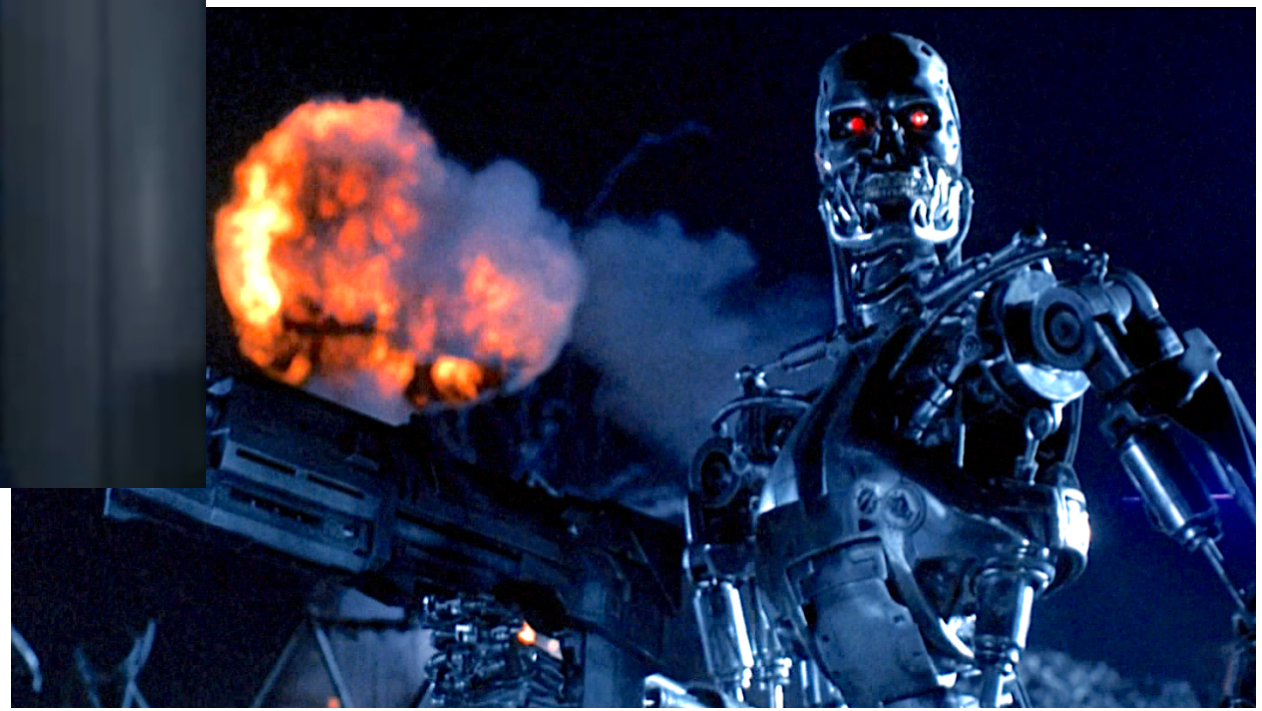
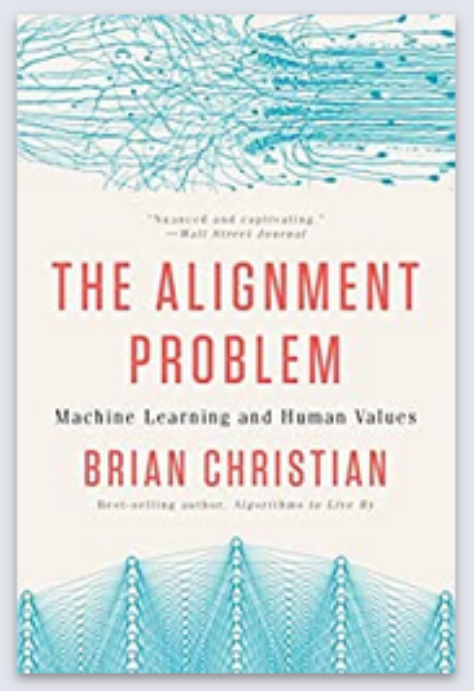
Ethics: The Alignment Problem

How do we harness artificial intelligence for the good of humanity?

