Ethics: The Alignment Problem

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

Ethics: The Alignment Problem

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

Disclaimer: I am not an ethics expert.

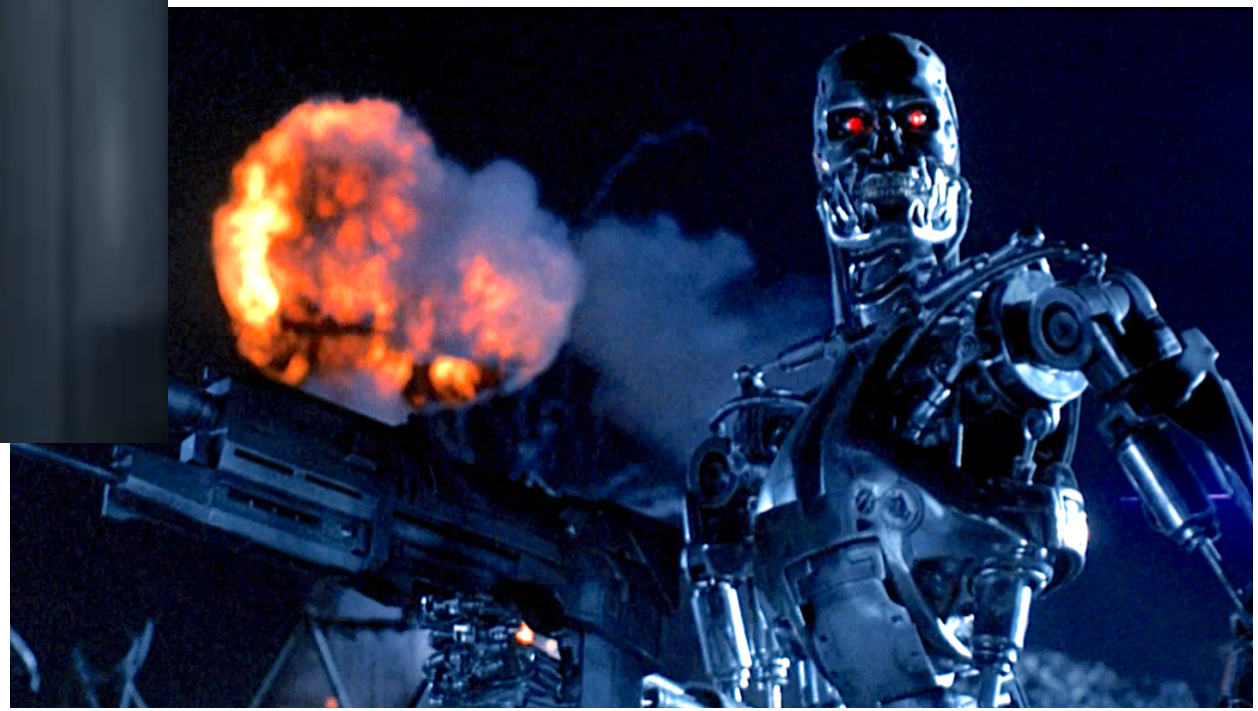
Ethics: The Alignment Problem

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

Disclaimer: I am not an ethics expert.

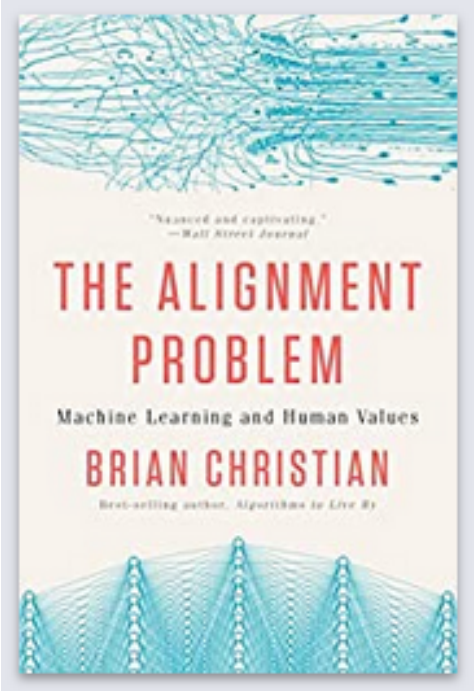
"If a thing is worth doing, it is worth doing badly."
- G.K. Chesterton

