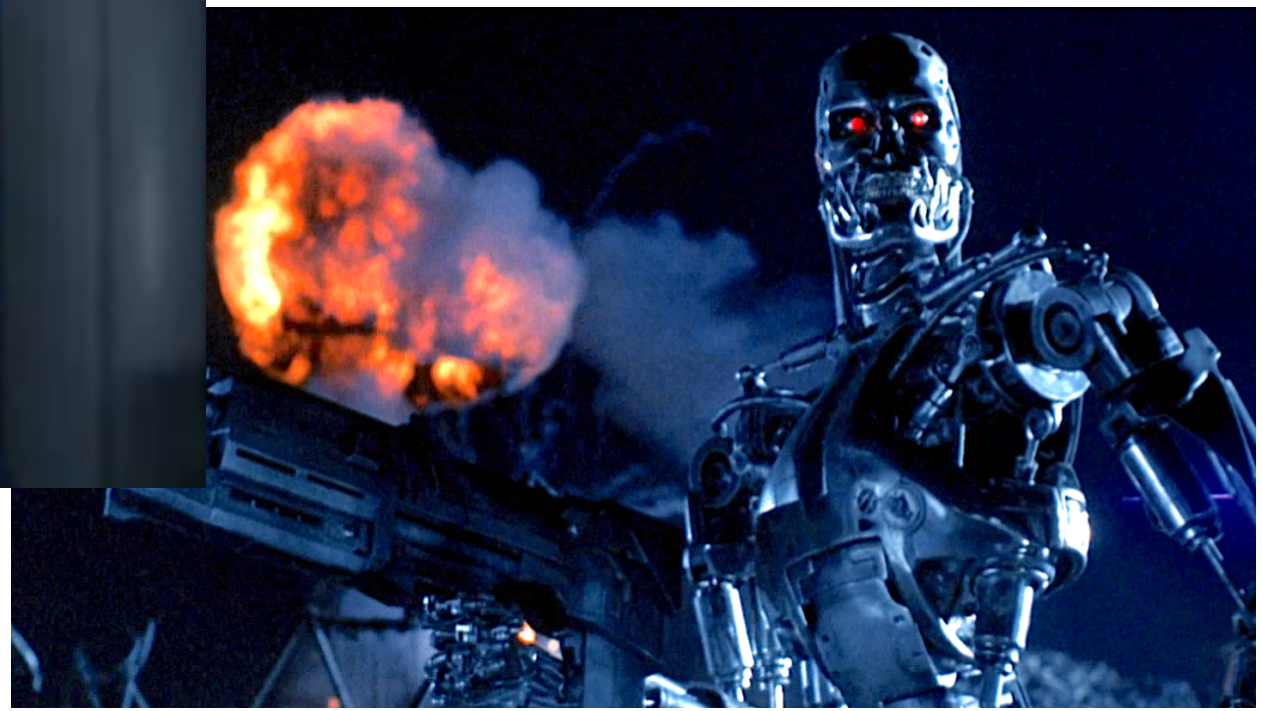
Ethics: The Alignment Problem

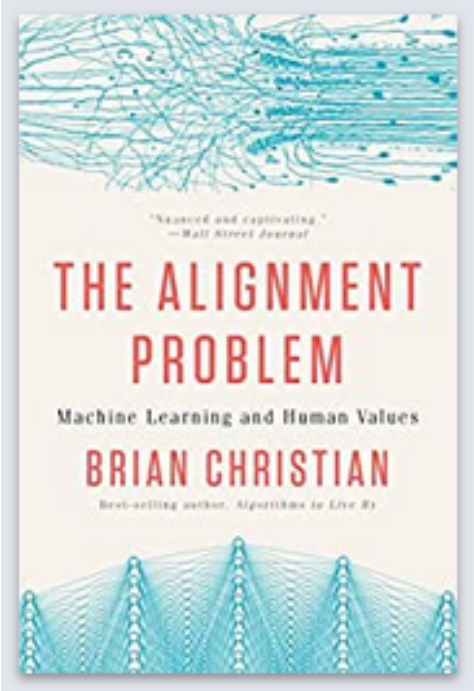
How do we harness artificial intelligence for the good of humanity?

