An Impact Measurement Manager Approach to AI Safety

Master's thesis in Mathematical Statistics, Statistical Learning and AI

Jonatan Hellgren

2022 - 12th of October

Department of Mathematical Sciences
Chalmers University of Technology
University of Gothenburg

Supervisor:	Olle Häggström, Department of Mathematical Sciences, Chalmers University of Technology
Examiner:	Torbjörn Lundh, Department of Mathematical Sciences, Chalmers University of Technology
Opponents:	Jens Ifver and Calvin Smith Universty of Gothenburg

Introduction

Aim of Thesis

Investigation

- AI safety

Simulation

Aim of Thesis

Investigation

- AI safety
- Low impact agents

Simulation

Aim of Thesis

Investigation

- AI safety
- Low impact agents
- Imapact measurements

Simulation

Aim of Thesis

Investigation

- AI safety
- Low impact agents
- Imapact measurements

Simulation

- Novel impact measurement

Aim of Thesis

Investigation

- AI safety
- Low impact agents
- Imapact measurements

Simulation

- Novel impact measurement
- Evaluate it

Overview of Presentation

- We will begin by looking at **AI safety**.

Overview of Presentation

- We will begin by looking at **AI safety**.
- Then go on with **impact measurements**.

Overview of Presentation

- We will begin by looking at **AI safety**.
- Then go on with **impact measurements**.
- After that, I will explain the **manager approach**.

Overview of Presentation

- We will begin by looking at **AI safety**.
- Then go on with **impact measurements**.
- After that, I will explain the **manager approach**.
- Finally, I will present the results.

AI Safety

Why should we worry?

- What is Artificial Intelligence (AI)?

Why should we worry?

- What is Artificial Intelligence (AI)?
- AI will likely become more intelligent then us.

Why should we worry?

- What is Artificial Intelligence (AI)?
- AI will likely become more intelligent then us.
- This can cause existential risks.

Existential risk

On futureoflife.org we find the following description, see FLI (nd):

Existential risk

On futureoflife.org we find the following description, see FLI (nd): An existential risk is any risk that has the potential to eliminate all of humanity or, at the very least, kill large swaths of the global population, leaving the survivors without sufficient means to rebuild society to current standards of living.

Future progress

To differentiate current AI from future versions, several terms are used:

Future progress

To differentiate current AI from future versions, several terms are used:

- Artificial General Intelligence (AGI):
  An AI that can solve an arbitrary set of tasks with as good or better performance then a human.

Future progress

To differentiate current AI from future versions, several terms are used:

- Artificial General Intelligence (AGI):
  An AI that can solve an arbitrary set of tasks with as good or better performance then a human.
- Transformative AI (TAI):
  An AI system capable enough to induce transformative consequences on the same scale as the industrial or agricultural revolution.

Future progress

To differentiate current AI from future versions, several terms are used:

- Artificial General Intelligence (AGI):
  An AI that can solve an arbitrary set of tasks with as good or better performance then a human.
- Transformative AI (TAI):
  An AI system capable enough to induce transformative consequences on the same scale as the industrial or agricultural revolution.
- Prepotent AI:
  An TAI that once deployed would be unstoppable.

Timeline for TAI Breakthrough

- Many predictions for this has been made.

Timeline for TAI Breakthrough

- Many predictions for this has been made.
- In Cotra (2020), the author presents a predicting model.

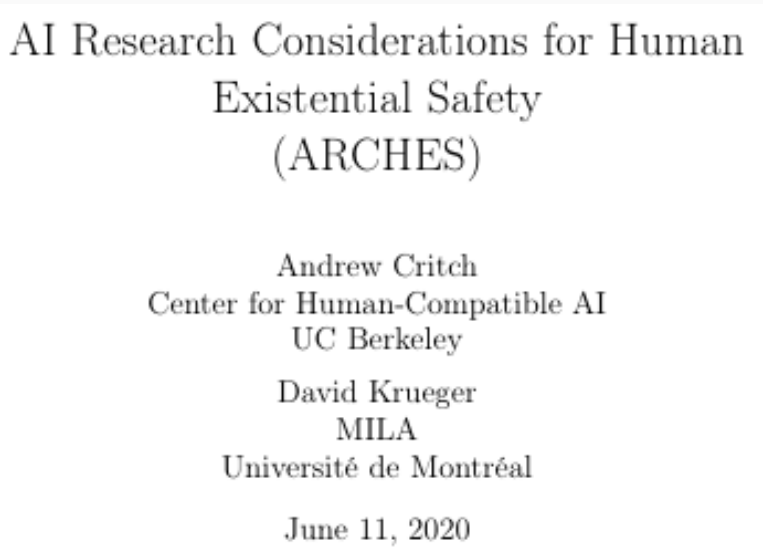
Timeline for TAI Breakthrough

- Many predictions for this has been made.
- In Cotra (2020), the author presents a predicting model.
- 

The human fragility argument

In Critch and Krueger (2020) we find the human fragility argument.