The problem we tend to think about: Skynet

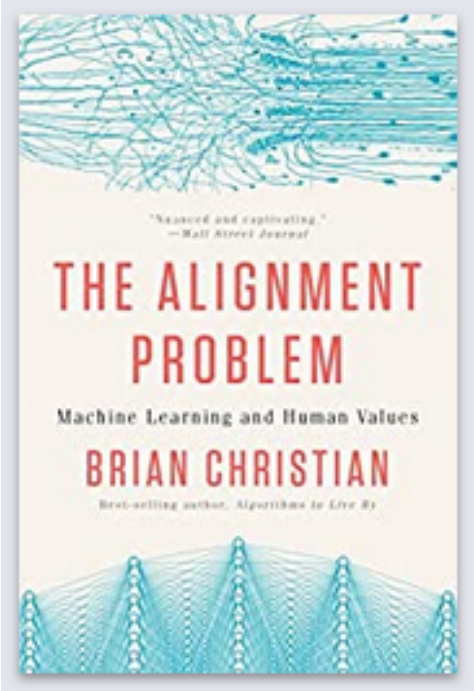


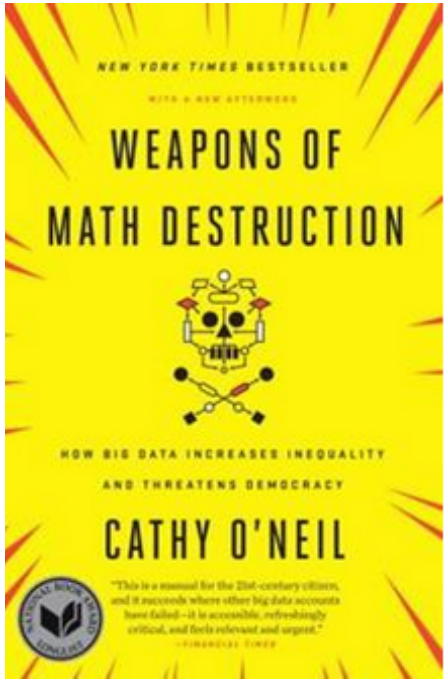
Some Problems Are Already Here

Some Problems Are Already Here

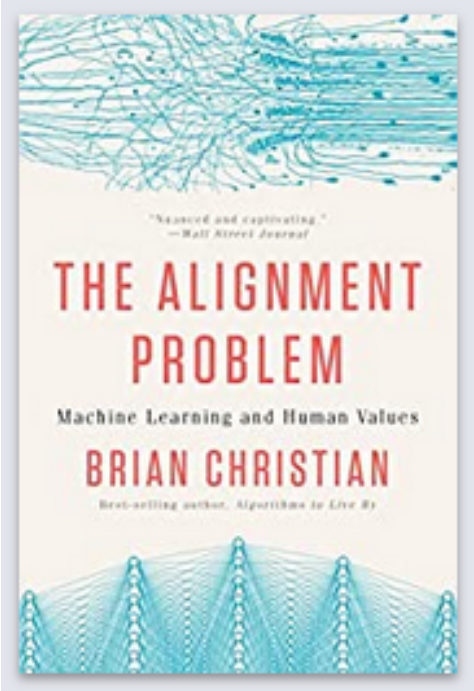


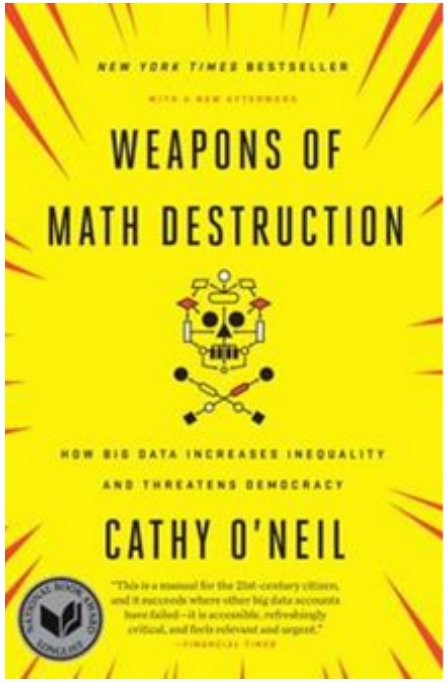
Some Problems Are Already Here

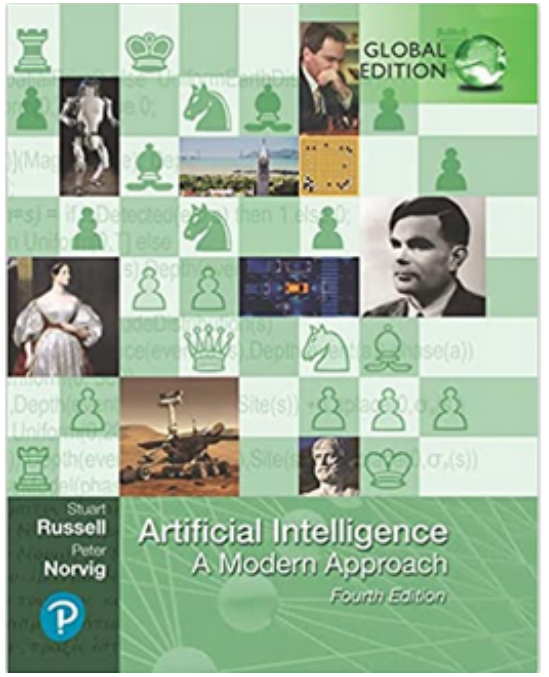




Some Problems Are Already Here







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

Sweeny, "Simple demographics often identify people uniquely"

Immediate Problem: Difficulty removing information from Data

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

Sweeny, "Simple demographics often identify people uniquely"

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

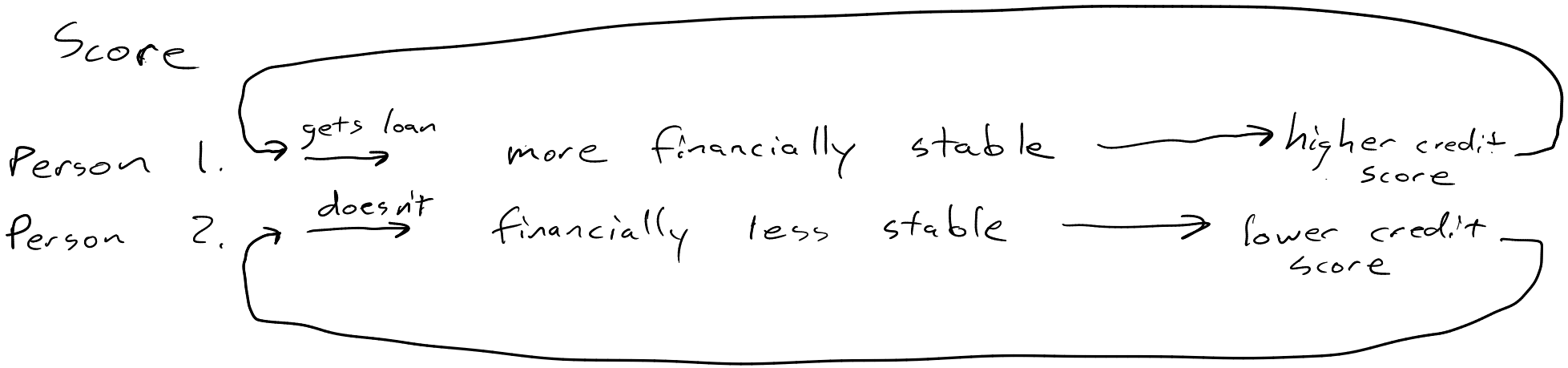
Suggested possible solution in AIMA:

"Equal Impact": assigning utility

Immediate Problem: Decision Feedback Loops

→ Lender: Lend to people who have the highest probability of making payments on time

→ Credit Score



Immediate Problem: Employment

- Bank Tellers / ATMs
- 1900 >40% people in agriculture 2000 <2%

- Differences
  - More General
  - Pace of change
    - 100 years vs 10 years vs 1 year
  - AI can do more desirable jobs
  - Zero marginal cost of replication
    - 10% better farmer → 20% more income
    - 10% better AI engineer → 1000% more income
  - Less cost for adoption

