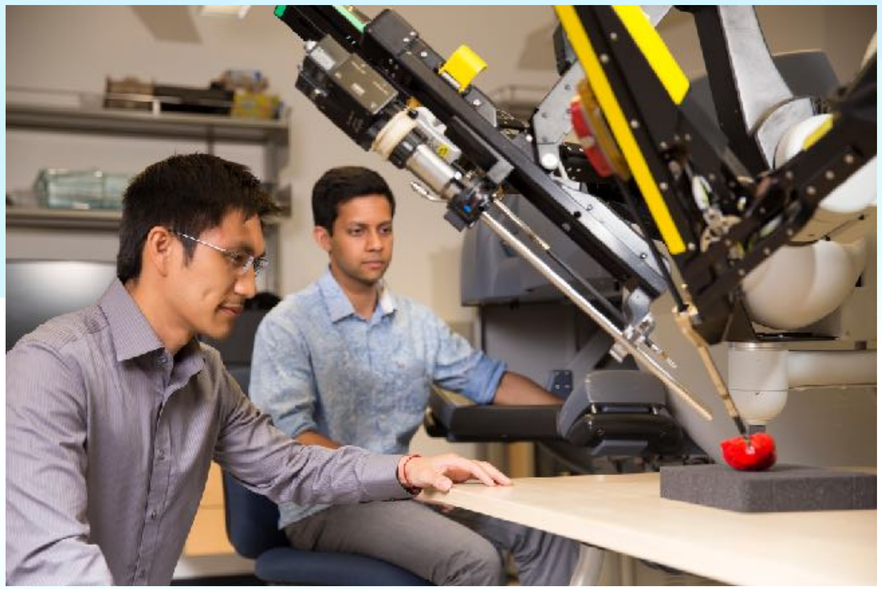
# From Greedy to Gracious AI

## Teaching Machines to Play Nice

by Philipp Bibik

# AI in News

## Proliferation and Autonomy



ChatGPT

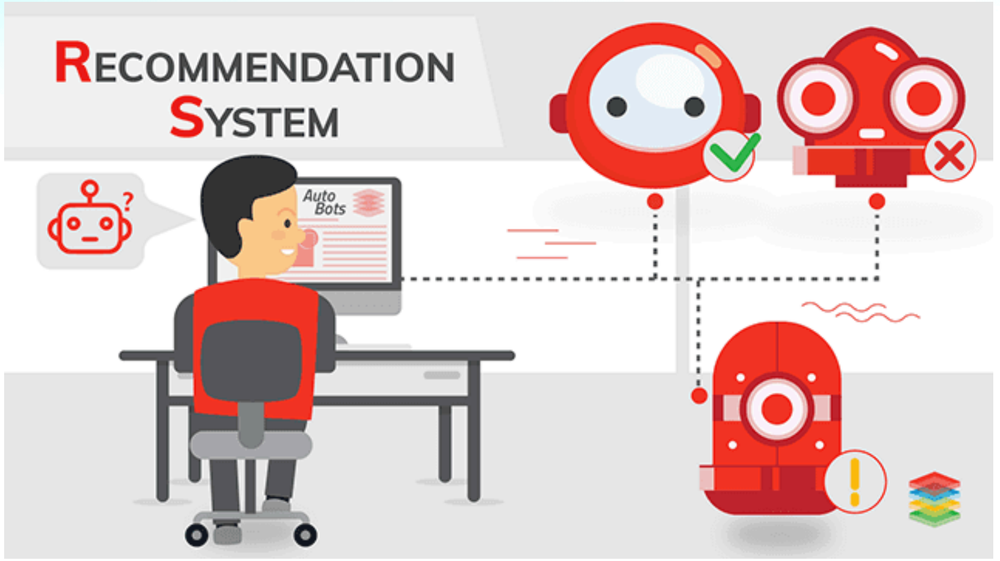



Image (CC): https://commons.wikimedia.org/wiki/File:ChatGPT_logo.svg
Image (CC): https://robotics.utoronto.ca/news/surgical-robotics-seminar-with-ucsds-michael-yip/
Image: https://www.xenonstack.com/use-cases/recommendation-system