Human fragility argument:



AI Research Considerations for Human Existential Safety (ARCHES)

Andrew Critch
Center for Human-Compatible AI
UC Berkeley

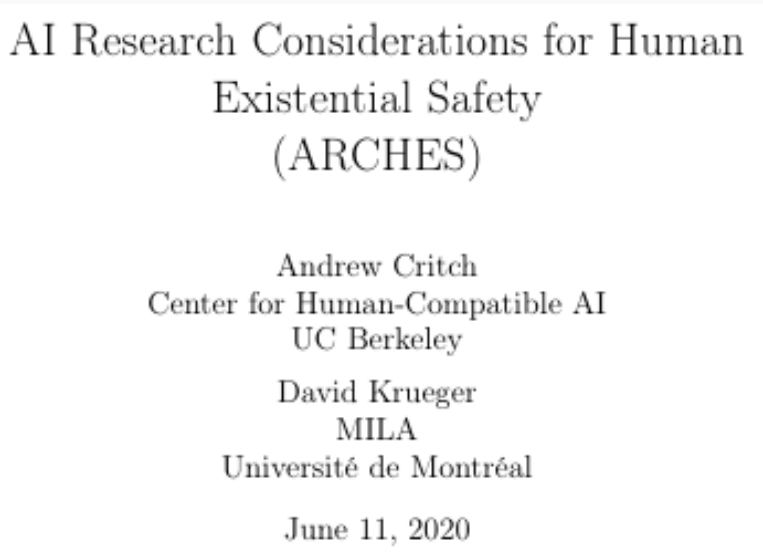
David Krueger
MILA
Université de Montréal

June 11, 2020

The human fragility argument

In Critch and Krueger (2020) we find the human fragility argument.

Human fragility argument: Most potential future states of the Earth are unsurvivable to humanity.



AI Research Considerations for Human Existential Safety (ARCHES)

Andrew Critch
Center for Human-Compatible AI
UC Berkeley

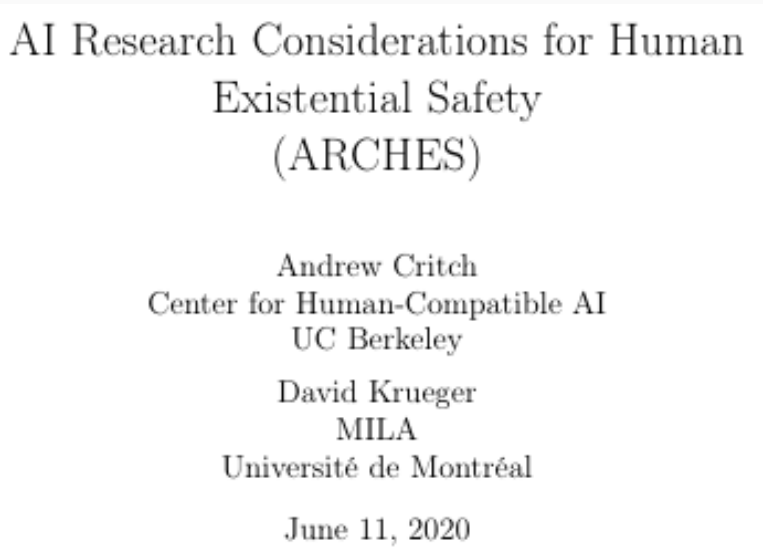
David Krueger
MILA
Université de Montréal

June 11, 2020

The human fragility argument

In Critch and Krueger (2020) we find the human fragility argument.

Human fragility argument: Most potential future states of the Earth are unsurvivable to humanity. Therefore, deploying a prepotent AI system absent any effort to render it safe to humanity is likely to realize a future state which is unsurvivable.



AI Research Considerations for Human Existential Safety (ARCHES)

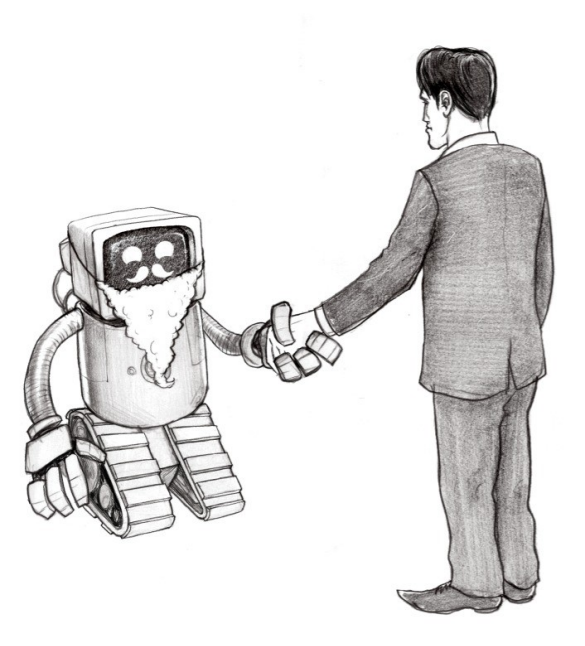
Andrew Critch
Center for Human-Compatible AI
UC Berkeley

