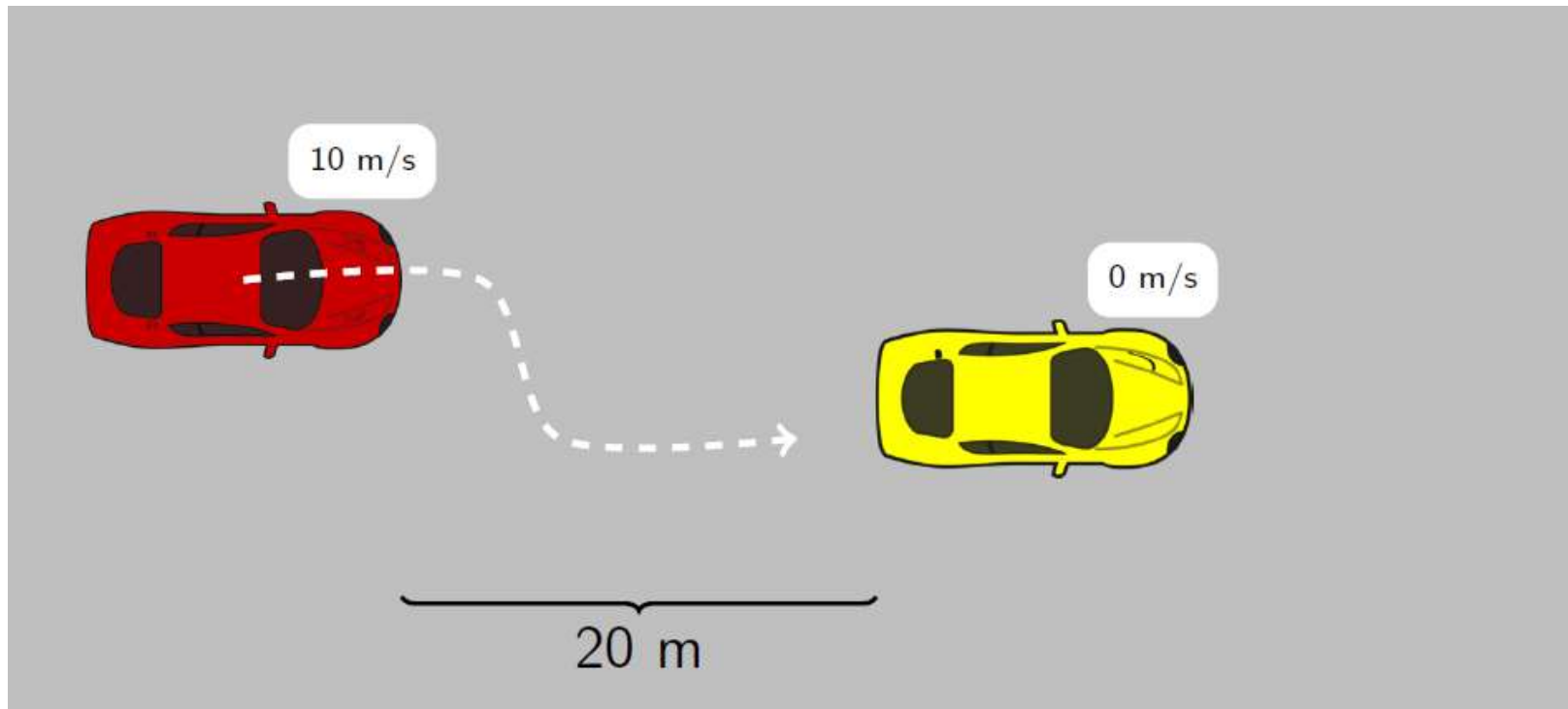
Why Driving Policy is Difficult?

