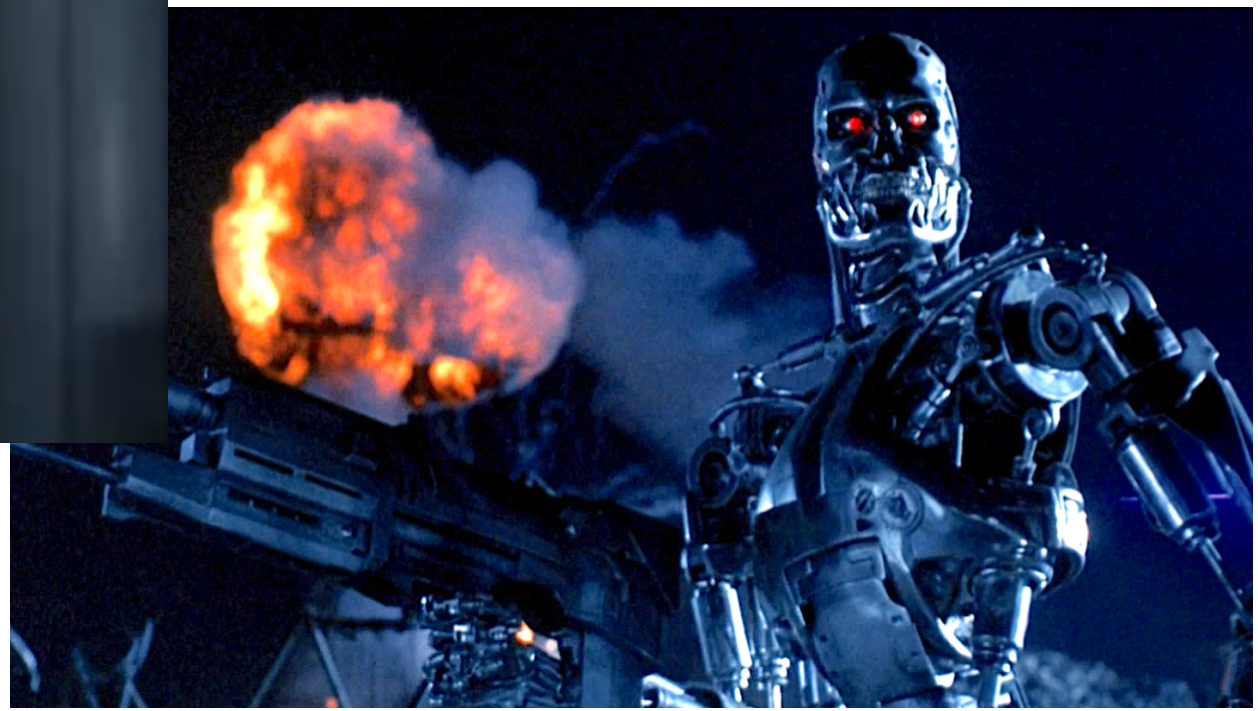
Ethics: The Alignment Problem

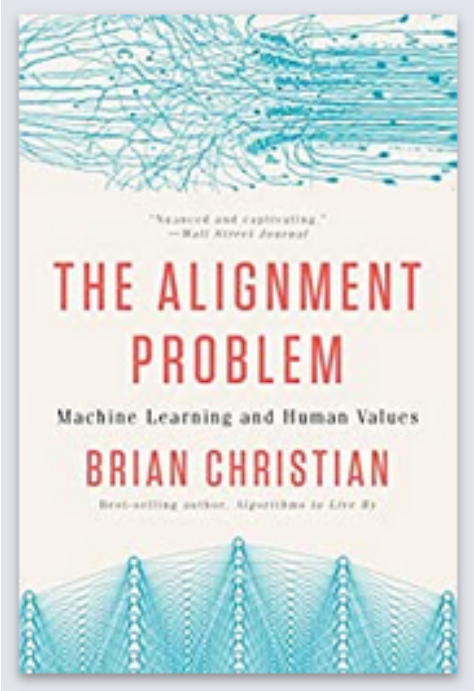
How do we harness artificial intelligence for the good of humanity?