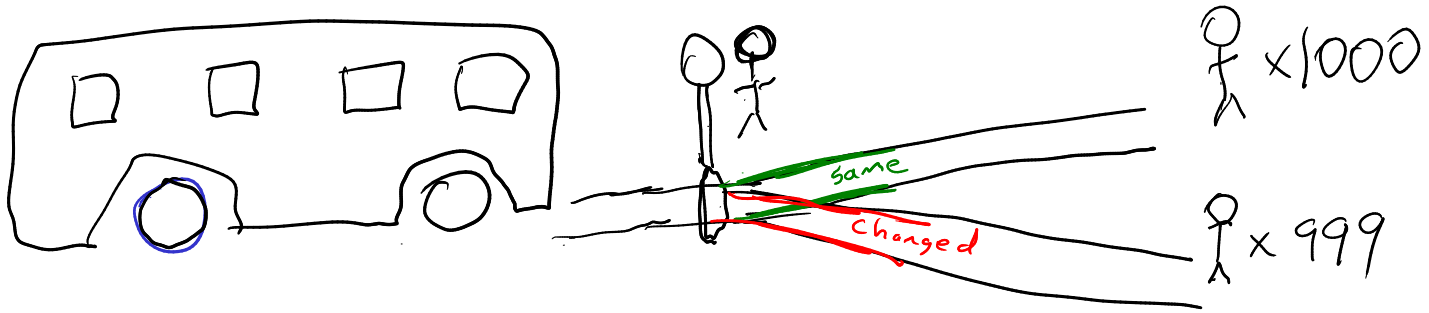
- Employment
  1. production of goods
  2. income
  3. sense of purpose, accomplishment, social integration

Computers are better than humans at well-defined mathematical optimization

ALPHAGO

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

Values: Trolley Problems



Transparency / Trust
- Release system specification
- Automated Explanation

IEEE P7001
- Standard for transparency in autonomous systems

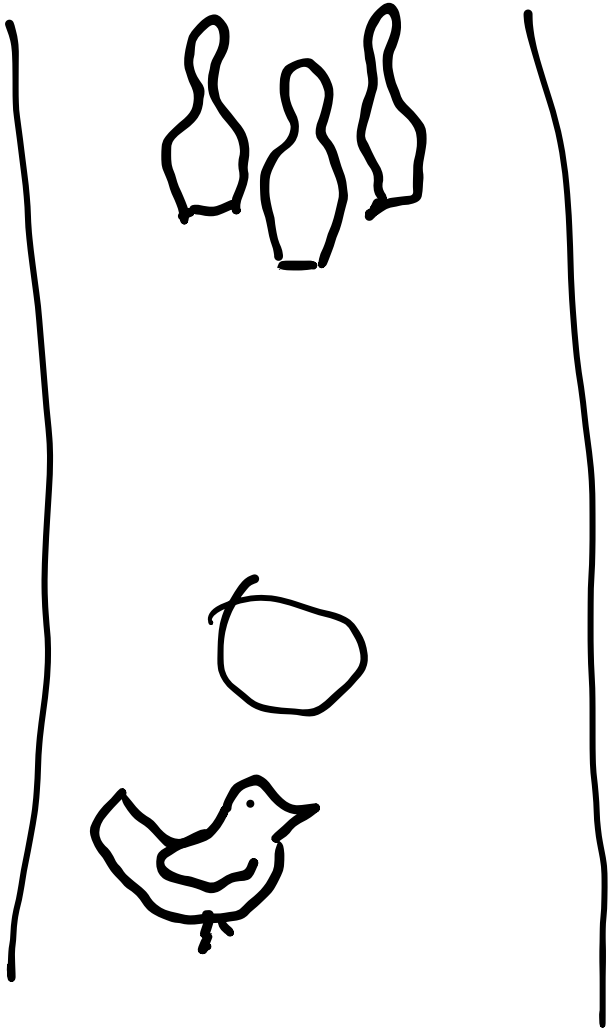
Reward Shaping

Reward Shaping

B. F. Skinner

Pigeon-guided bombs, 1943

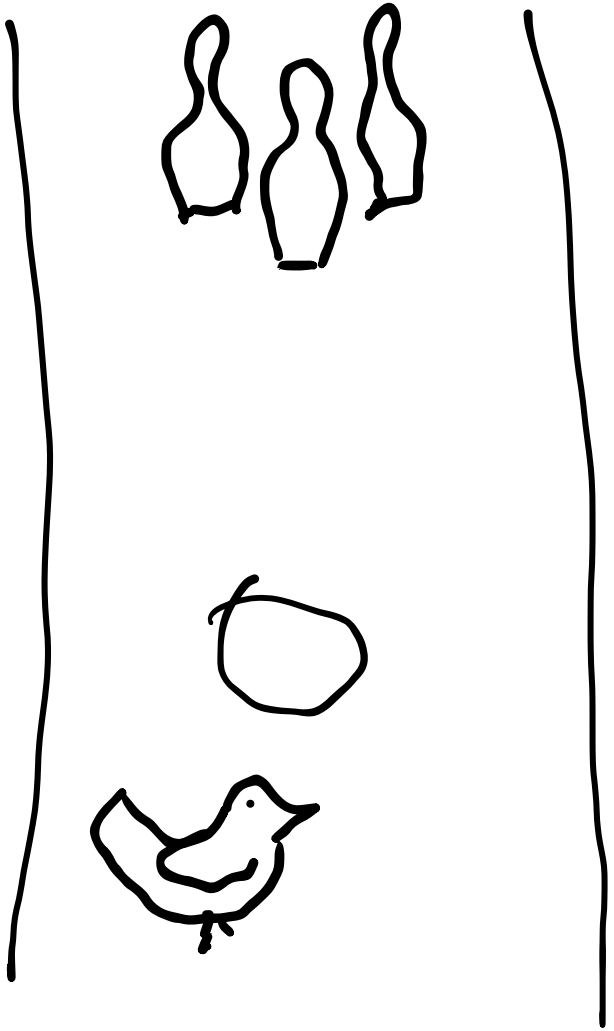
Reward Shaping



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/embed/tlOIHko8ySg?enablejsapi=1