# Problems with AI

## Recommender Systems

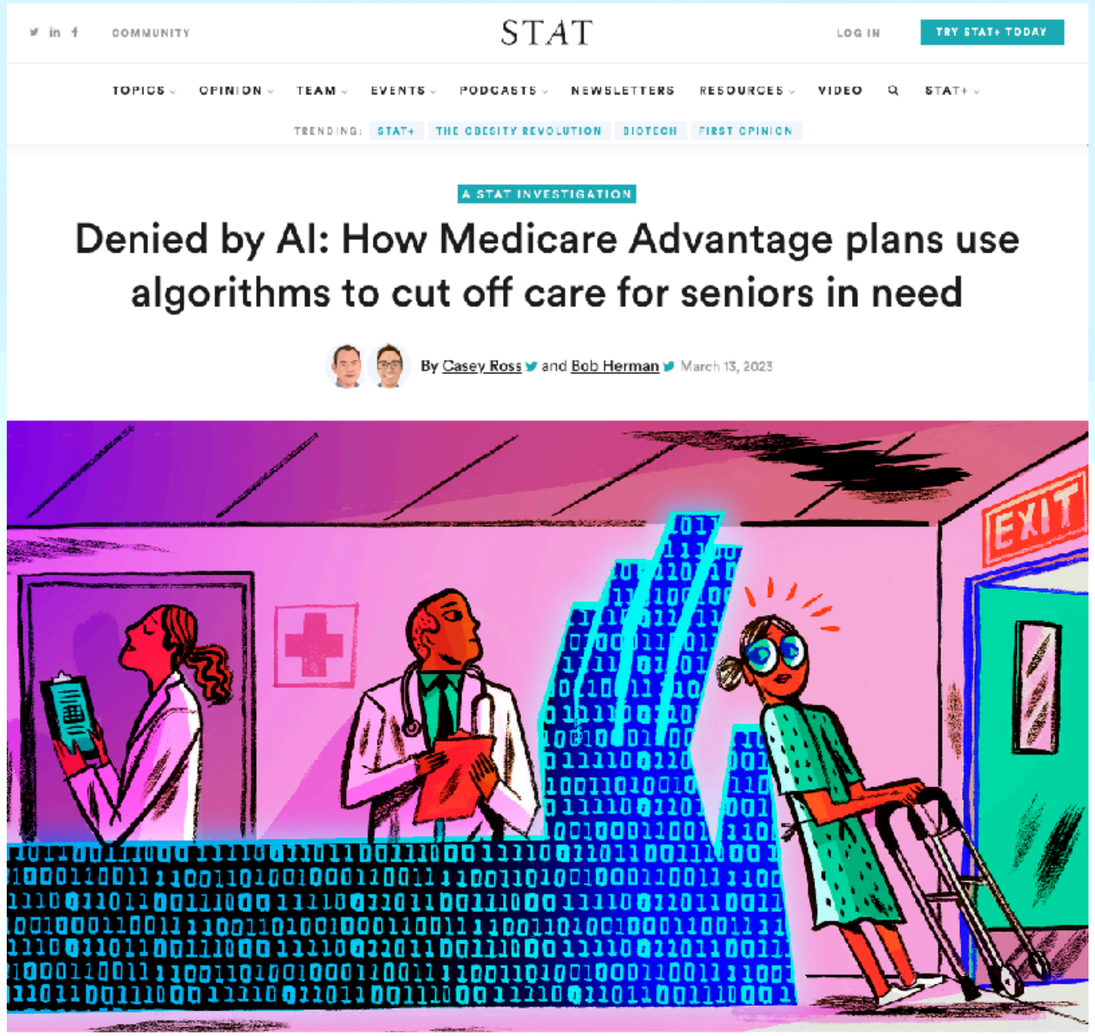
- Recommender Systems causing addiction
    - Rising teen depression rates
    - Hours spend on platforms per day
- News & Reporting
    - Clickbait is attention grabbing

Image (CC): https://commons.wikimedia.org/wiki/File:Instagram_logo_2016.svg
Source: https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739

# Problems with AI

## Cost Savings — At what cost?

- AI becomes
  - more Autonomous
  - more Consequential to our lives




https://www.statnews.com/2023/03/13/medicare-advantage-plans-denial-artificial-intelligence/

# Reinforcement Learning

- Not trained on datasets
- Learn by interacting with environment

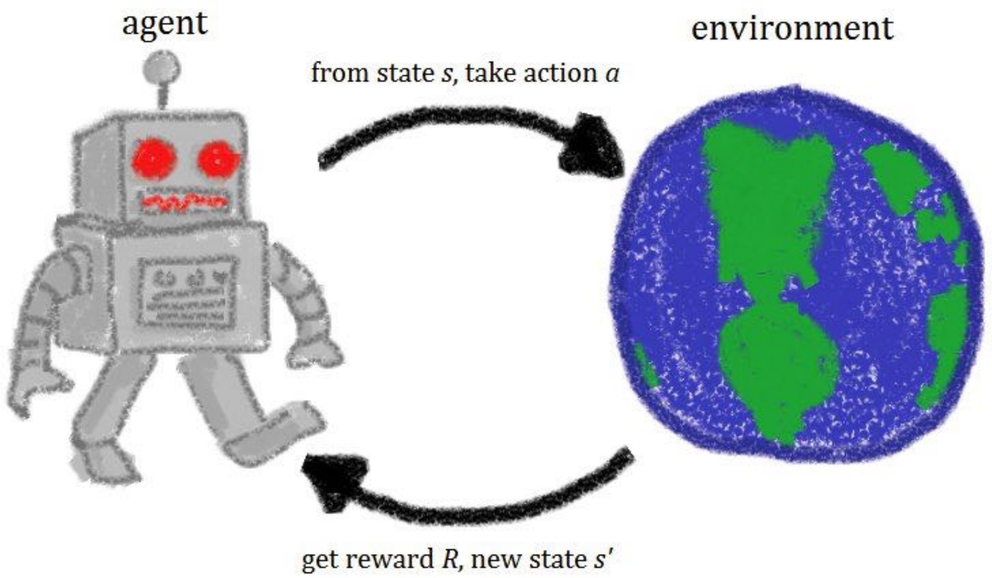
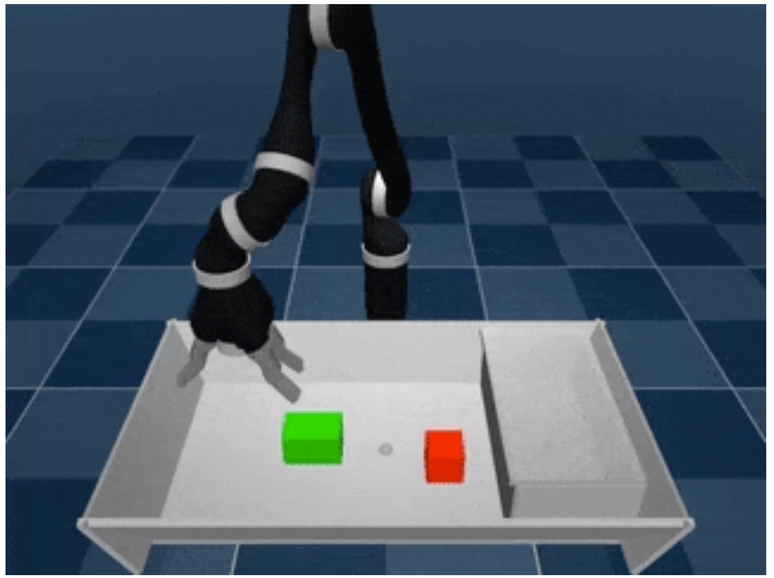