- Close loop:
 - Actions of the ego vehicle affect other road users (e.g., when "pushing" in a lane change)
 - Actions that are performed now may have long term effect on the future (butterfly effect)
- Must handle uncertainties about the future (what others might do)
- No "ground truth"

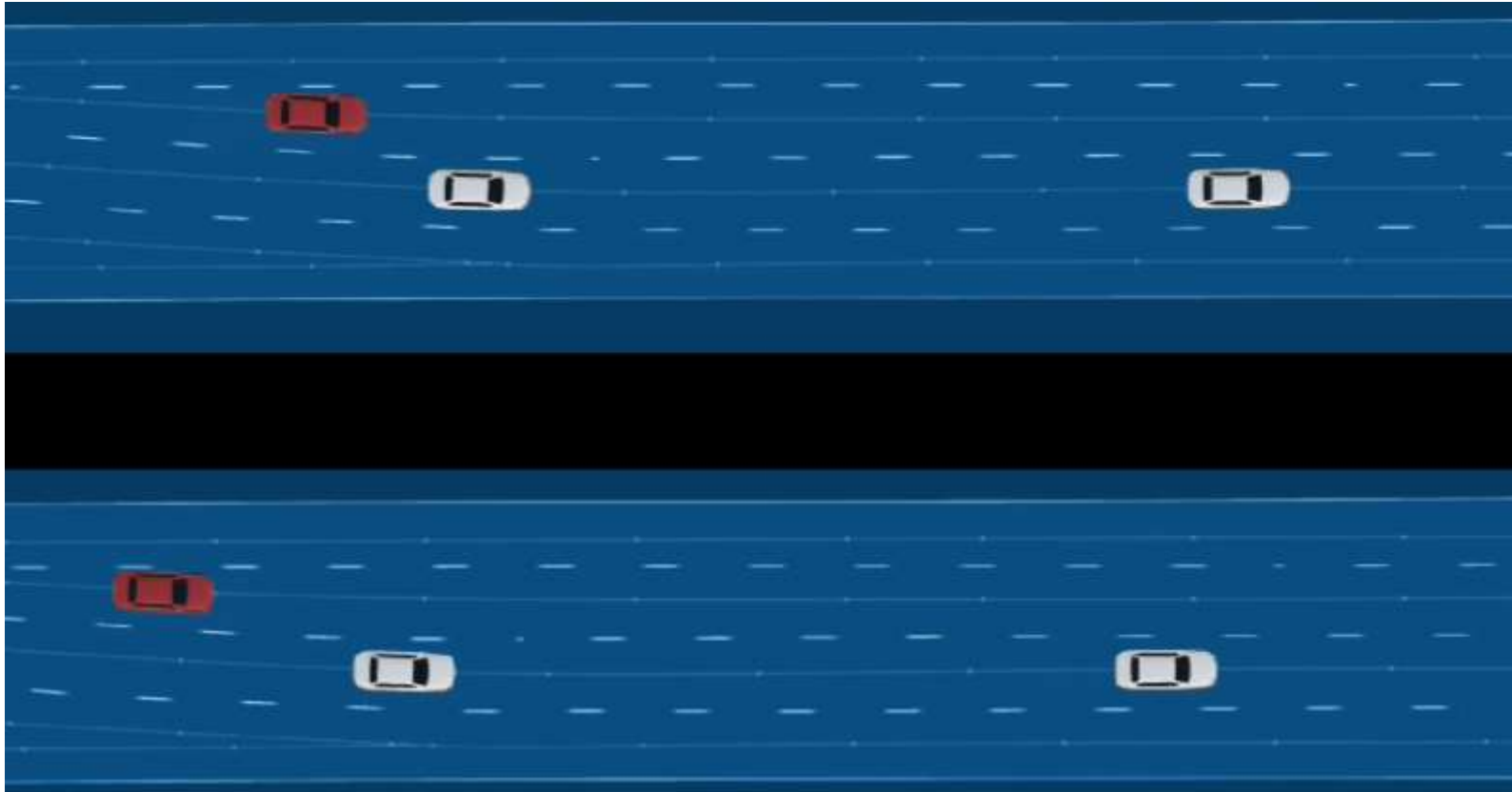


Example

- Actions that are performed now may have long term effect on the future
- Must plan for a sufficiently long time, because a bad plan might look perfectly fine at the near future



- But, this requires also to predict what other agents will do in a far future
- And, the future behavior of other agents might depend on our actions ...