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

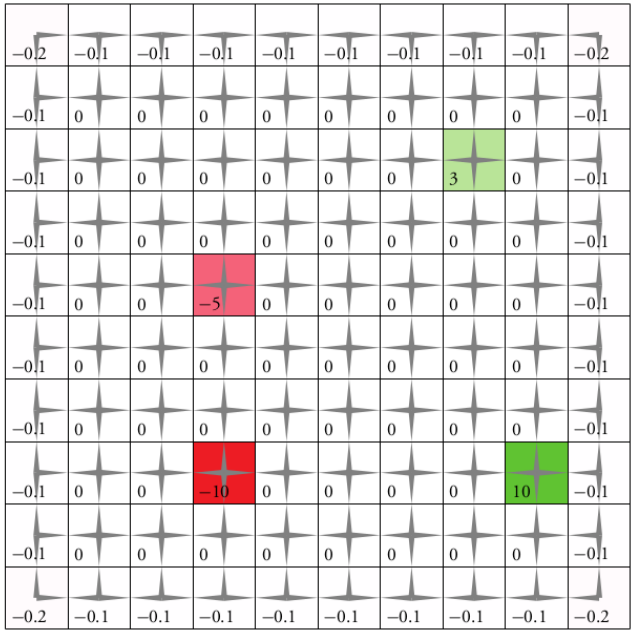
Reward Shaping

"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

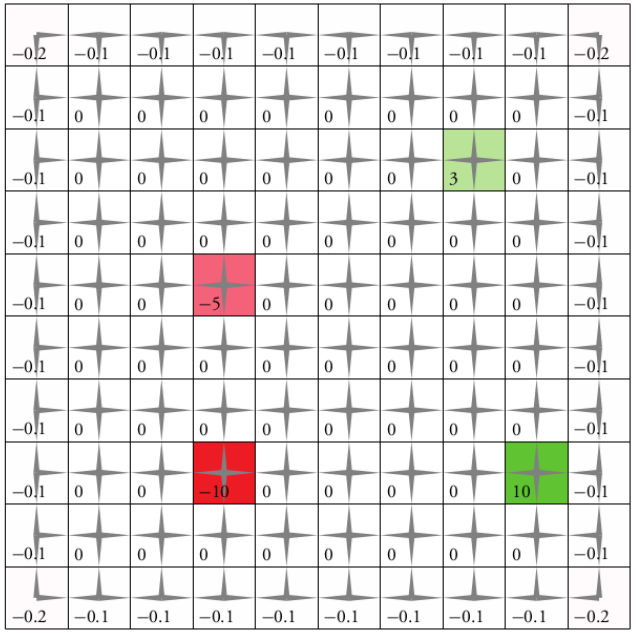
Reward



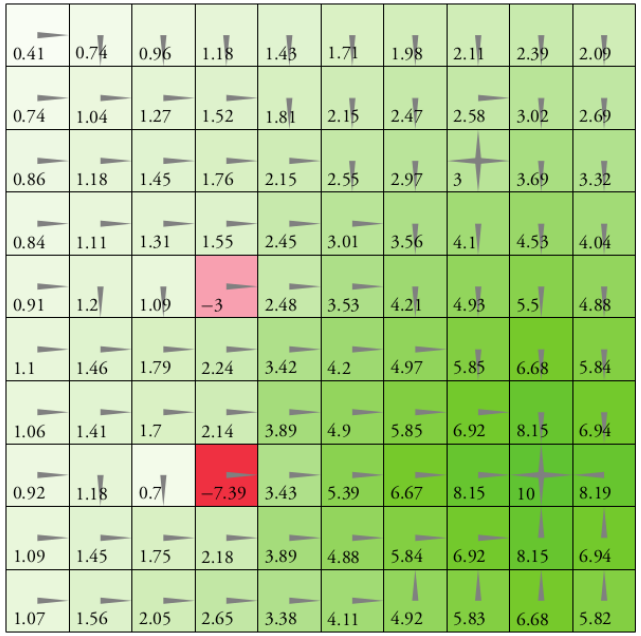
Reward Shaping

"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?

Emerging best practices (AIMA)

What can we do?

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

Emerging best practices (AIMA)

What can we do?

Emerging best practices (AIMA)

- 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

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

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)

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

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

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

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

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

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
  - Prevent accidents

Bad

- Misuse of power
- Alignment problem
- No way to check a-priori if solution is good or bad

Objectiveness

- Is it ethical to create a superintelligence
- Supplant Humanity

- Transhumanism

Thought Experiment: Paperclip Maximizer

(Bostrum, 2003)

- Take all resources to make paper clips
- Too many paperclips

- Shutoff switch

- Constrain inputs

The Thermodynamic Objection

Marc Andreeson: By the way, there's a very practical objection to all this, which is kind of sometimes called the thermodynamic objection, which, again, sort of connects this back to reality, which is: Look, we're sitting here today and let's say that GPT develops whatever you want to call it--a mind of it might of its own or its own goals or whatever. Like, it can't get chips. Right? So, now it has its evil plan to take over the world. It needs, like, more chips to be able to run its evil plan. NVIDIA is out of chips. And so, what--

Russ Roberts: They have a story for that. They explain: they'll get some poor low-IQ person--not you or me, Marc, because we're too smart--but they'll get a low-IQ person, an employee of some lower level, and they'll convince him to go buy chips for them.

Marc Andreessen: No, no. But, the chips literally don't exist. Like NVIDIA can't make the chips. There's chip shortages all throughout the AI ecosystem.

Russ Roberts: Oh. Well, they'll fix that. That's easy.

Marc Andreessen: Exactly. So, basically--

Russ Roberts: They'll get Senators, the Congress people to vote for subsidies to things that the chips need and then in a week or two, that'll go away.

Marc Andreessen: So, this is what's called the thermodynamic objection, which is: Okay, you're the AI, you're the sentient artificial intelligence. To accomplish your evil planting over the world, you need the chips, you need the electricity, you need to go buy the votes in Congress, you need to do this, you need to do all of these things.

And, that somehow these things are going to happen basically overnight, very quickly, very easily without putting--at this point, neither one of us are steel manning, by the way--but without putting a footprint into the world. Right? And this is this sort of takeoff idea, and this all happens in 24 hours.

