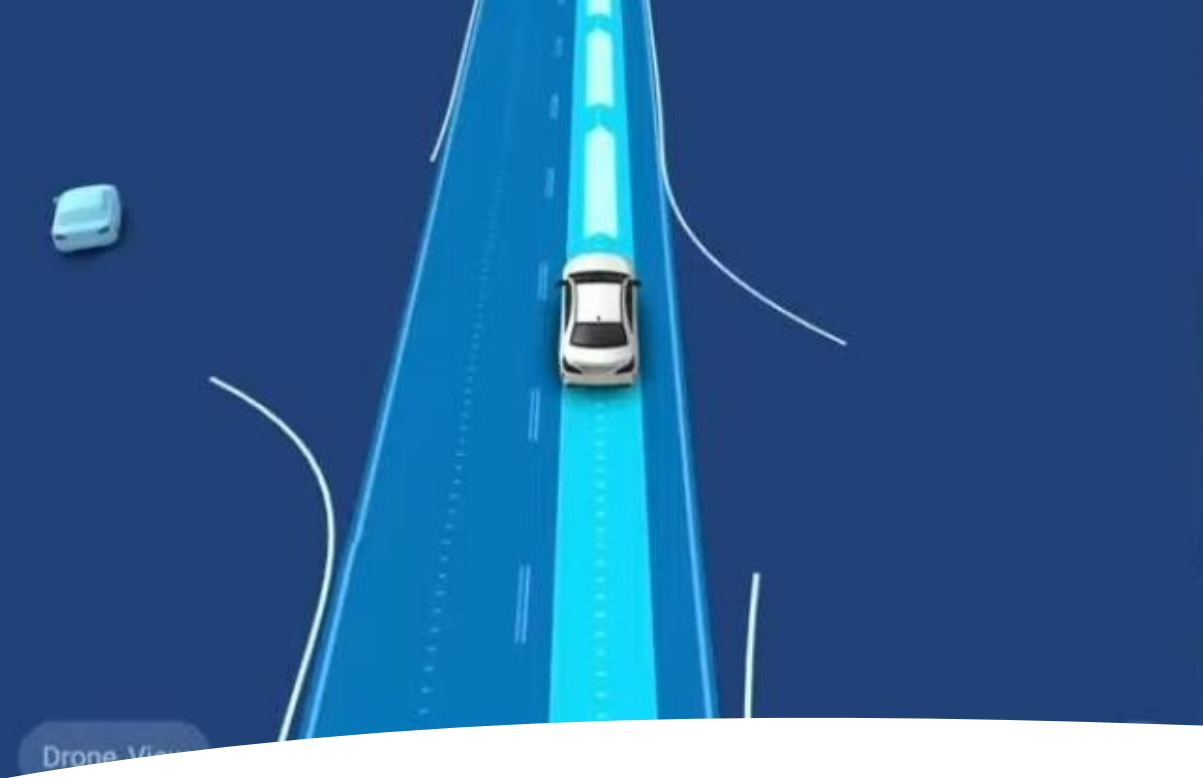


On the Ethics of Building AI Responsibly

Shai Shalev-Shwartz

Joint work with S. Shammah and A. Shashua





The AI Revolution

A screenshot of the Wordtune app interface. The background is purple. In the top left, there is a logo with a white 'W' and a star, followed by the text "wordtune". Below the logo, the text "Your thoughts in words" is displayed in yellow and white. On the right side, there are four white text boxes stacked vertically, each containing a sentence with one word highlighted in a different color. The top box has a blue highlight on "exciting", the second has a purple highlight on "great", the third has a purple highlight on "got", and the fourth has a purple highlight on "something". At the top right of the text boxes, there are icons for undo, redo, and other editing functions.

Is AI Dangerous?
And if so, why
aren't we
scared?



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)




“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
 - No strong regulation on AI development
 - 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 ...