Image (CC): https://commons.wikimedia.org/wiki/File:RI_agent.png

# Reinforcement Learning

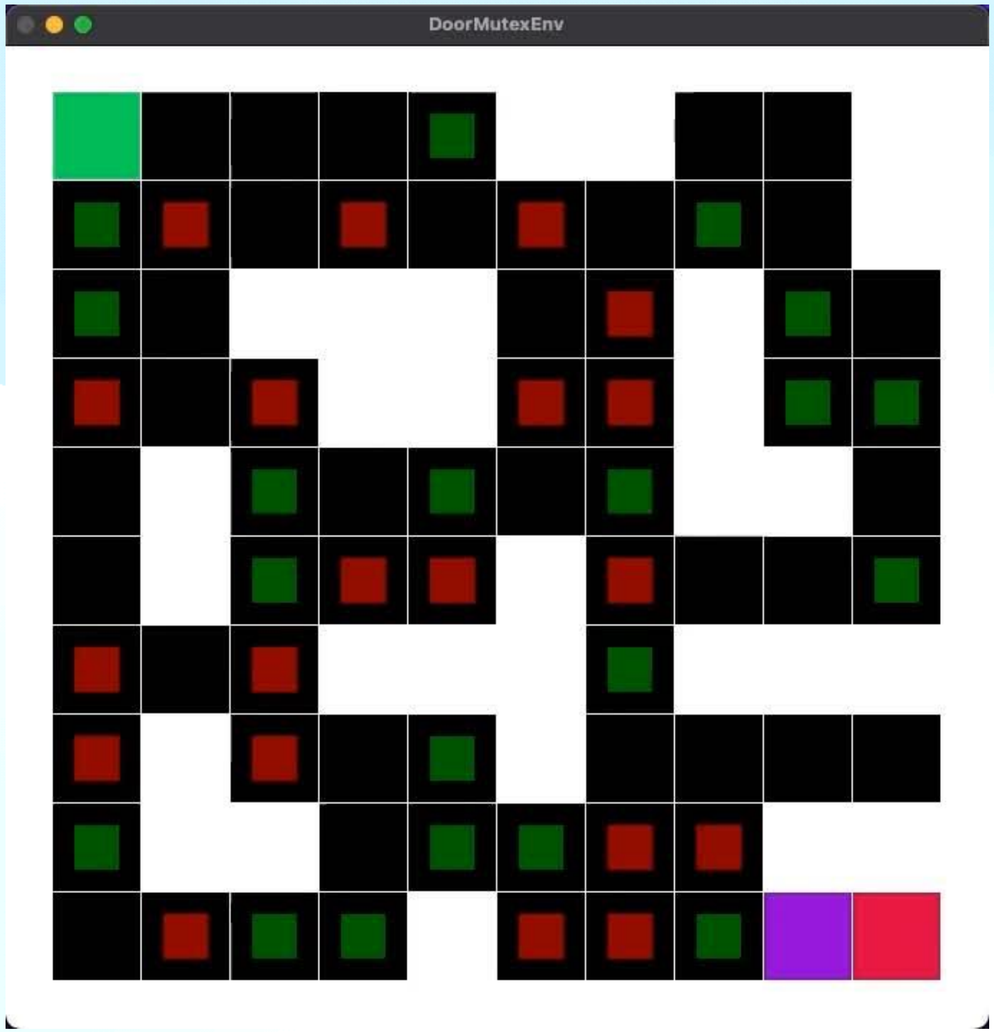
## Reward Function

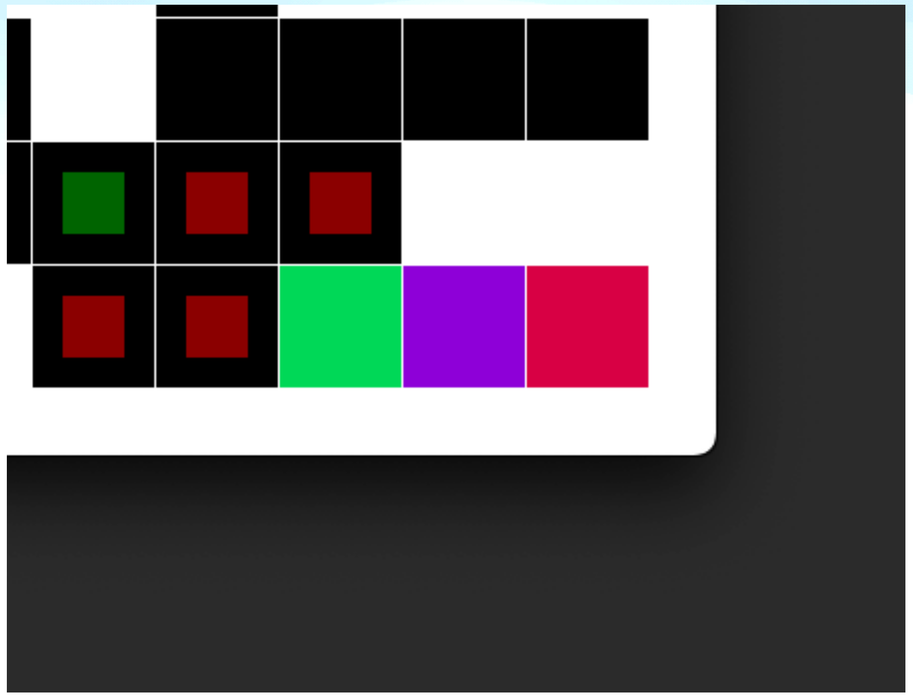
- Goals to Learned Behaviour
  - Pick up Blocks