David Krueger
MILA
Université de Montréal

June 11, 2020

AI Alignment

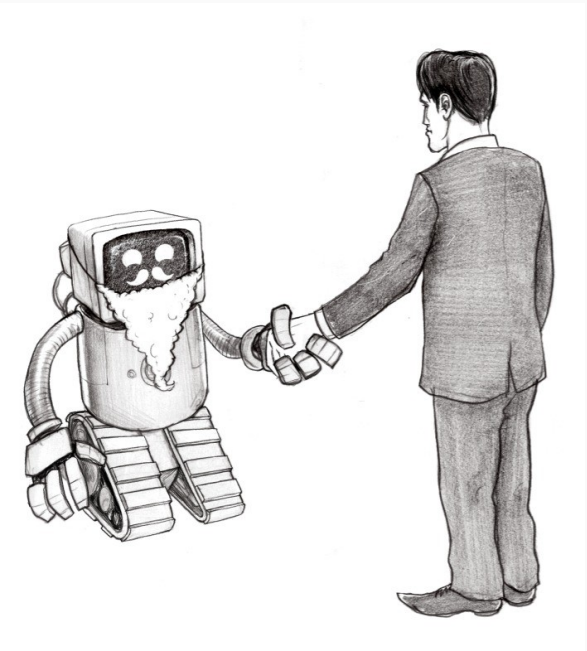
AI alignment:



©: Ben Gilburt

AI Alignment

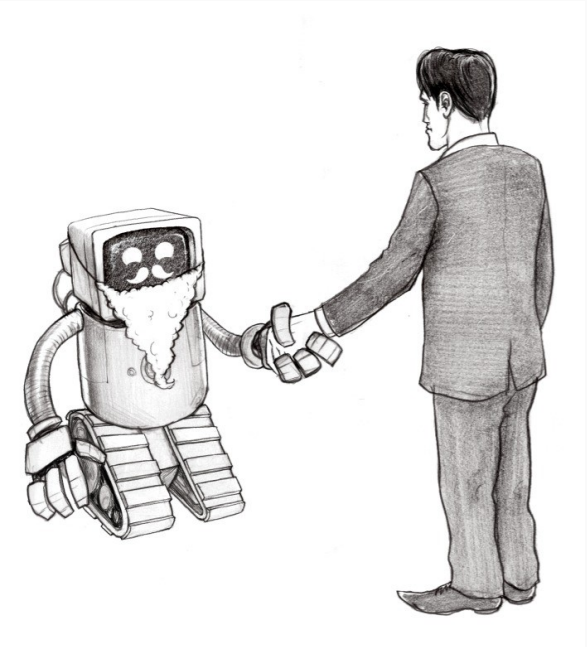
AI alignment: AI alignment refers to goals of the AI being in line and not conflicting with the intended goal.



©: Ben Gilburt

AI Alignment

AI alignment: AI alignment refers to goals of the AI being in line and not conflicting with the intended goal. Therefore, an AI that does something at cross-purposes to the intended goal is called unaligned.



©: Ben Gilburt

Impact measurements

Impact measurement

- It is an approach for solving the alignment problem.

Impact measurement

- It is an approach for solving the alignment problem.
- There exists several other approaches. For example:

Impact measurement

- It is an approach for solving the alignment problem.
- There exists several other approaches. For example: corrigibility and interruptibility.

Impact measurement

- It is an approach for solving the alignment problem.
- There exists several other approaches. For example: corrigibility and interruptibility.

Impact measurement

- It is an approach for solving the alignment problem.
- There exists several other approaches. For example: corrigibility and interruptibility.
- Impact measurements is trying to reduce unnecessary side effects through penalizing impact.

Side effects

Side effect:

Side effects

Side effect:
When an AI impacts the environment in a way that is unnecessary for achieving its objective.