The problem we tend to think about: Skynet

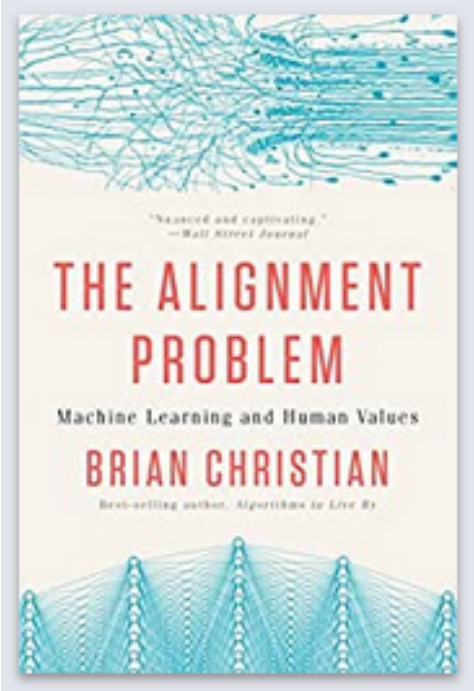


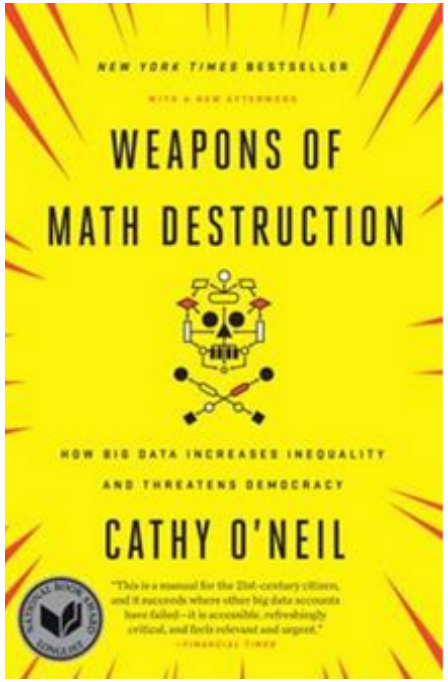
Some Problems Are Already Here

Some Problems Are Already Here

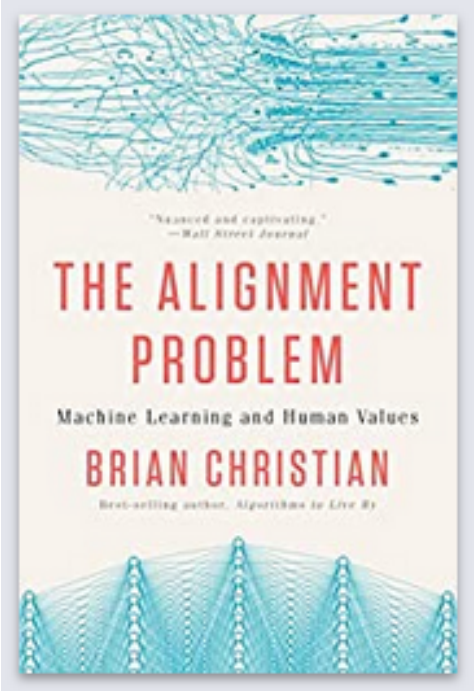


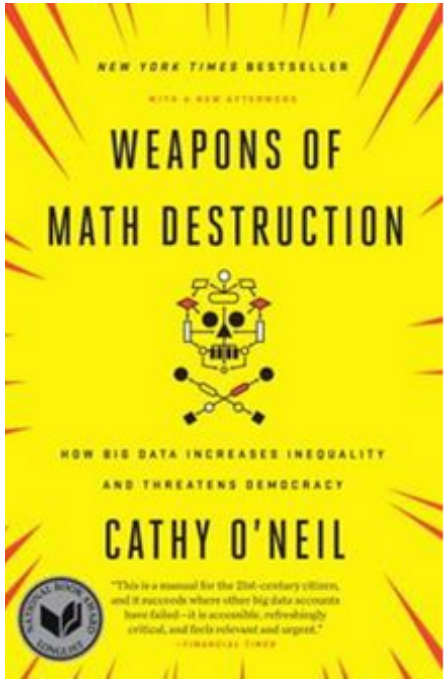
Some Problems Are Already Here

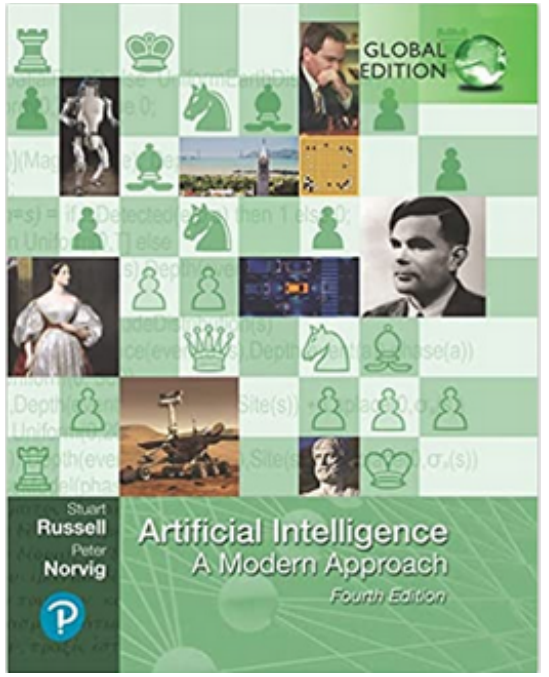




Some Problems Are Already Here







Two Categories

Two Categories

Immediate Problems

Two Categories

Immediate Problems

Long-Term Problems

Two Categories

Immediate Problems

- Weak AI

Long-Term Problems

Two Categories

Immediate Problems

- Weak AI
- Subtle Challenges

Long-Term Problems

Two Categories

Immediate Problems

- Weak AI
- Subtle Challenges

Long-Term Problems

- Strong AI

Two Categories

Immediate Problems

- Weak AI
- Subtle Challenges

Long-Term Problems

- Strong AI
- Existential Threats

Immediate Problem: Bias in Data

Immediate Problem: Bias in Data

word2vec

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

Immediate Problem: Bias in Data

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*
```

Immediate Problem: Bias in Data

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
```

Immediate Problem: Bias in Data

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
```

Immediate Problem: Bias in Data

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
```

Immediate Problem: Bias in Data

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

Immediate Problem: Difficulty removing information from Data

Immediate Problem: Difficulty removing information from Data

- date of birth + gender + zip code = % uniquely identified

Immediate Problem: Difficulty removing information from Data

- date of birth + gender + zip code = 87% uniquely identified

Immediate Problem: Fairness

COMPAS: predicting recidivism

Immediate Problem: Fairness

COMPAS: predicting recidivism

- Well-calibrated: among people with risk score of 7/10, 60% of whites and 61% of blacks re-offend

Immediate Problem: Fairness

COMPAS: predicting recidivism

- Well-calibrated: among people with risk score of 7/10, 60% of whites and 61% of blacks re-offend
- Proportion of those who did *not* re-offend, but were falsely rated high risk was 45% for blacks and 23% for whites

Immediate Problem: Fairness

COMPAS: predicting recidivism

- Well-calibrated: among people with risk score of 7/10, 60% of whites and 61% of blacks re-offend
- Proportion of those who did *not* re-offend, but were falsely rated high risk was 45% for blacks and 23% for whites

Suggested possible solution in AIMA:

"Equal Impact": assigning utility

Immediate Problem: Decision Feedback Loops

Immediate Problem: Employment

Values: Trolley Problems

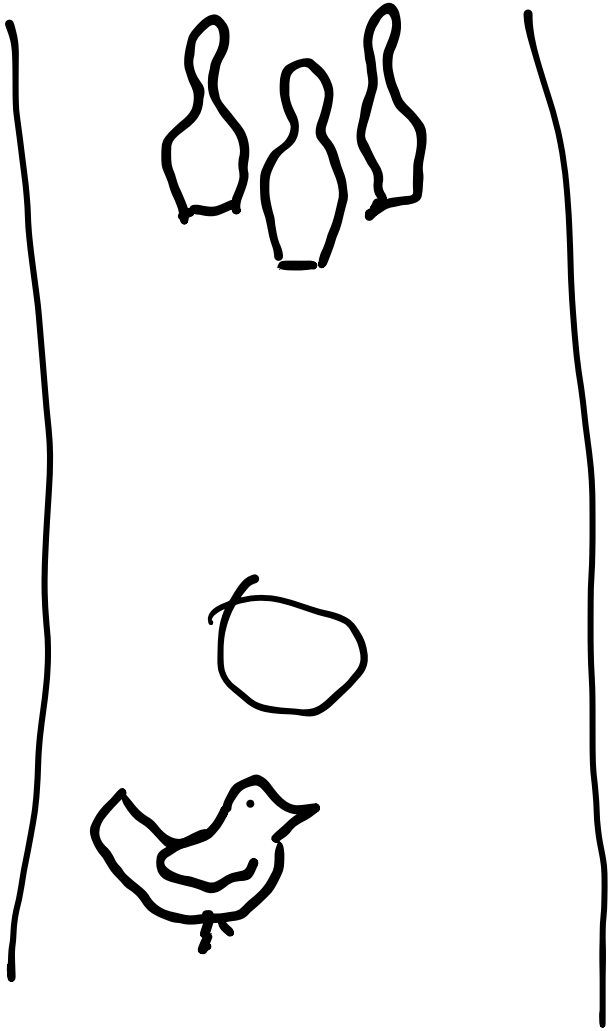
Reward Shaping

Reward Shaping

B. F. Skinner

Pigeon-guided bombs, 1943

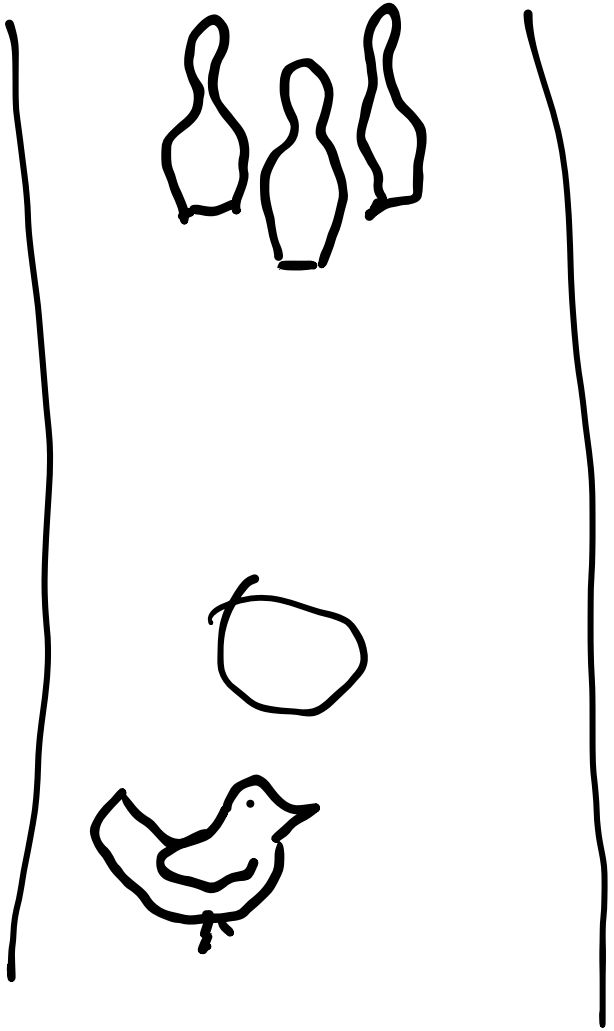
Reward Shaping



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

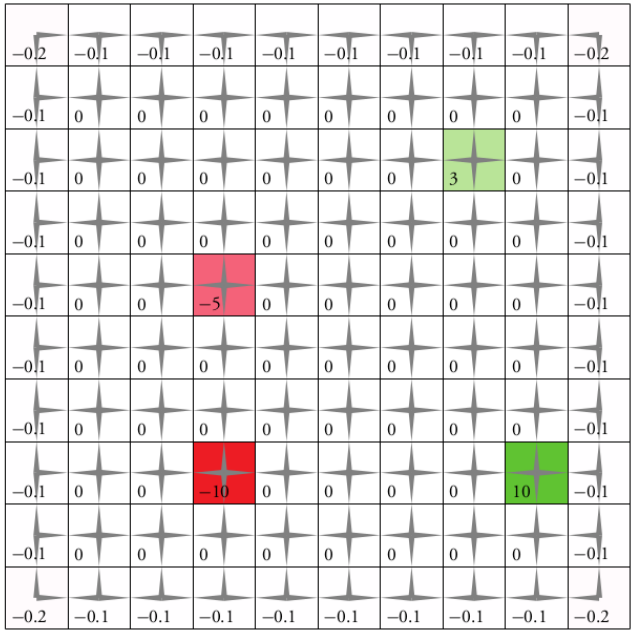
Reward Shaping

"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

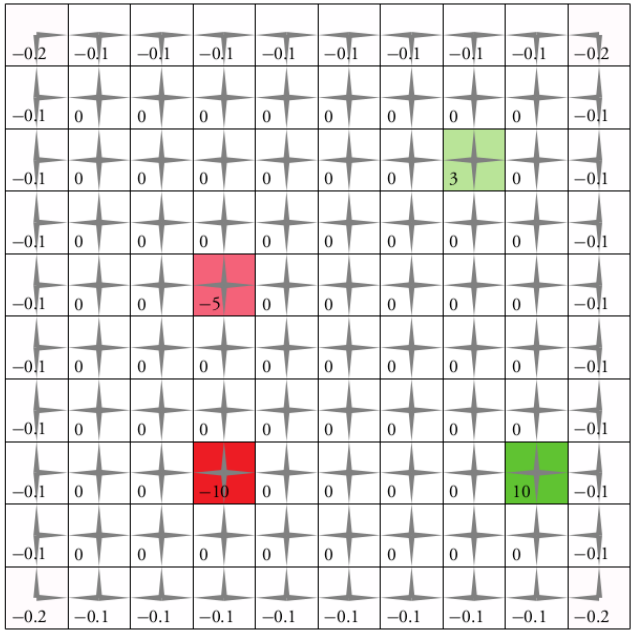
Reward



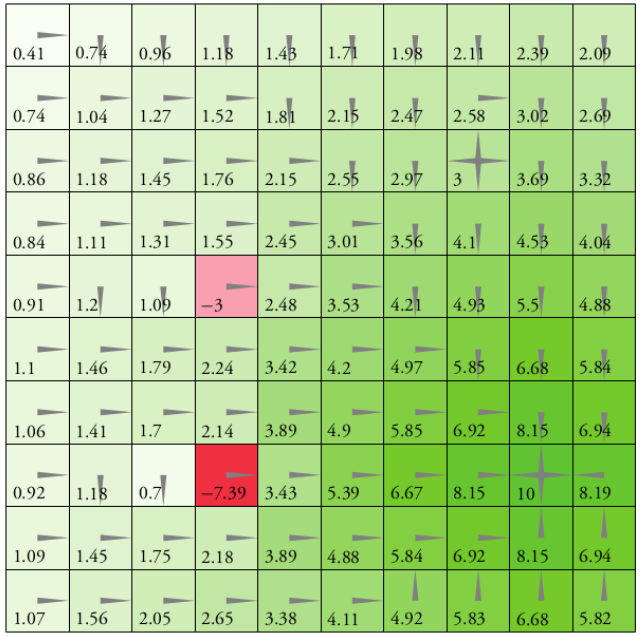
Reward Shaping

"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



Reward Shaping

- $R(s, a, s') += F(s) - \gamma F(s')$
- any other transformation may yield sub optimal policies unless further assumptions are made about the underlying MDP

What can we do?

What can we do?

- Transparency (this is hard because it opens you up to criticism)
  - IEEE P7001

What can we do?

- Transparency (this is hard because it opens you up to criticism)
  - IEEE P7001
- Understand the problem, especially what you don't know
  - What uncertainties can you quantify?
  - What problems are likely to arise?
  - Keep formulations as simple as possible - do not use band-aid fixes
  - Test often

What can we do?

- Transparency (this is hard because it opens you up to criticism)
  - IEEE P7001
- Understand the problem, especially what you don't know
  - What uncertainties can you quantify?
  - What problems are likely to arise?
  - Keep formulations as simple as possible - do not use band-aid fixes
  - Test often

Emerging best practices (AIMA)

- Software engineers talk to social scientists and domain experts
- Foster diverse pool of software engineers representative of society
- Define what groups your system will support (language, age, abilities)