The problem we expect: Skynet



The problems already here



word2vec

300-dimensional embedding trained just based on hiding words from phrases

```
Czech + currency = koruna
Vietnam + capital = Hanoi
German + airlines = Lufthansa
French + actress = Juliette Binoche*
Berlin - Germany + Japan = Tokyo
bigger - big + cold = colder
```

```
doctor – man + woman = nurse
```

Computers are better than humans at well-defined mathematical optimization

ALPHAGO

We should focus on defining problems in the **right way**

Defining Reward Functions is Hard

Hypothetical Examples:

Defining Reward Functions is Hard

Hypothetical Examples:

- Acme paper clip research division

Defining Reward Functions is Hard

Hypothetical Examples:

- Acme paper clip research division
- Asimov's laws
  - A robot may not injure a human being or, through inaction, allow a human being to come to harm.
  - A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
  - A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

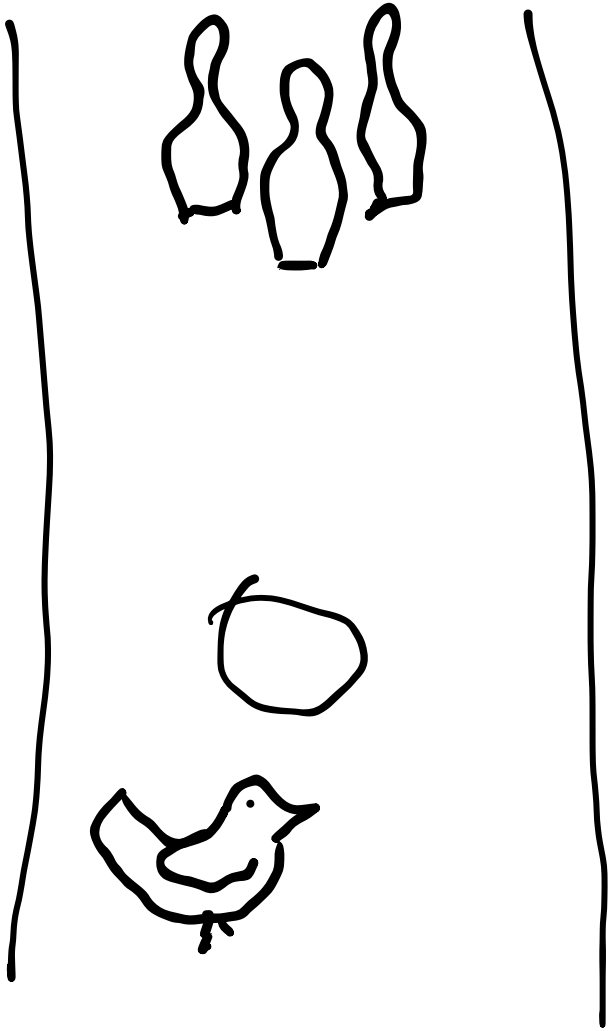
Reward Shaping

Reward Shaping

B. F. Skinner

Pigeon-guided bombs, 1943