The problem we think about: Skynet

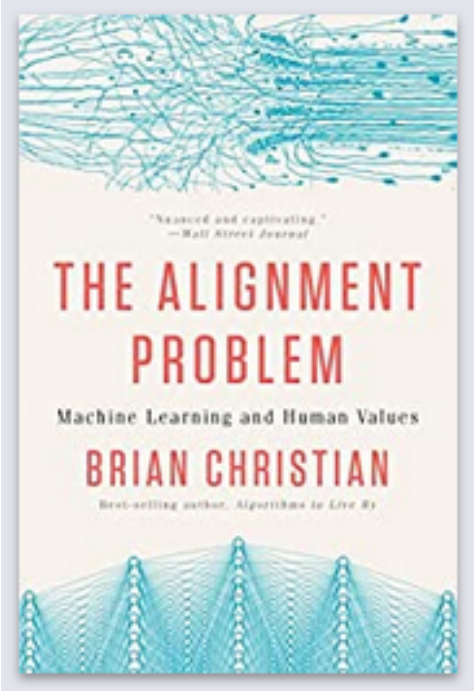


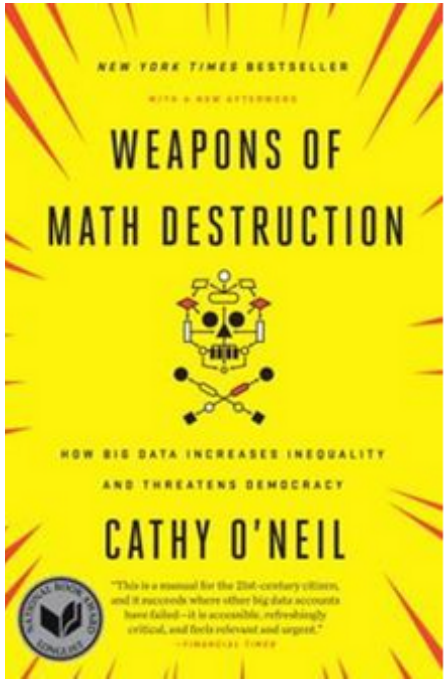
Some Problems Are Already Here

Some Problems Are Already Here

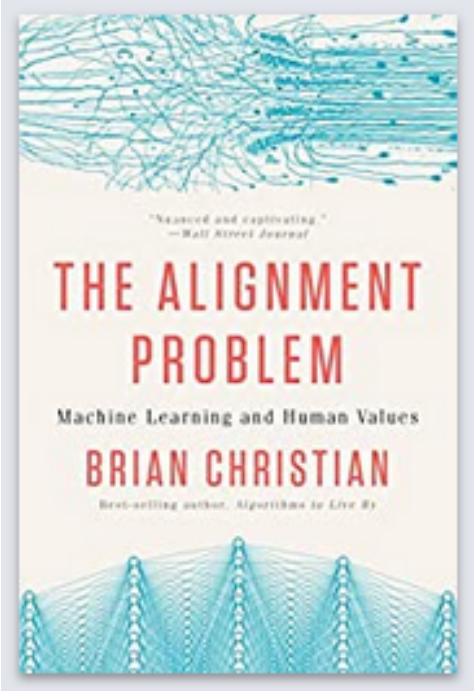


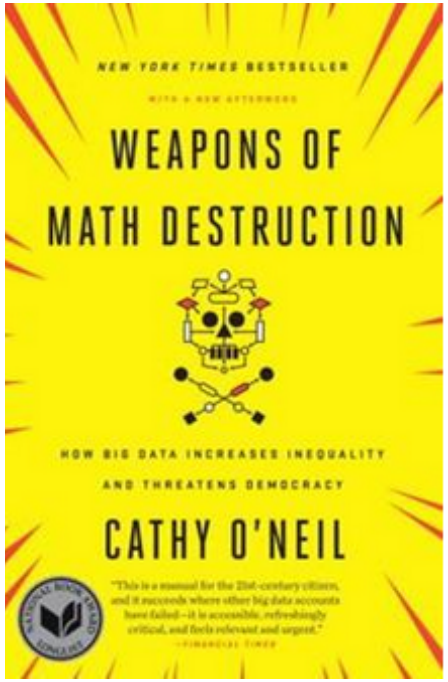