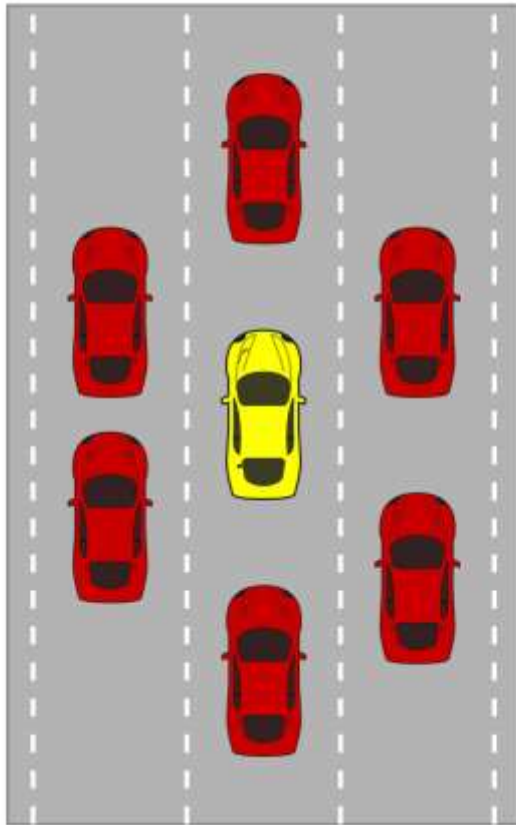


Slow truck ahead ---
must plan for a
sufficiently long
time, while other
agents respond to
our behavior

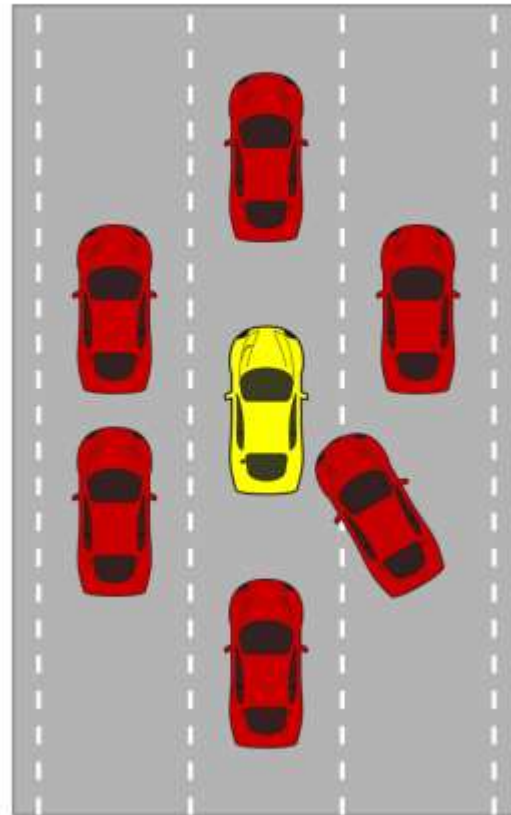


Absolute Safety is Impossible

Before



After



**Yellow car can't prevent
the accident**

Approaches to Safety of Driving Policy

- **“Statistically better than a human driver”**
 - Manifested as the “miles driven argument” , but this approach is practically impossible to validate rigorously
 - We need to re-validate after every software update
 - Close loop nature of driving prevents offline validation, unless
 - Using simulators, but who will validate the simulator
- **“Scenarios validation”**
 - Like how Autonomous-Emergency-Braking (AEB) is validated
 - Main problem: lack of completeness and over-fitting to the tests

Our Methodology: Responsibility- Sensitive-Safety (RSS)


- **Assume:** define a small set of reasonable assumptions on how other agents behave on the road, (e.g. maximal acceleration, braking, and speed in various situations)
- **Safety-net:**
 - A metric, defined on the present, that determines if the current state is “safe”
 - Proper response: what to do if the state is not safe
- **Guarantee:**
 - If all other agents comply with the assumptions, AV will not cause a collision
 - If all agents apply proper response in unsafe states, there will be no collisions

RSS advantages

- **Transparency and soundness:**
 - We clearly define what assumptions we make on other road agents
 - Society (through regulators) can affect the assumptions
- **Completeness:**
 - Beyond the RSS assumptions, all possible futures are considered. Hence, all scenarios are covered
- **Efficiency**
 - Pair-wise property without contradictions
 - Decoupling all possible futures into a property of the present state
 - One can apply **any** driving policy he likes (including extremely complicated ones), if the driving policy is override by RSS's proper response when the state is un-safe.