- Objective function incorporating fairness
- Examine data for prejudice and for correlation with protected attributes
- Understand human annotation process, verify annotation accuracy
- Track metrics that for vulnerable subgroups
- Include system tests that reflect experience of vulnerable users
- Have a feedback loop so that problems are dealt with

Long-Term Problems

Superintelligence

- Eventually (perhaps very soon), we will most likely create AI systems that are more intelligent than humans according to some metric
- Is this a good thing?

Good

- Solve really hard problems
  - Cure diseases
  - Eliminate Accidents
- Transhumanism

Bad

- No way to check if solution is good
- Is it ethical to create a superintelligence
- Supplant humanity

Thought Experiment: Paperclip Maximizer

(Bostrum, 2003)

- Too many paperclips
- All of earth's resources to produce paperclips
- Shutoff switch

What can we do about it?

What can we do about it?

- **Asimov's laws**
  - A robot may not injure a human being or, through inaction, allow a human being to come to harm. ← what is 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.

What can we do about it?

- **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.

Experience with other superintelligent entities

What can we do about it?

- **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.

Experience with other superintelligent entities

- NASA/Spacex

What can we do about it?

- **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.

Experience with other superintelligent entities

- NASA/Spacex
- Other corporations

What can we do about it?

- **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.

Experience with other superintelligent entities

- NASA/Spacex
- Other corporations
- Countries (liberal democracy recognizes human limitations with freedom of speech)