Some Problems Are Already Here

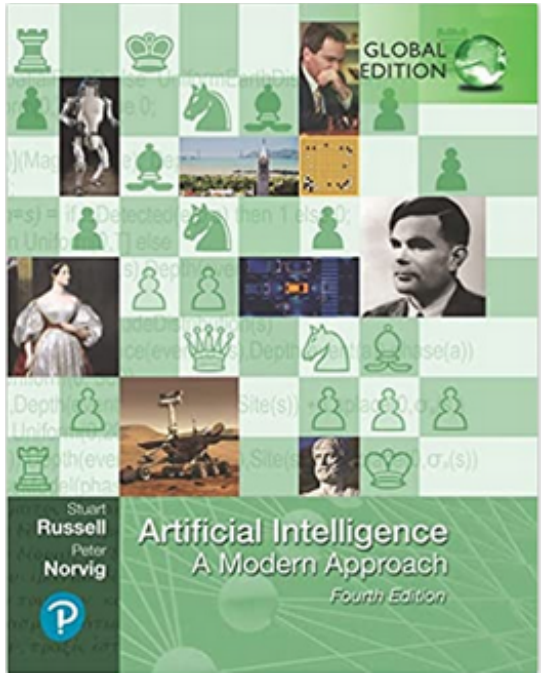




Some Problems Are Already Here







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

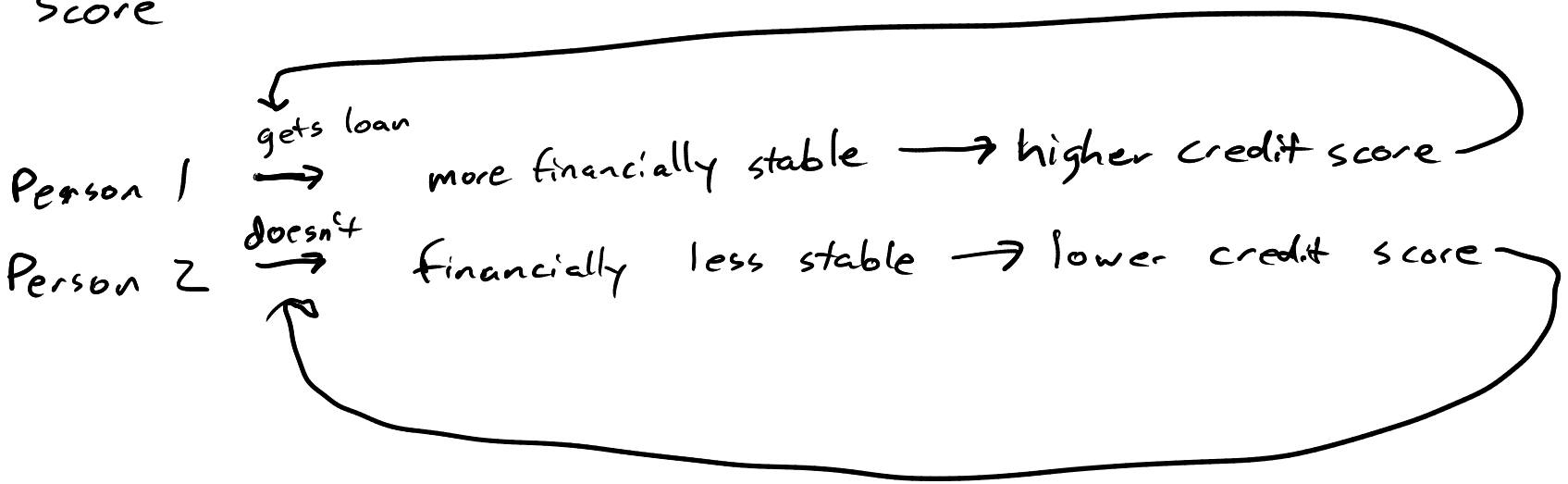
Suggested possible solution in AIMA:

"Equal Impact": assigning utility

Immediate Problem: Decision Feedback Loops

Lender: Lend to people who have highest prob. making payments on time

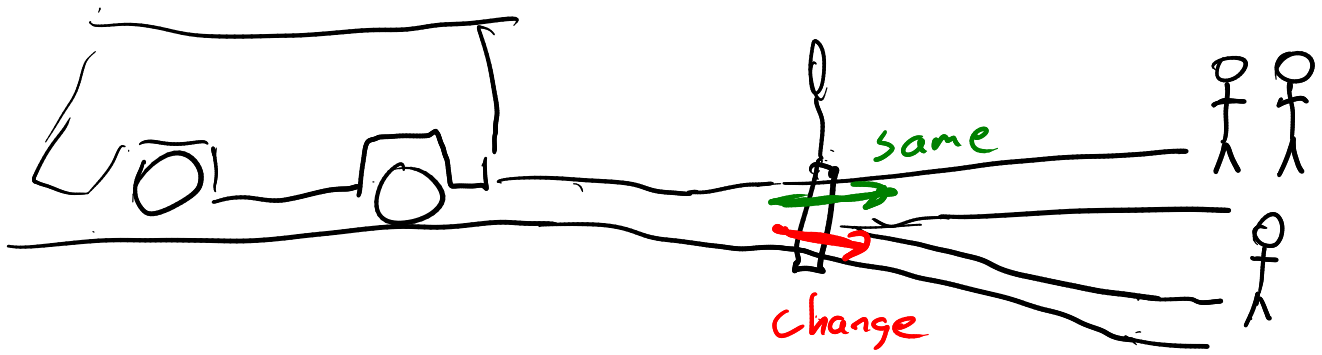
Credit Score



Immediate Problem: Employment

- Bank Tellers / ATM
  - 1900: >40% in agriculture 2000: <2%
    - Differences:
      - AI can do interesting things
      - Pace of Change
        - 100 years vs 10 years
      - Zero-marginal cost of replication
        - 10% better farmer → 20% more income
        - 10% better AI engineer → 1000% more income ←
      - Less cost for adoption
- Employment
  1. production of goods ← X
  2. income ←
  3. sense of purpose, accomplishment, social integration ←