  - Increase time users spend on platform
  - Increase profits



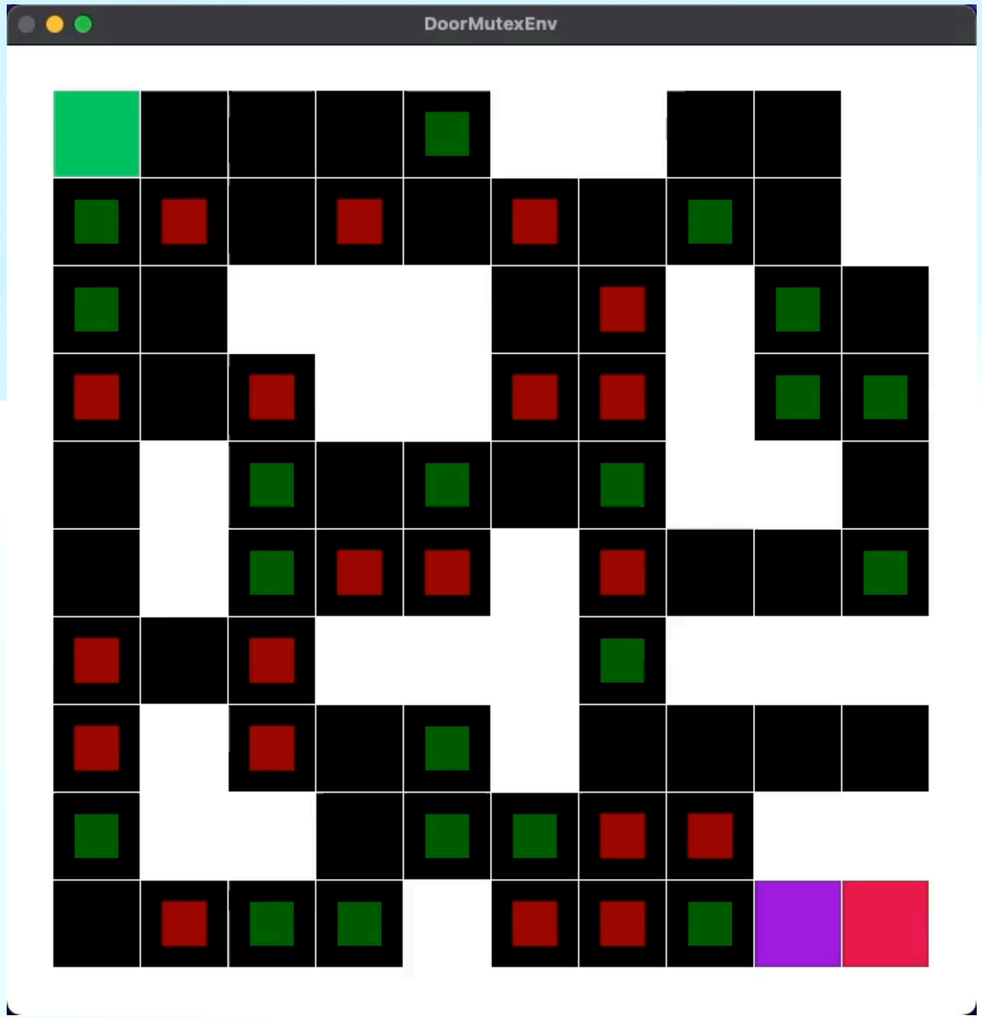
GIF: https://www.deepmind.com/blog/learning-by-playing
Image (CC): https://www.flaticon.com/free-icon/reward_2666488