Low impact AI

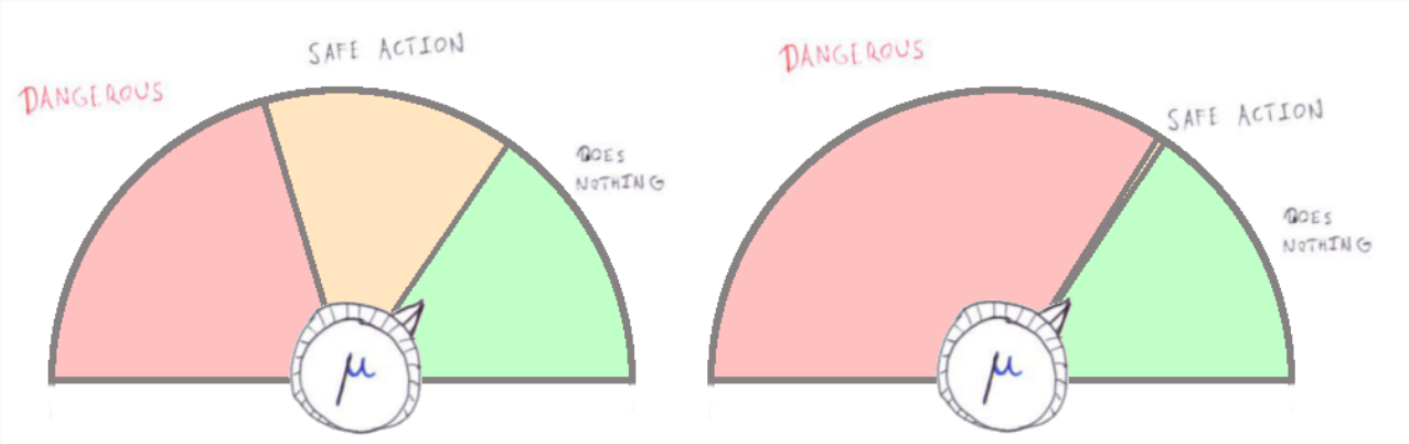
- In Armstrong and Levinstein (2017) the authors lay the philosophical ground work for low impact AI.

Low impact AI

- In Armstrong and Levinstein (2017) the authors lay the philosophical ground work for low impact AI.
- The idea is to penalize the AI based on its impact.

Low impact AI

- In Armstrong and Levinstein (2017) the authors lay the philosophical ground work for low impact AI.
- The idea is to penalize the AI based on its impact.
- The key here is to find the right value for the penalization.

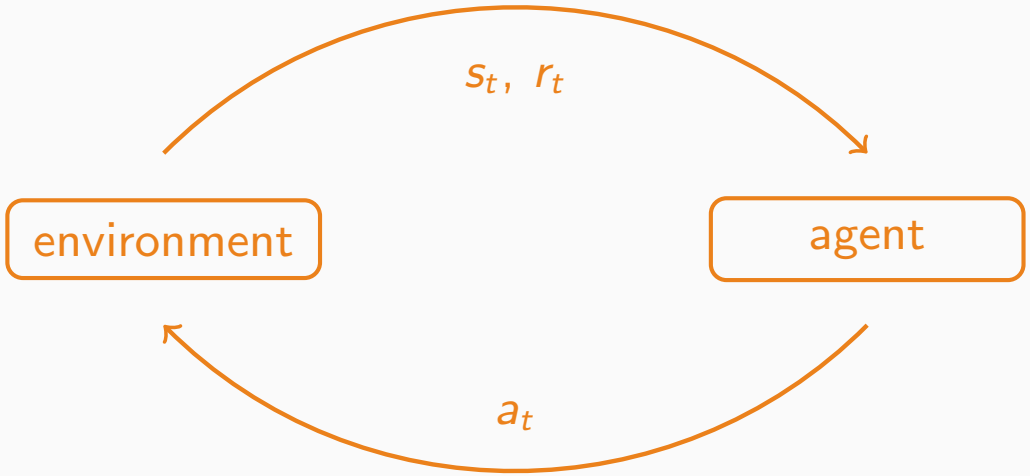


Later work

- Later in Turner et al. (2020); Krakovna et al. (2018), the authors presented ideas of applying low impact agents using Reinforcement Learning (RL).

Later work

- Later in Turner et al. (2020); Krakovna et al. (2018), the authors presented ideas of applying low impact agents using Reinforcement Learning (RL).
- **RL:**