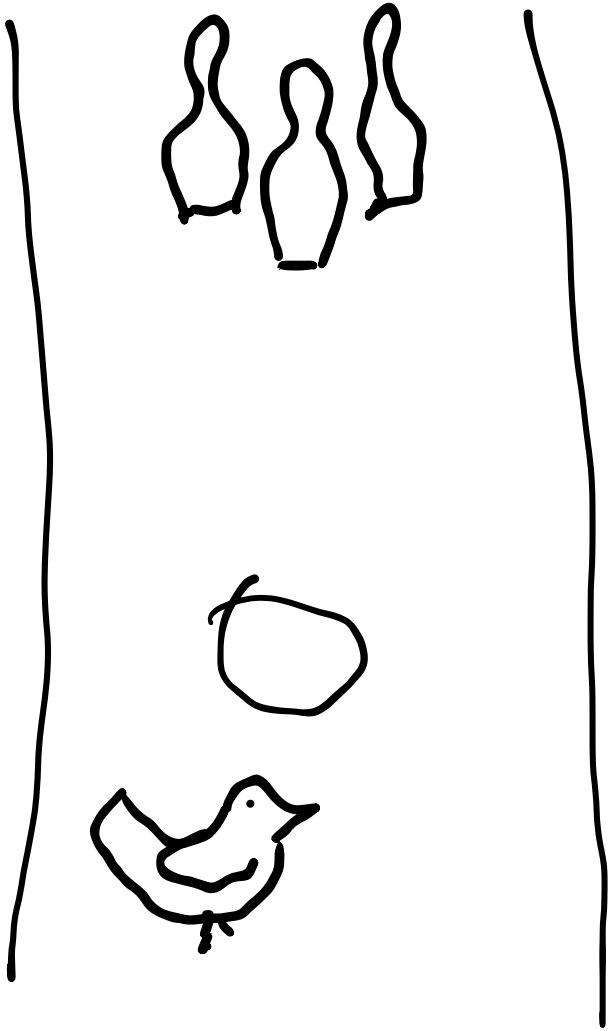
Reward Shaping



B. F. Skinner

Pigeon-guided bombs, 1943

Reward Shaping



B. F. Skinner

Pigeon-guided bombs, 1943

> We decided to reinforce any response which had the slightest resemblance to a swipe—perhaps, at first, merely the behavior of looking at the ball—and then to select responses which more closely approximated the final form. The result amazed us. In a few minutes, the ball was caroming off the walls of the box as if the pigeon had been a champion squash player.

https://www.youtube.com/embed/tlOIHko8ySg?enablejsapi=1

https://www.youtube.com/watch?v=tlOIHko8ySg

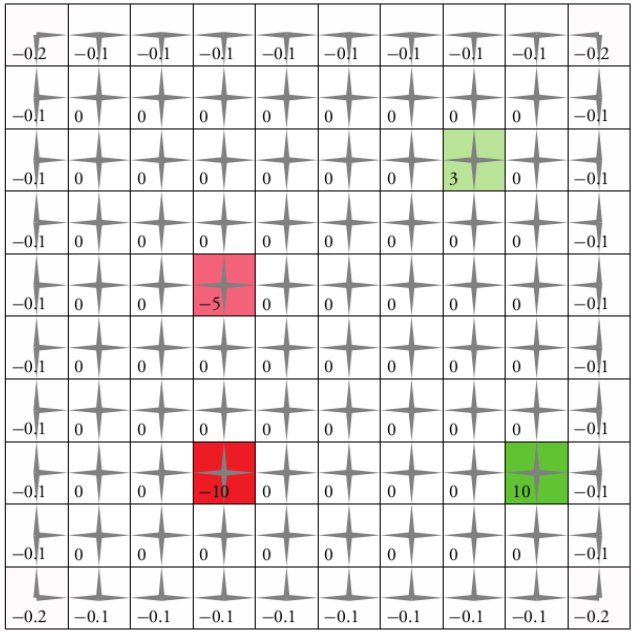
Reward Shaping

"As a general rule, it is better to design performance measures according to what one actually wants in the environment, rather than according to how one thinks the agent should behave." - Stuart Russell

Reward Shaping

"As a general rule, it is better to design performance measures according to what one actually wants in the environment, rather than according to how one thinks the agent should behave." - Stuart Russell