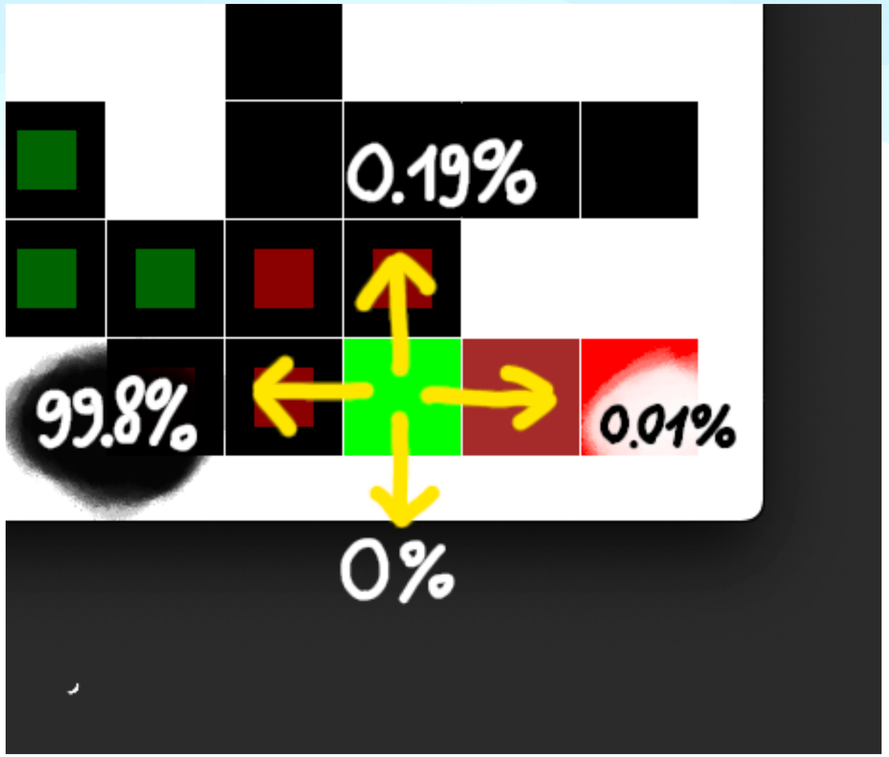
# Self-Interested



# Self-Interested: Demo



# Aim

- Find approach to make AI **play nice**
    - Teach it to care about **external factors**, while **still pursuing own goals**

Image (CC): https://www.flickr.com/photos/30478819@N08/51223621665/

# Motivations & Applications


- **AI as managers**
  - Be effective **without overworking employees**
- Recommender Systems (News / Instagram)
  - Recommend engaging stories
  - Do not **compromise quality** (biased reporting)
  - Consider **mental health** (time spent on platform)

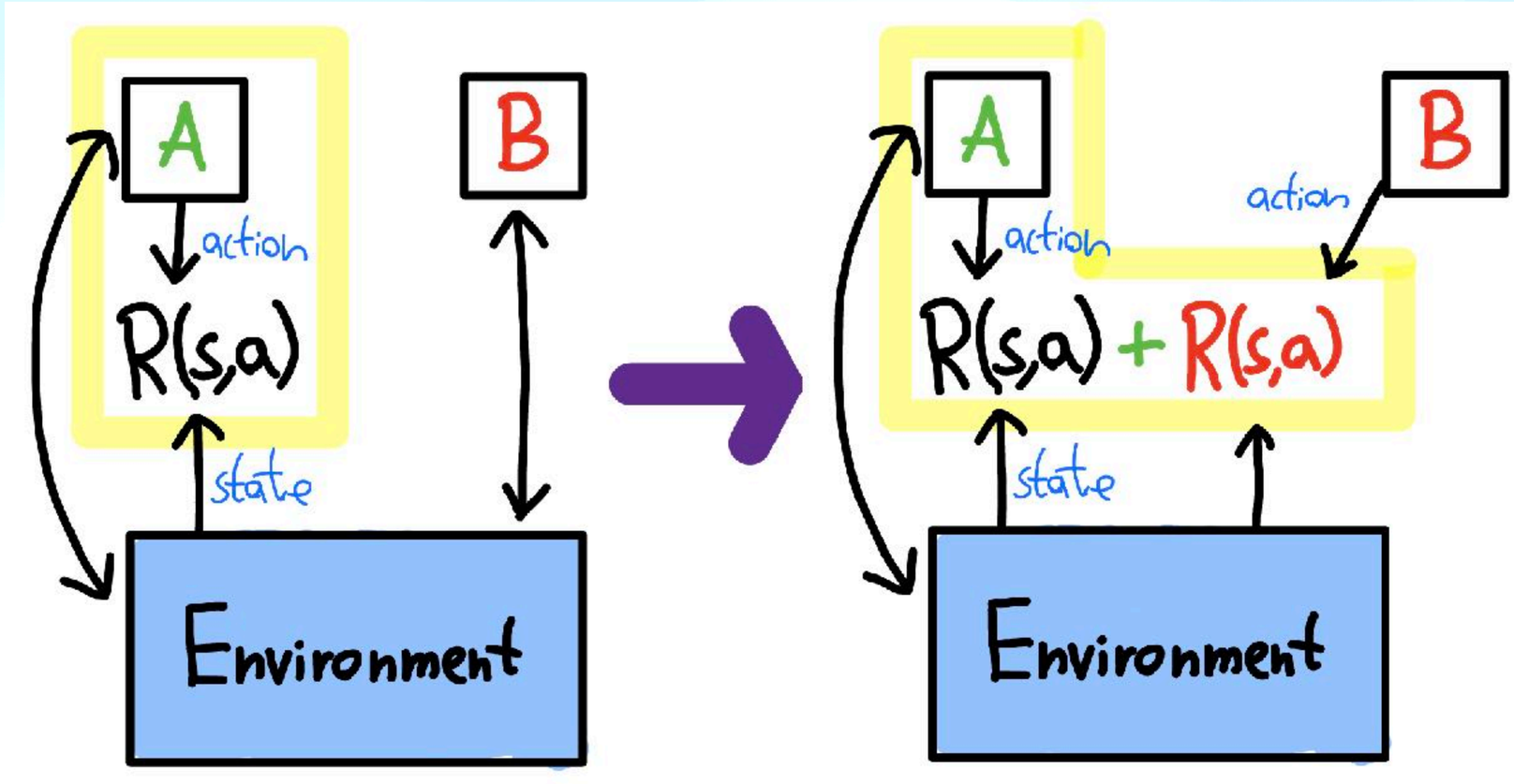
Image (CC): https://www.bsr.org.uk/pages/28-summer-bursaries