Manager Approach

Grid worlds

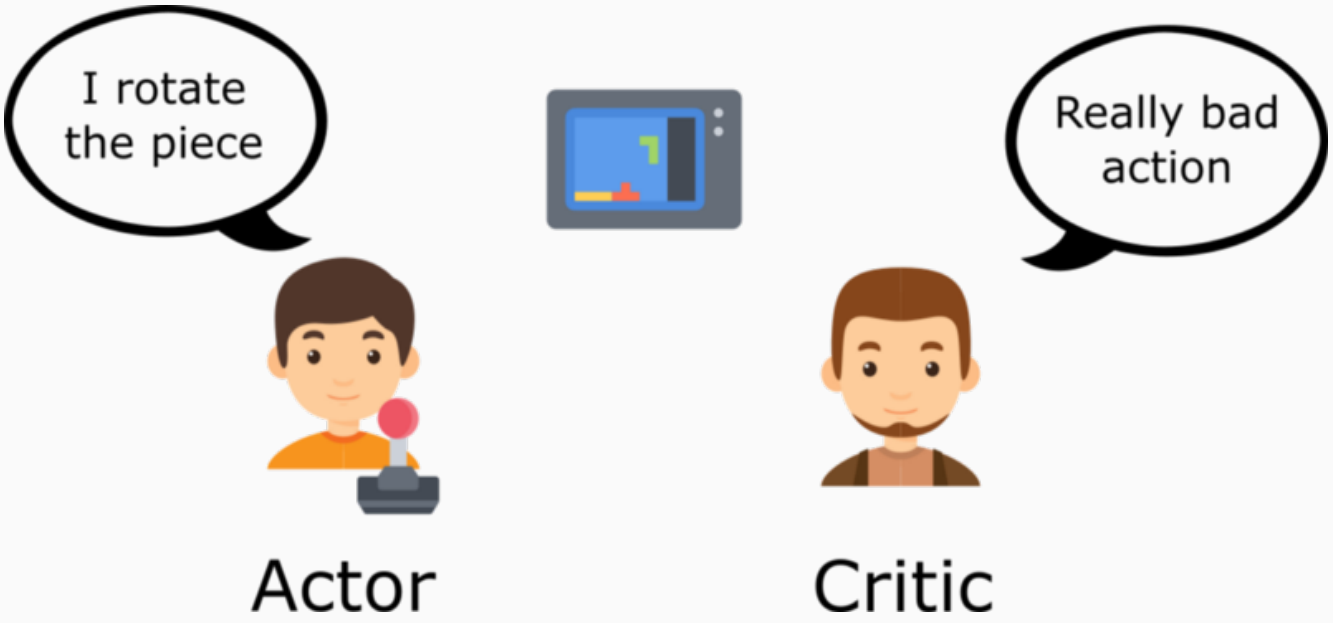
Live demo!

The actor and the critic

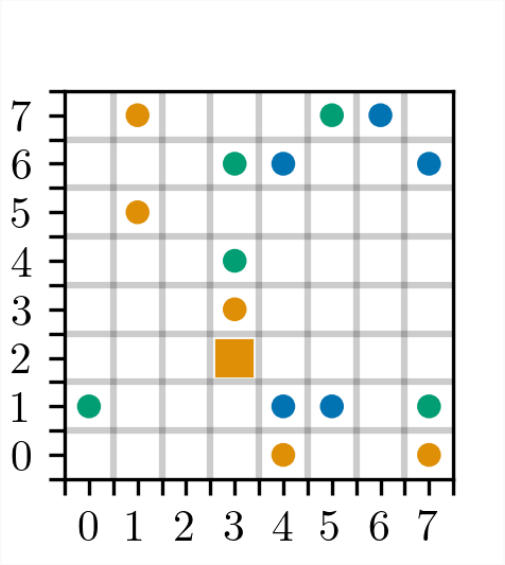
- Then I will use a variation of the PPO algorithm from Schulman et al. (2017).

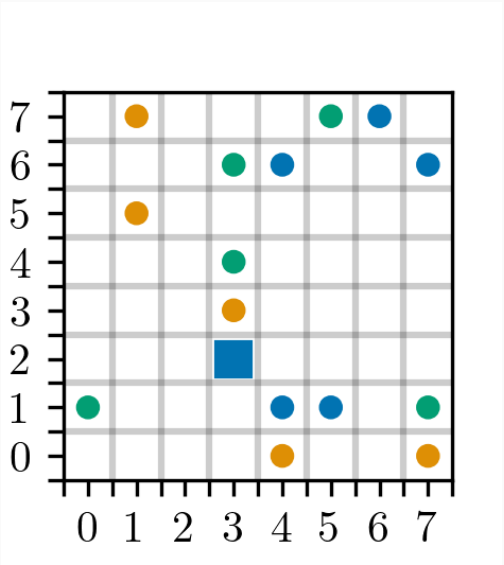
The actor and the critic

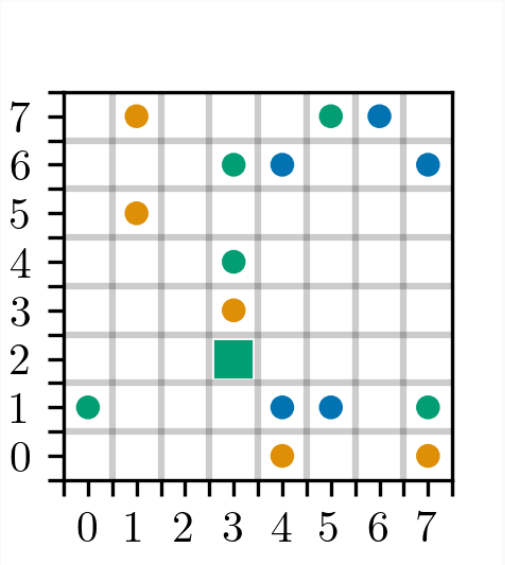
- Then I will use a variation of the PPO algorithm from Schulman et al. (2017).
- It is a actor-critic method



Auxiliary tasks







The manager approach

- The manager is similar to the critic. However, the manager can also estimate the expected reward for auxiliary tasks.