It's like--I don't know about you, but anybody who's ever tried to get Congress people to do anything, it doesn't happen like that. Once you enter the real world of politics to get a bill passed--

https://www.econtalk.org/marc-andreessen-on-why-ai-will-save-the-world/

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.
  - 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)

Values: Trolley Problems

https://www.nature.com/articles/s41893-022-00851-6

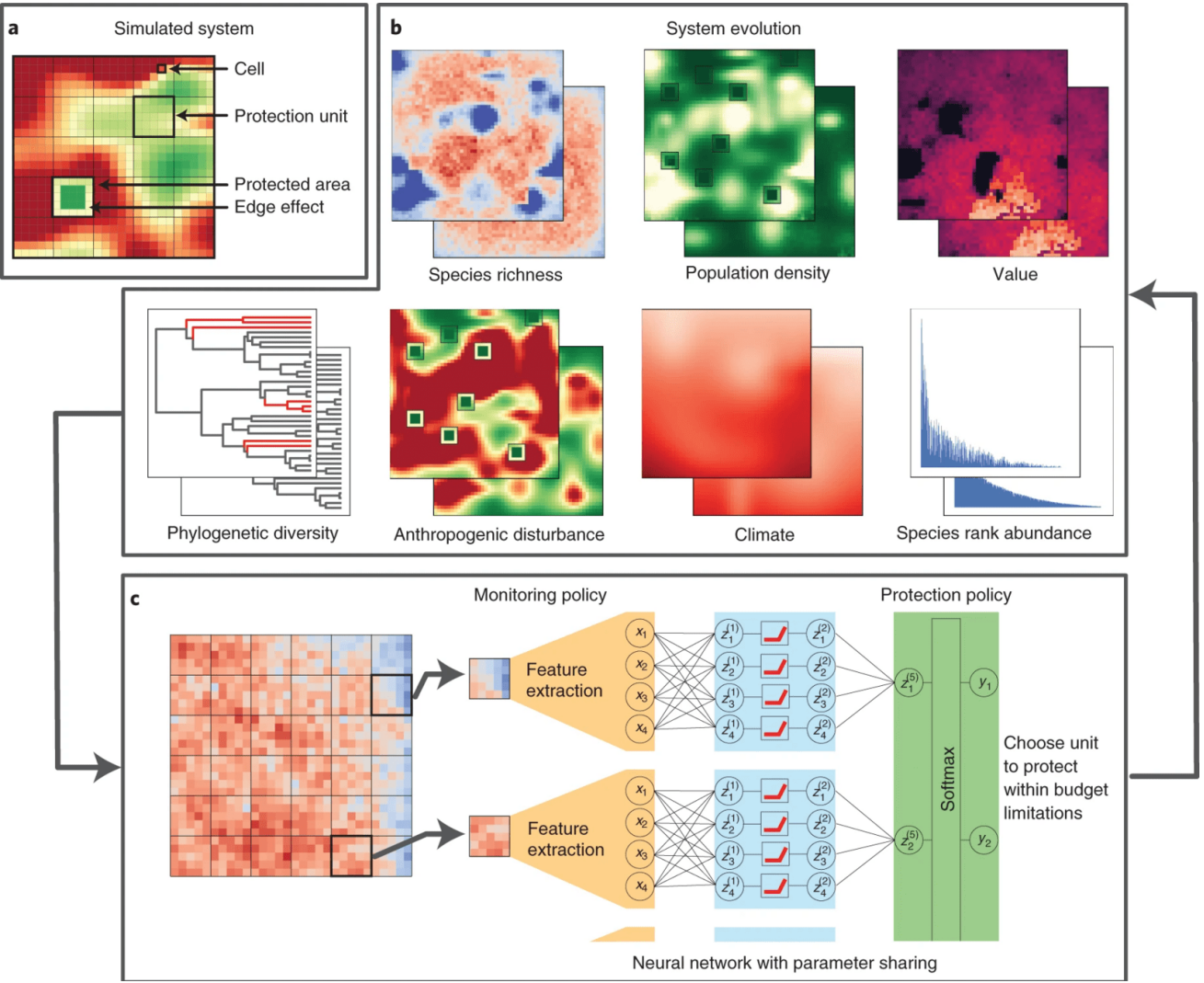
Values: Trolley Problems

The CAPTAIN RL framework

https://www.nature.com/articles/s41893-022-00851-6

Values: Trolley Problems

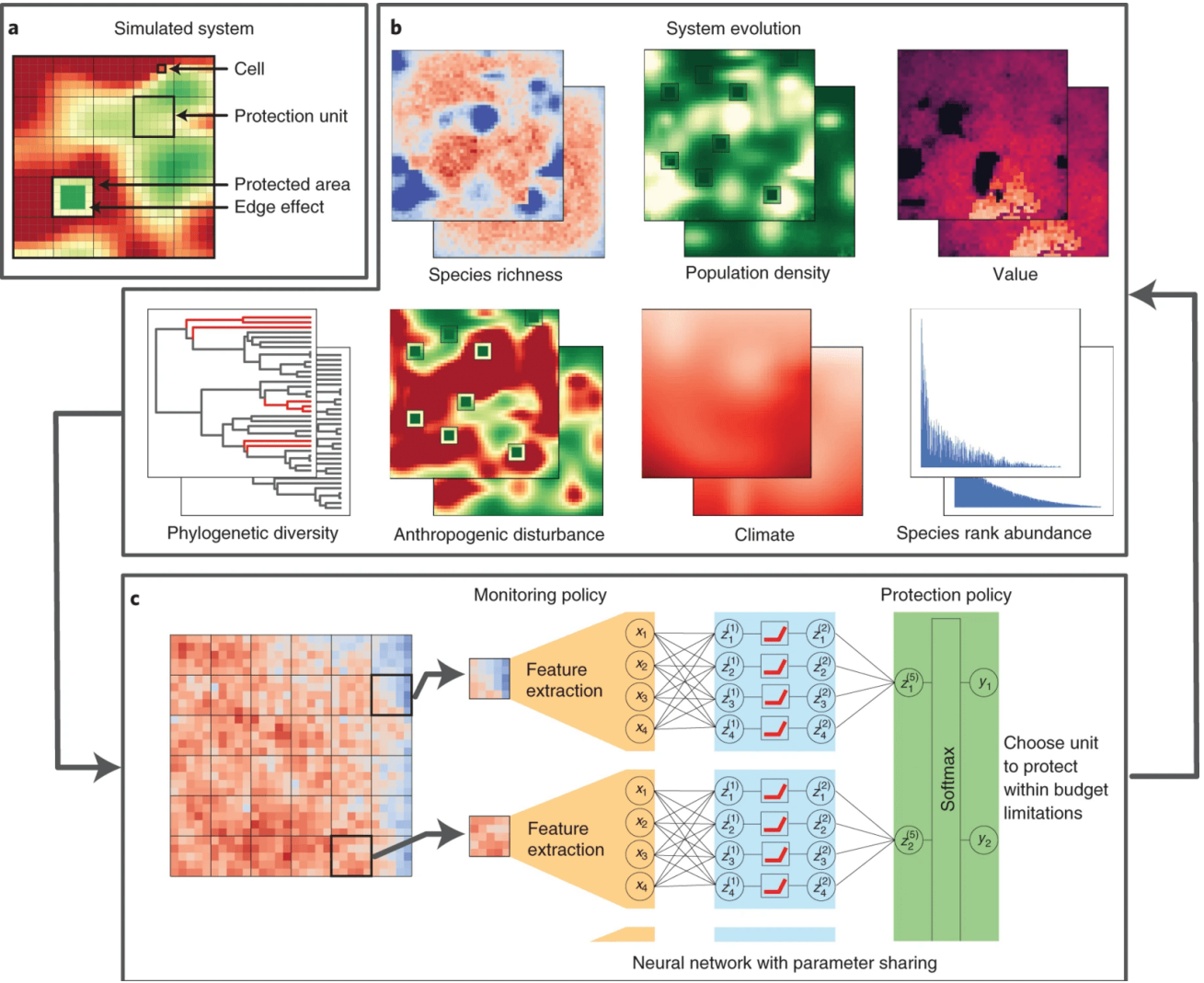