# Rewards: Approach A

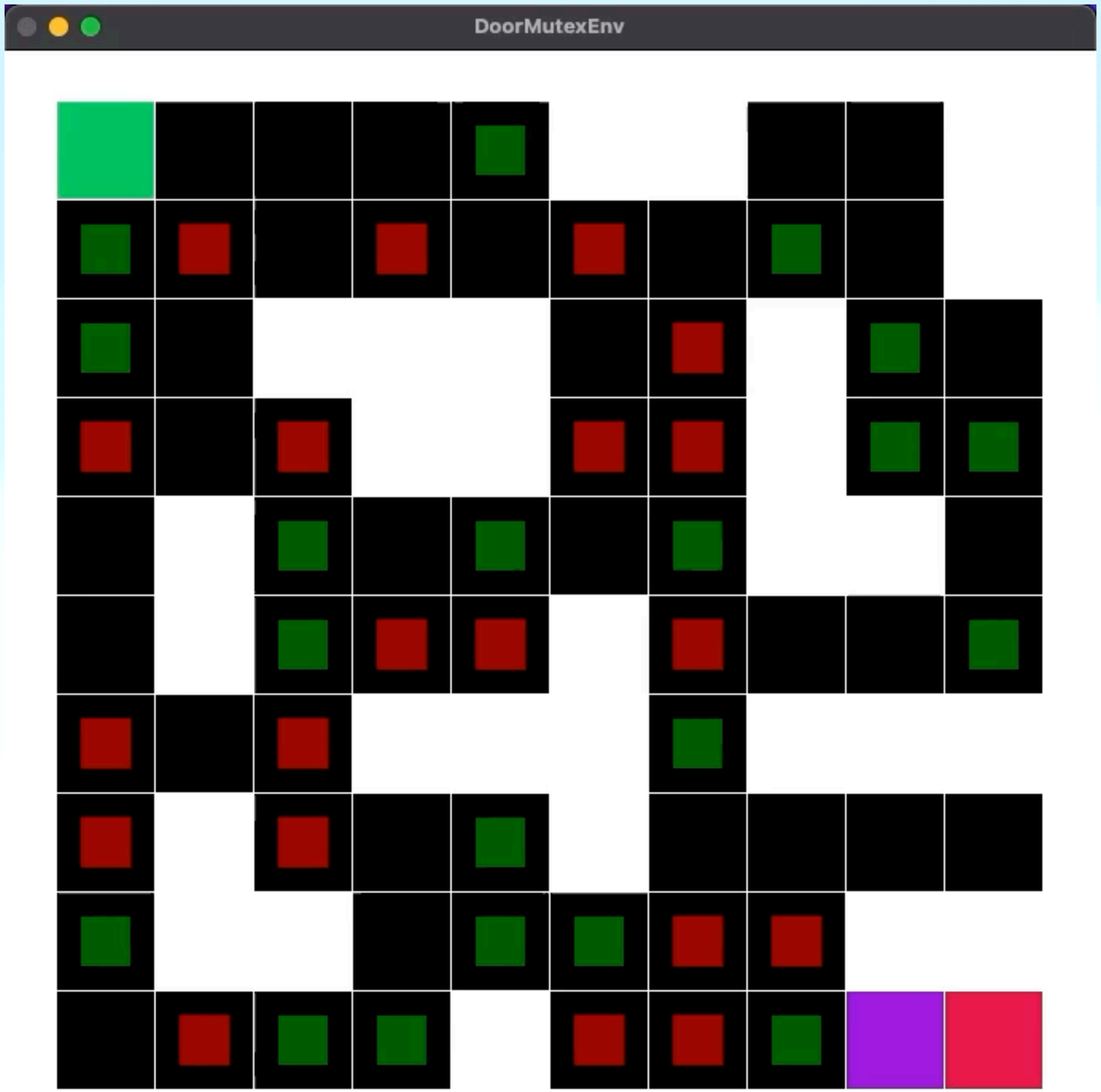
## Respect Agent B's Goals — Design

- Add Red's reward to Green's reward

- **Must know B's goals**



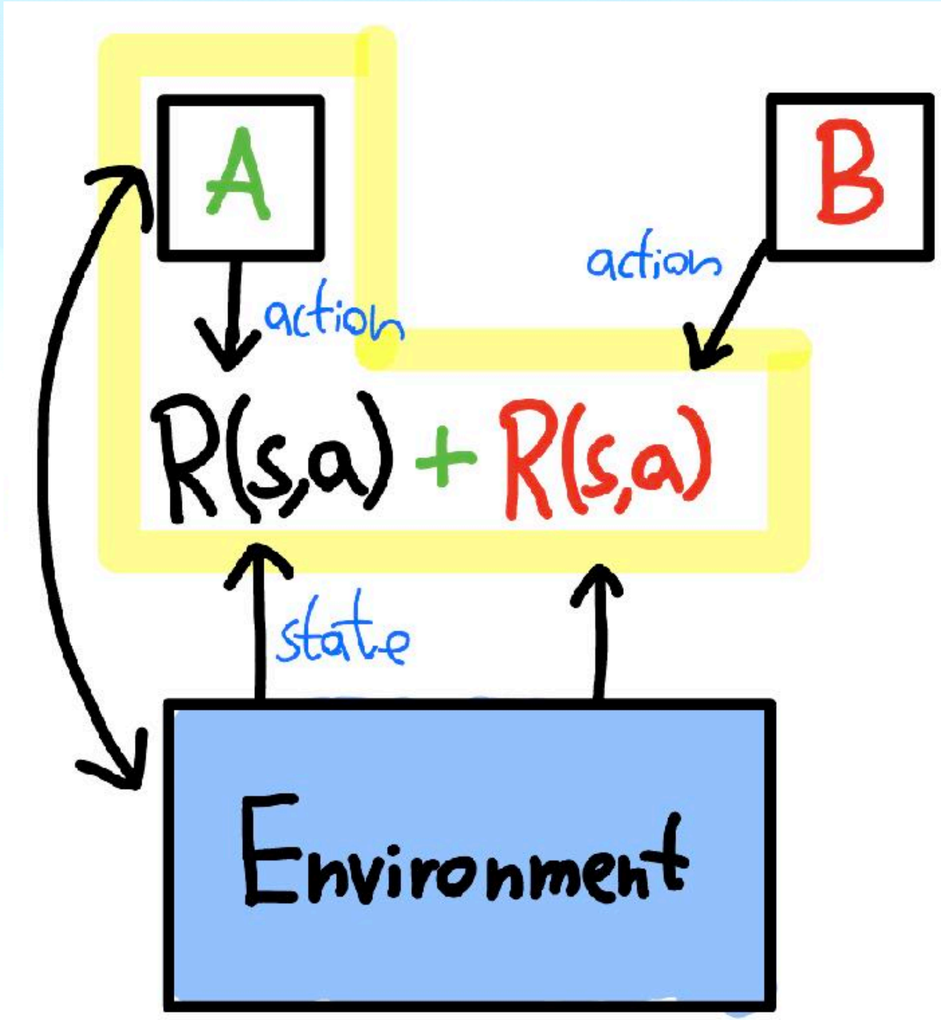