Values: Trolley Problems



Transparency / Trust
- Releasing system specification
- Automated explanation

IEEE P7001

Standard for transparency in autonomous systems

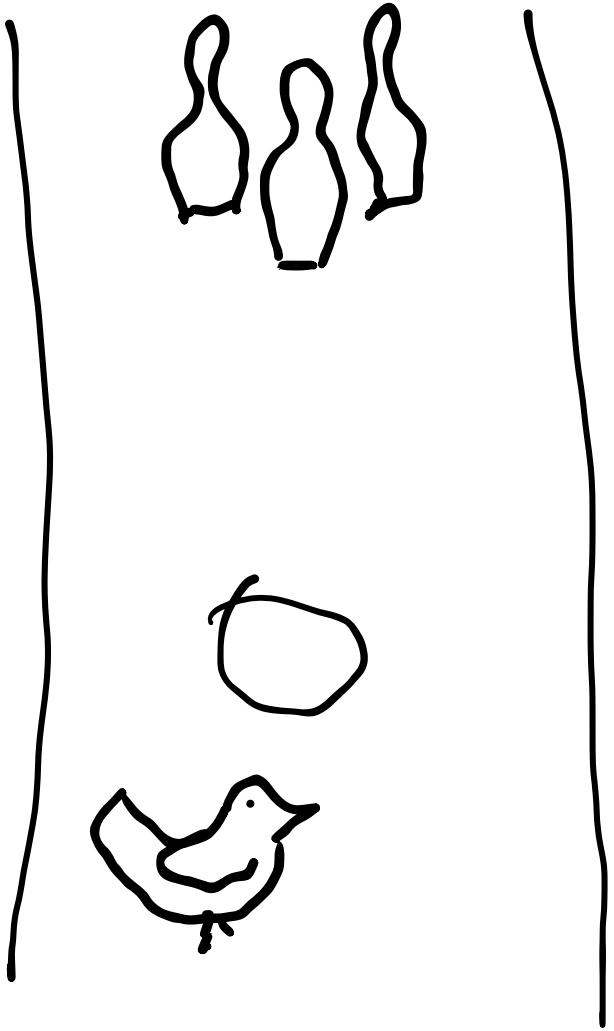
Reward Shaping

Reward Shaping

B. F. Skinner

Pigeon-guided bombs, 1943

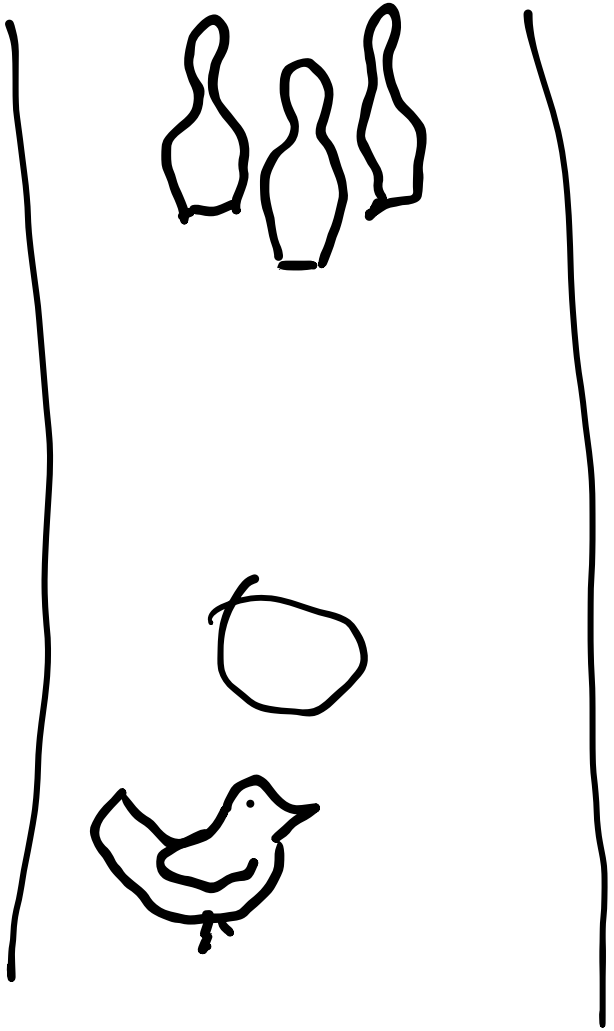
Reward Shaping



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/watch?v=tlOIHko8ySg

Reward Shaping

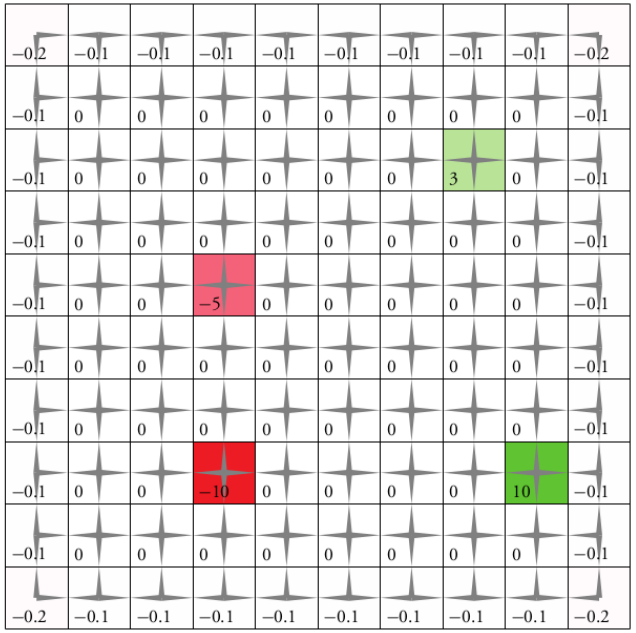
"As a general rule, it is better to design performance measures according to what one actually wants in the environment, rather than according to how one thinks the agent should behave." - Stuart Russell