The CAPTAIN RL framework

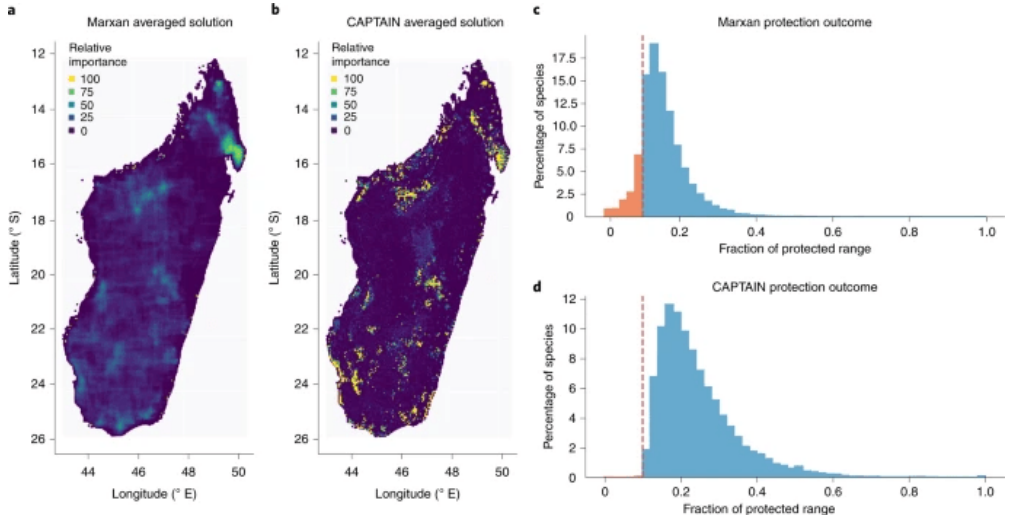


https://www.nature.com/articles/s41893-022-00851-6

Values: Trolley Problems

The CAPTAIN RL framework

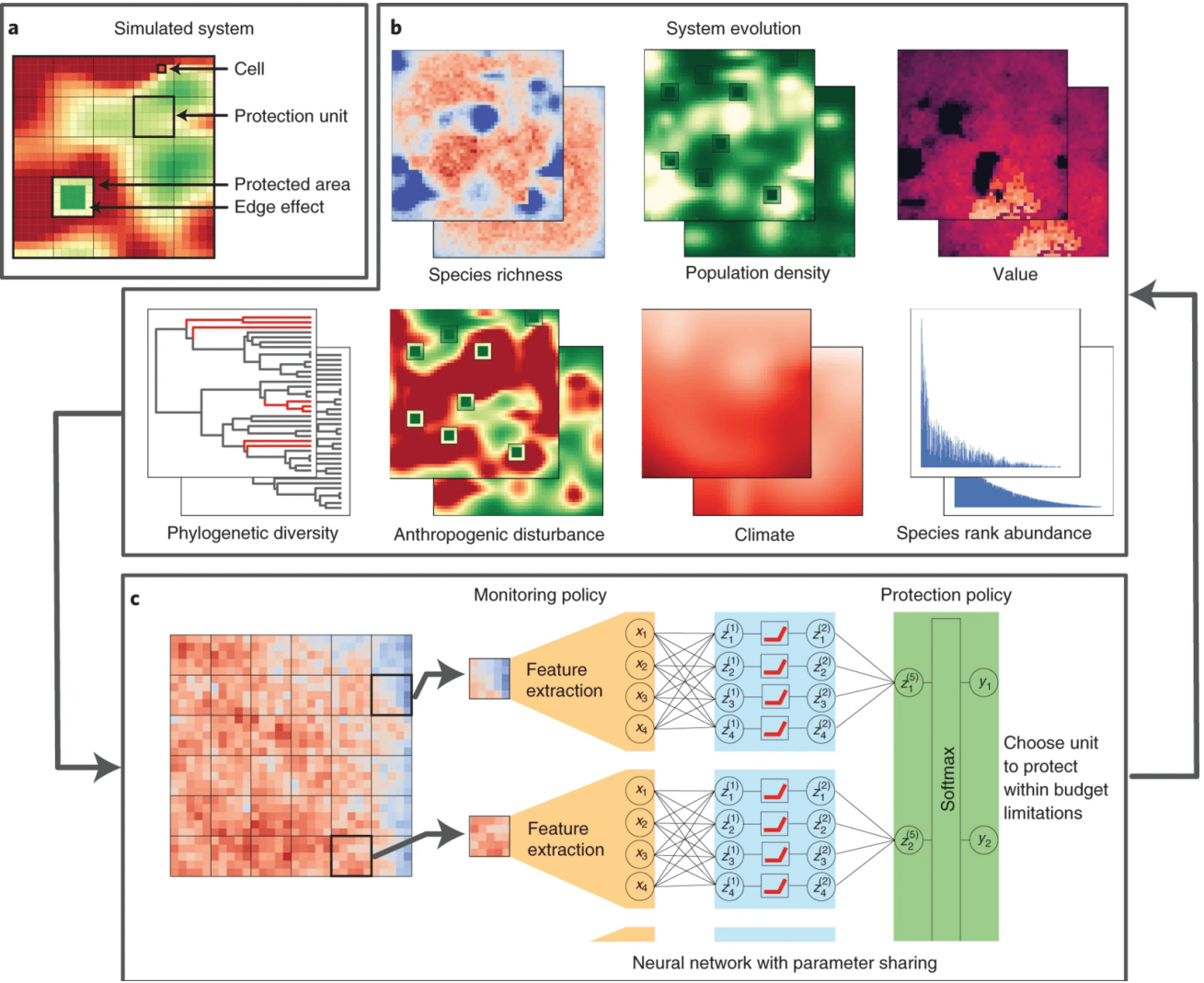


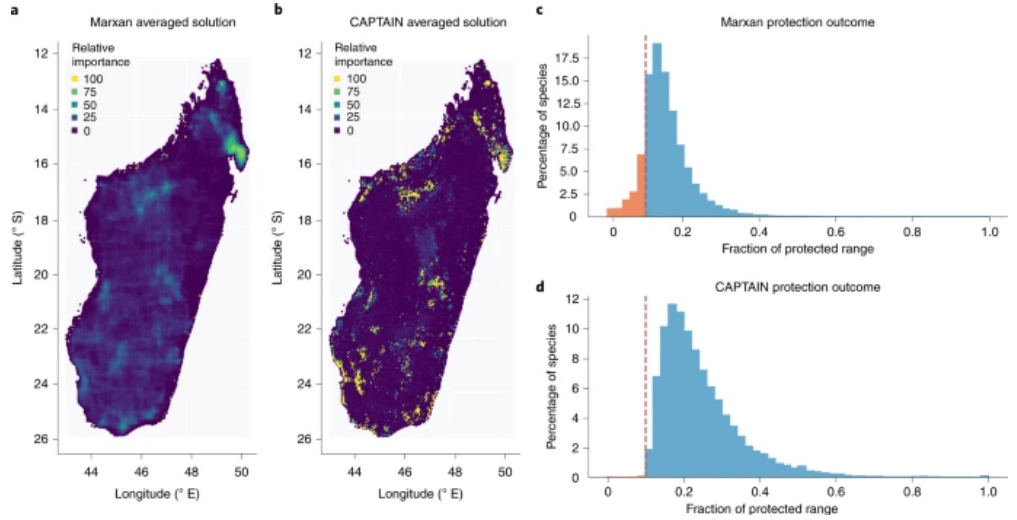


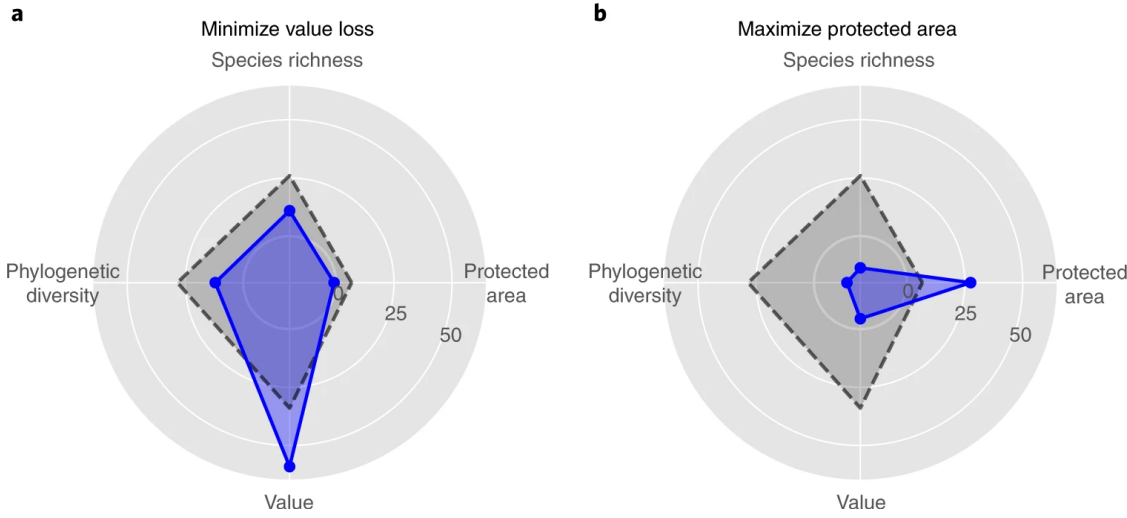
https://www.nature.com/articles/s41893-022-00851-6