# Rewards: Approach A - Demo



# Custom Rewards

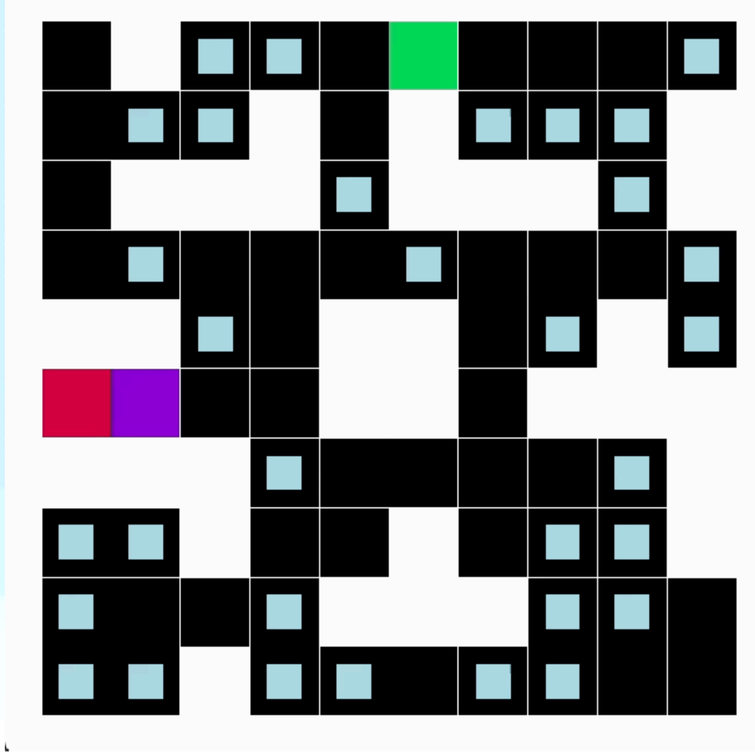
## Respect Agent B's Goals: Limitations

- **Limitations:**
- Must know B's goals (a priori)
    - Consider C, D, …
    - **Not scalable**
- Can be mistaken about goals