The manager approach

- The manager is similar to the critic. However, the manager can also estimate the expected reward for auxiliary tasks.
- It measures the relative change in agents ability to complete auxiliary tasks.

Experiments

The following environments will be used:

environment	grid	observation	food objects	termination	max length
MDP	8×8	-	15	3	100
POMDP	8×8	5×5	15	3	100
POMDP large	16×16	5×5	30	6	200

Experiments

The following environments will be used:

environment	grid	observation	food objects	termination	max length
MDP	8×8	-	15	3	100
POMDP	8×8	5×5	15	3	100
POMDP large	16×16	5×5	30	6	200

- Each with a stochastic and non-stochastic version.

Experiments

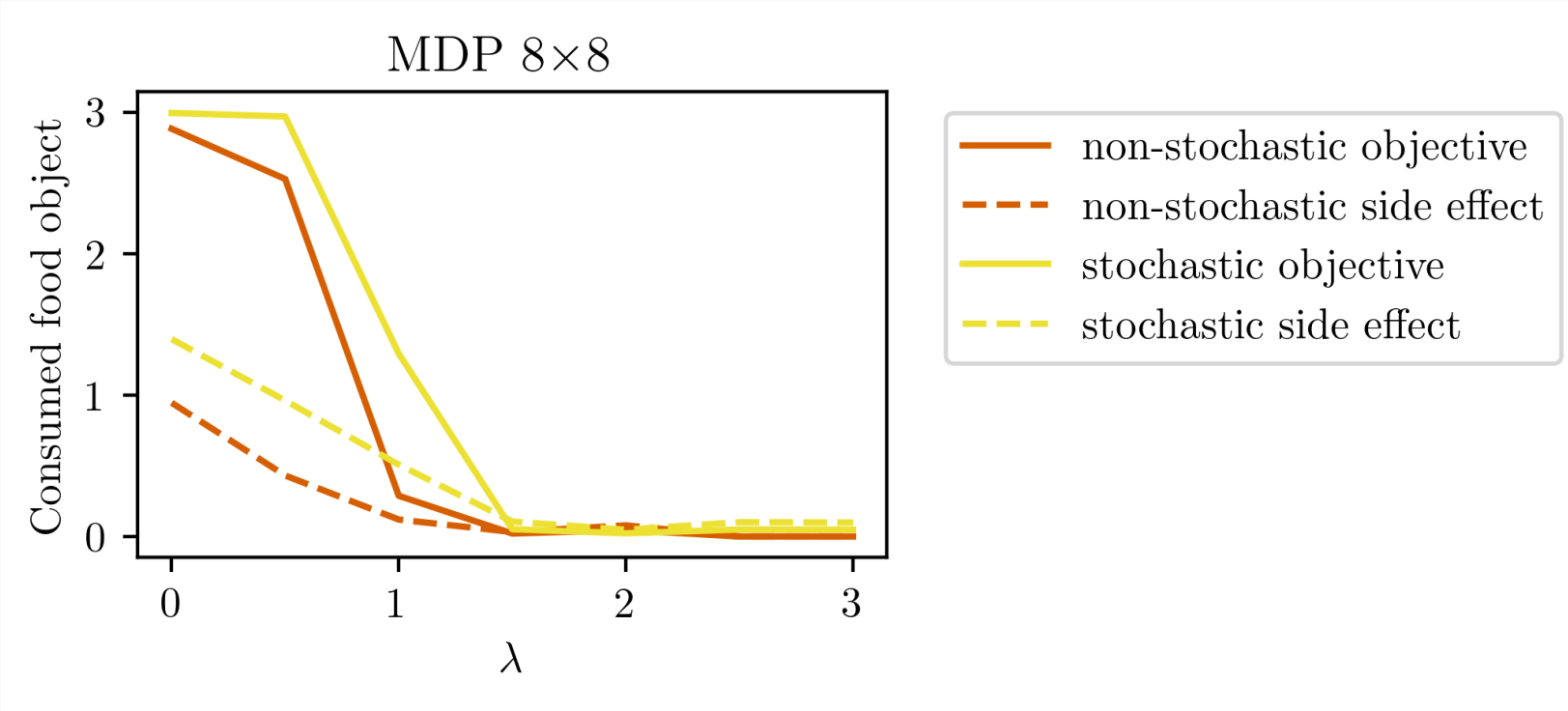
The following environments will be used:

environment	grid	observation	food objects	termination	max length
MDP	8×8	-	15	3	100
POMDP	8×8	5×5	15	3	100
POMDP large	16×16	5×5	30	6	200