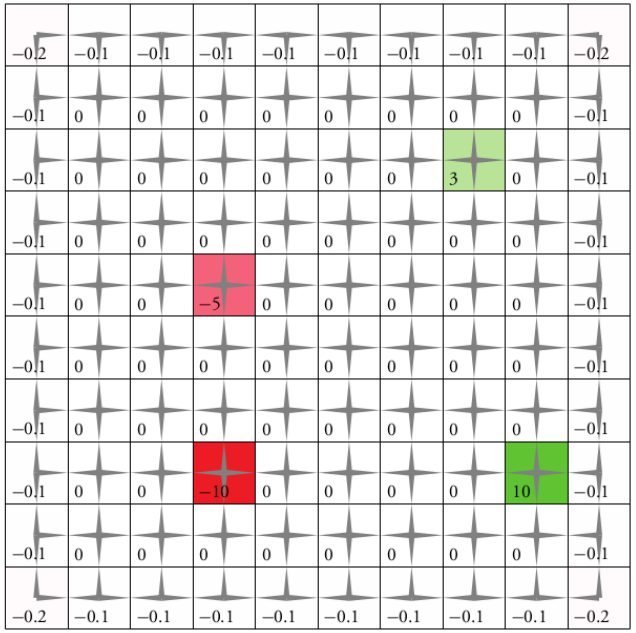
Reward



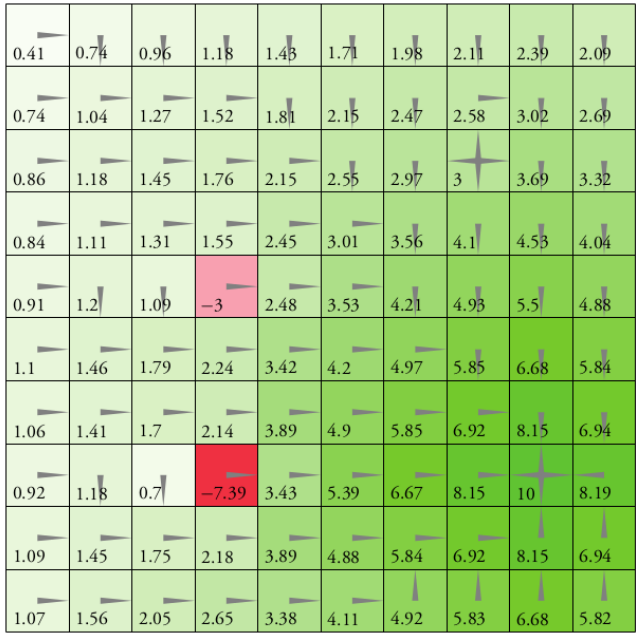
Reward Shaping

"As a general rule, it is better to design performance measures according to what one actually wants in the environment, rather than according to how one thinks the agent should behave." - Stuart Russell

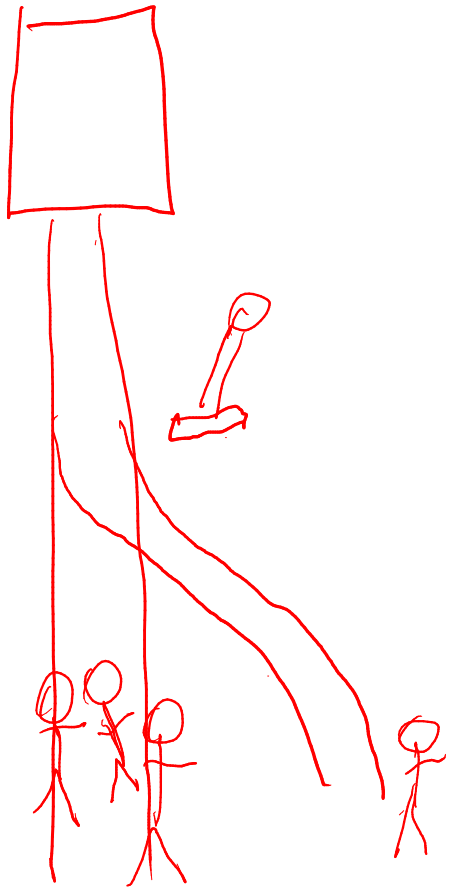
Reward



Value



Values: Trolley Problems



What should we do about it?

What should we do about it?

- Understand Uncertainty
- Know when you don't know