.

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

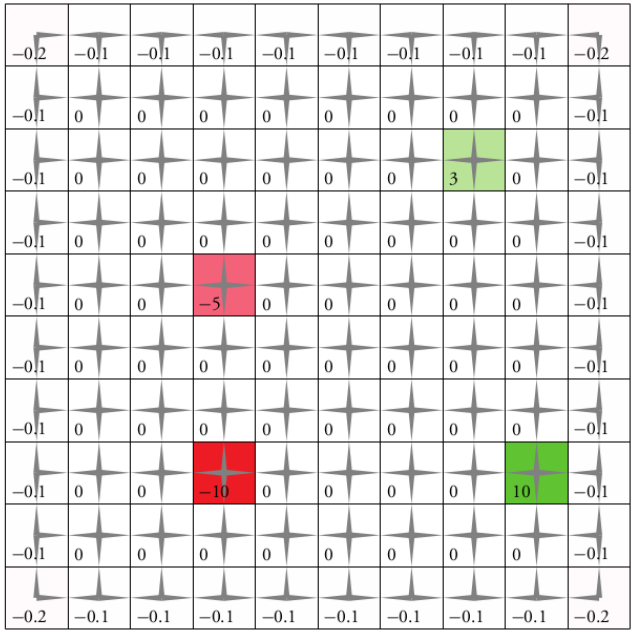
Reward



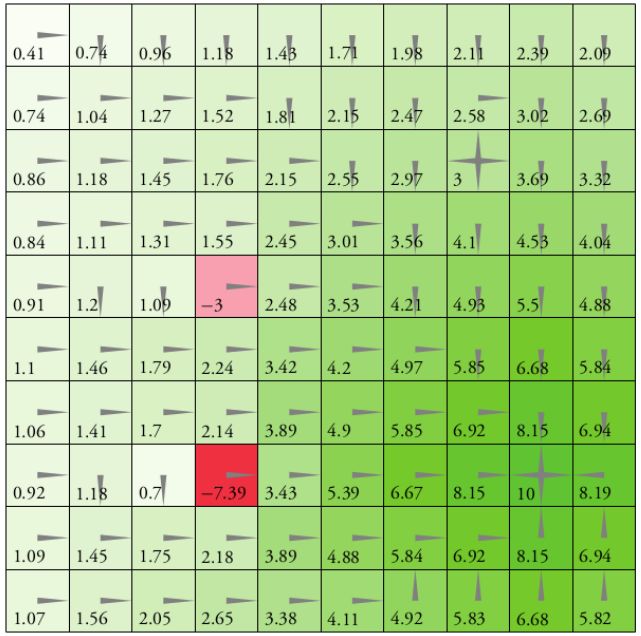
Reward Shaping

"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

Potential-Based Reward Shaping

$F(s') - \gamma F(s'')$

- $R(s, a, s') += F(s) - \gamma F(s')$ ← will not change optimal policies
- any other transformation may yield sub optimal policies unless further assumptions are made about the underlying MDP

Long-Term Problems

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.

Gorilla Problem: Super-Human Intelligence