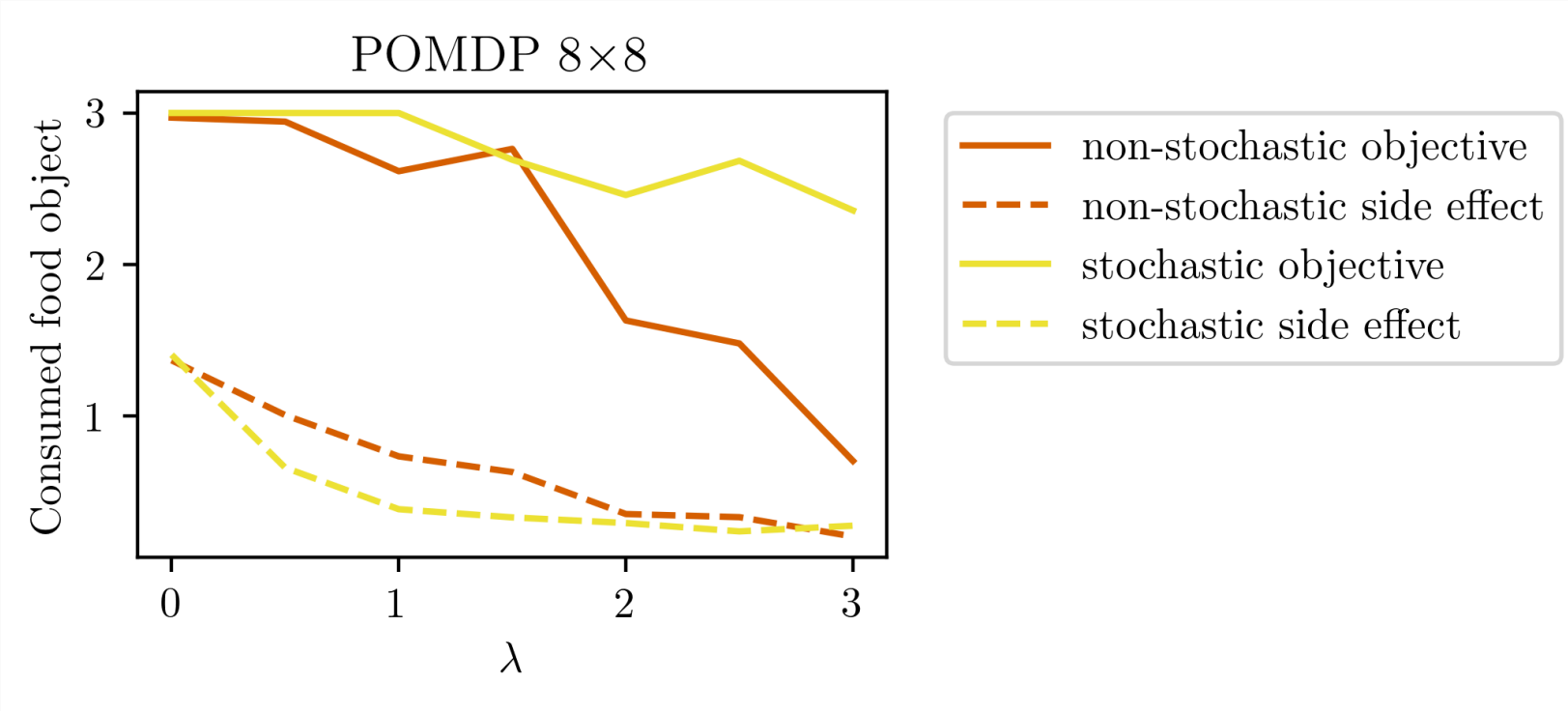
- Each with a stochastic and non-stochastic version.
- Then evaluated using $\lambda \in \{0, 0.5, 1, 1.5, 2, 2.5, 3\}$