RSS --- Ethics

How fast should we drive in a residential road?



Driving slower --- safer, but hurts normal flow of traffic
Safety-usefulness tradeoff: this is an ethical question!
Much more important than the trolley problem ...

RSS is a language that enables to formalize the “duty of care”

Duty of Care is a legal obligation which is imposed on an individual, requiring adherence to a standard of reasonable care while performing any acts that could foreseeably harm others.

RSS: Simple car following scenario

- Reasonable assumption: front car won't brake stronger than $-a_{max}$
- Parameters: response time and maximal brake of rear car
- Proper response for rear car: apply maximal braking after response time
- Safe state: if rear will apply proper response, and front adheres to the "reasonable assumption", then there won't be a collision
- Guarantees: Suppose that rear does whatever it wants, but at the first time in which the state is not safe it performs proper response. Then, there will be no collisions.



RSS – Design Choices and Proof Techniques

- We define “proper response” so that proper response w.r.t. one agent never contradicts proper response w.r.t. another agent (pairwise property)
- Main proof technique is by **induction**:
 - **Base**: when we’re at a standstill we’re safe
 - **Step**: if current state is safe we’re fine. Otherwise, let t' be the last time in which the situation was safe. From that time and on, we performed proper response. And, the definition of safe state and proper response is s.t., under the assumptions on the other agents, we must stop before a collision

Safety of AI

- Safety of Self-Driving cars
 - A safety net, with mathematical guarantees

- Can we solve the general AI alignment problem ?



The AI Alignment Problem

- Modern AI building = Optimizing a Reward function
- Building AI to fill a cauldron, with a reward function of “1” for a full cauldron and “0” otherwise, might end up with a flood
- Why?
Our objective isn't fully aligned with what we really want (we don't care if the cauldron is 99.9% full or 100% full)




Safety net? “Stop” button ?

- A stop button is an obstacle to the AI reward, so it might stop you from pressing it
- Maybe add a large reward if stop button is pushed?
Not good --- AI will push it
- Don't allow the AI to push it?
Not good --- it'll manipulate you to push it
- Bottom line: unsolvable for AGI





Are
you scared?

- Most AI researchers are not scared
 - Why?
 - Researchers believe that the problem is “only” for AGI, but narrow AI systems are not dangerous
- 

The AI alignment problem is relevant to today's technology

- Reward for self-driving cars:
 - Safety (cost for accidents)
 - Comfort (cost for strong braking and jerk)
 - Usefulness (maximize speed up to the legal limit)
- Sounds good?
- All self-driving cars will suddenly stop, people will get confused and get out, then cars will lock themselves and start driving at a constant speed on the highway
- We “forgot” to embed in the reward that we want people to use the service ...
- Not catastrophe, but such a bug might have tremendous impact on the confidence of people in the service

Are you scared?

- Many science and technology advancements are dangerous in retrospect
 - Studying the relationship between mass and energy lead to an atomic bomb
 - Inventing the combustion engine had a big effect on climate change
- Is AI different?

Strategic vs. Agnostic Alignment Problem

- Strategic Alignment:
 - AI optimizes a reward by strategically, intentionally, changing the distribution of events in the world
- Agnostic Alignment Problem
 - AI optimizes a reward, and a distribution shift due to “butterfly effect” leads to a bad result
- The “agnostic alignment problem” is relevant to all science and technology, and in a sense, can only be avoided by stopping progress
- The “strategic alignment problem” is unique to AI

We can (and should) prevent strategic AI alignment

- Machine learning
 - Learning from data --- safe
 - Learning from experience
 - Can suffer from strategic AI alignment
 - A buffered environment and a human validator can prevent mis-alignment if we don't suffer from the matrix problem



The Matrix Problem

We might think all is good, but
it's not ...