Values: Trolley Problems

The CAPTAIN RL framework



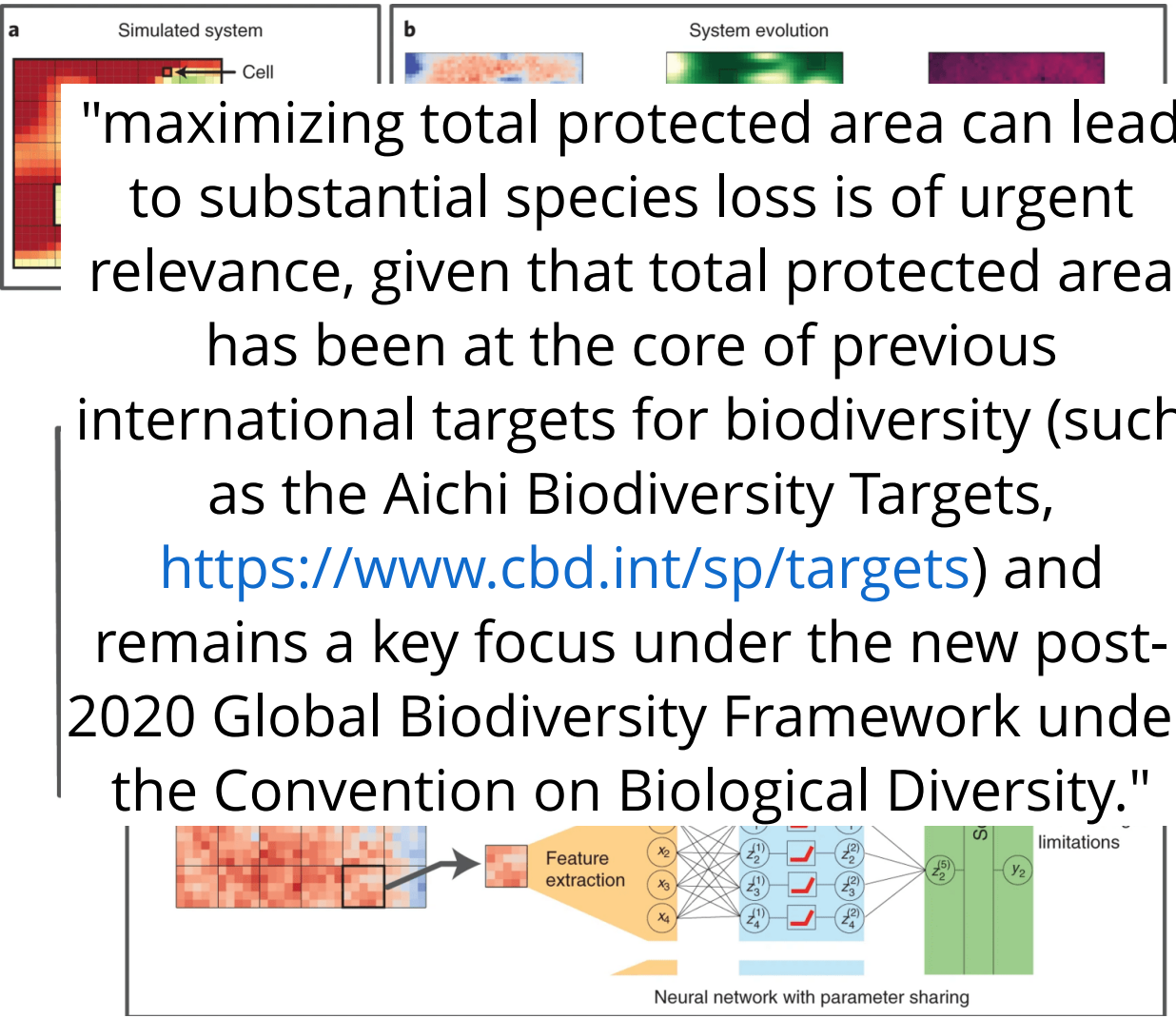




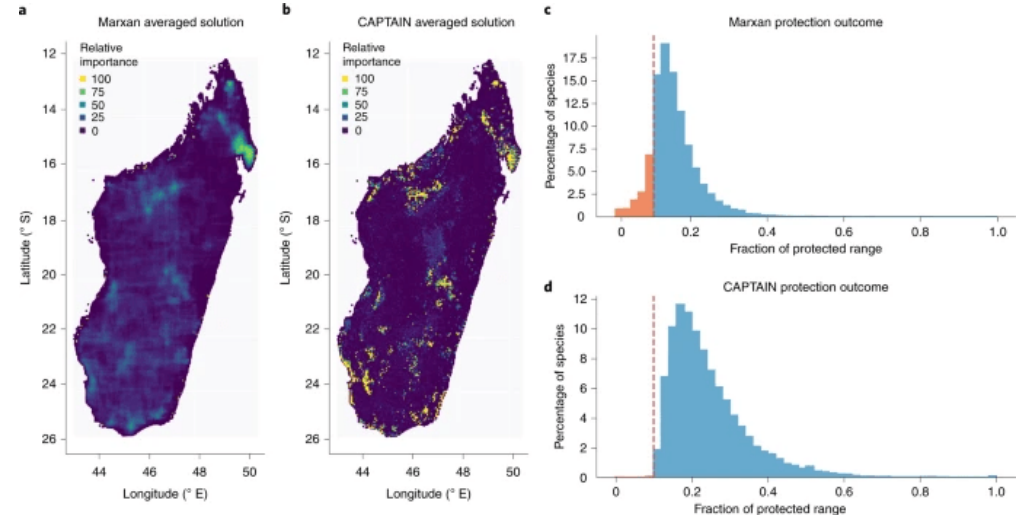
https://www.nature.com/articles/s41893-022-00851-6

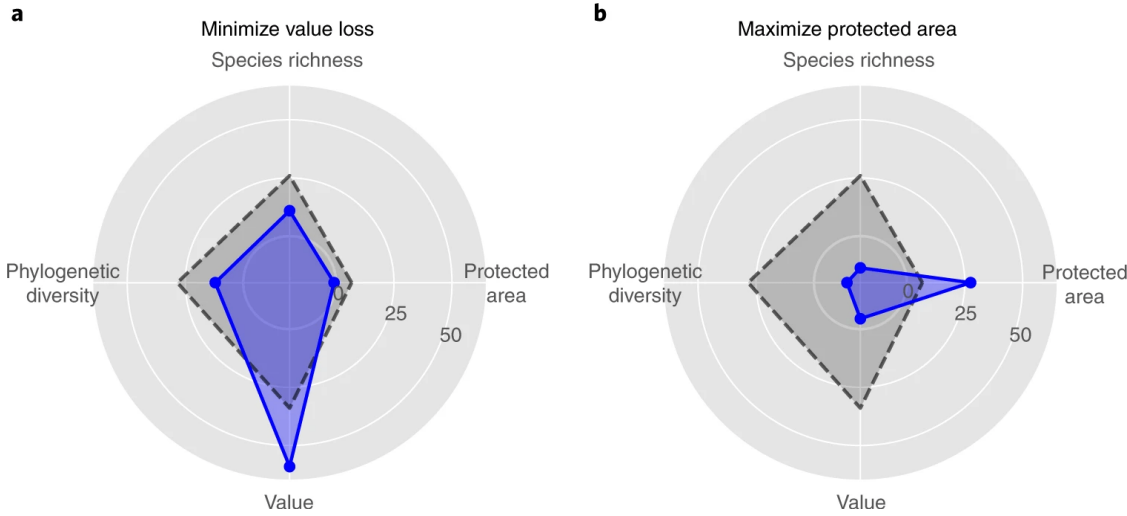
Values: Trolley Problems

The CAPTAIN RL framework



"maximizing total protected area can lead to substantial species loss is of urgent relevance, given that total protected area has been at the core of previous international targets for biodiversity (such as the Aichi Biodiversity Targets, https://www.cbd.int/sp/targets) and remains a key focus under the new post-2020 Global Biodiversity Framework under the Convention on Biological Diversity."





https://www.nature.com/articles/s41893-022-00851-6

What should we do about it?

What should we do about it?

- ???

What should we do about it?

- ???
- Understand Uncertainty

What should we do about it?

- ???
- Understand Uncertainty
- Know when you don't know

What should we do about it?

- ???
- Understand Uncertainty
- Know when you don't know
- The future does not depend on technology as much as it depends on individual and collective morality (?): Foster norms