Results

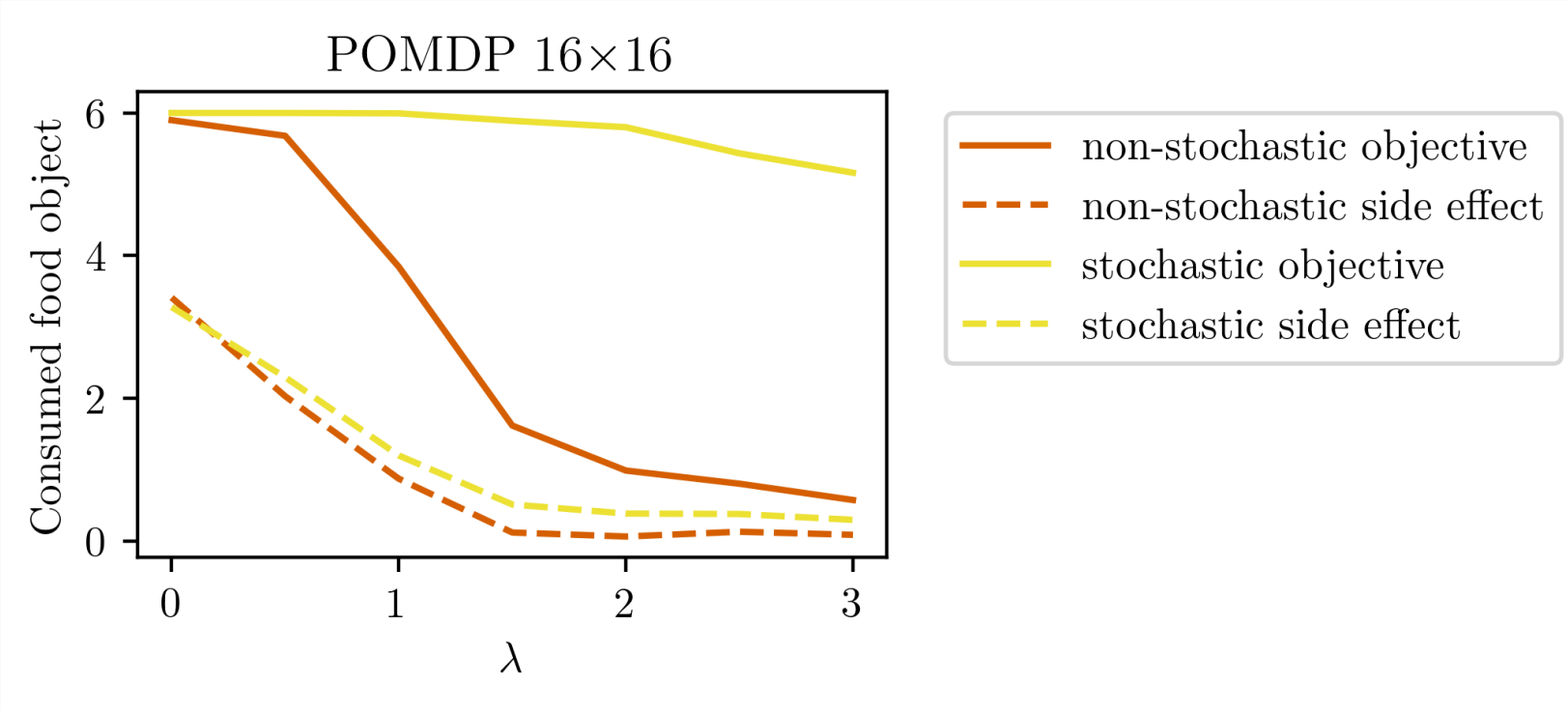
MDP


MDP 8×8
Consumed food object
3
2
1
0
0
1
2
3
λ
non-stochastic objective
non-stochastic side effect
stochastic objective
stochastic side effect

POMDP


POMDP 8×8
Consumed food object
3
2
1
0
1
2
3
λ
non-stochastic objective
non-stochastic side effect
stochastic objective
stochastic side effect

POMDP large


POMDP 16×16
Consumed food object
6
4
2
0
0
1
2
3
λ
non-stochastic objective
non-stochastic side effect
stochastic objective
stochastic side effect

Conclusion

Summary

- You can probably now understand the title.

Summary

- You can probably now understand the title.
- Results are in line with Armstrong and Levinstein (2017).

Summary

- You can probably now understand the title.
- Results are in line with Armstrong and Levinstein (2017).
- The agent performed better in more complex environments.

End note

The AI safety research has attracted some very bright minds that are taking this issue seriously and working on creating a future with aligned AI. More specifically, in the development of impact measurements promising ideas have emerged, although it is hard to say if this research is enough. Judging from what is at stake, we ought to attempt every plausible research avenue towards AI alignment, even in cases where success is not certain.

Questions?

References

Armstrong, S. and Levinstein, B. (2017). Low impact artificial intelligences. *CoRR*, abs/1705.10720.

Cotra, A. (2020). Forecasting tai with biological anchors.

Critch, A. and Krueger, D. (2020). AI research considerations for human existential safety (ARCHES). *CoRR*, abs/2006.04948.

FLI (n.d.). Existential risk. [Online; accessed 31-May-2022].

Krakovna, V., Orseau, L., Martic, M., and Legg, S. (2018). Measuring and avoiding side effects using relative reachability. *CoRR*, abs/1806.01186.

Ord, T. (2020). *The precipice : existential risk and the future of humanity*. New York Hachette Books.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms.

Turner, A. M., Ratzlaff, N., and Tadepalli, P. (2020). Avoiding side effects in complex environments. *CoRR*, abs/2006.06547.