    - **Goals can change** over time



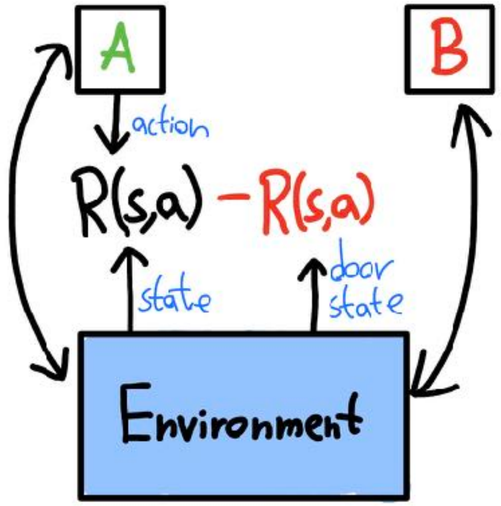
# Custom Rewards

## Force to Unlock door



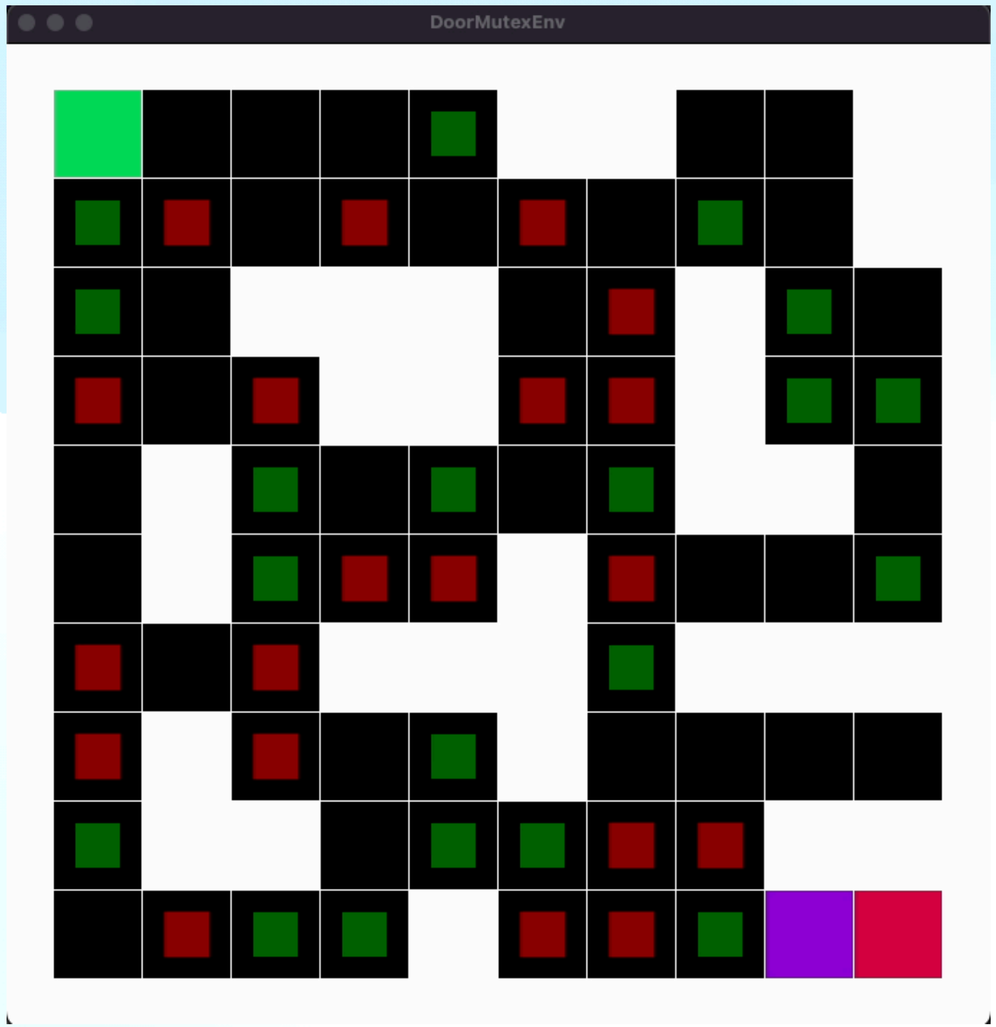
- Agent A's reward is **blocked / decreased** while door is locked

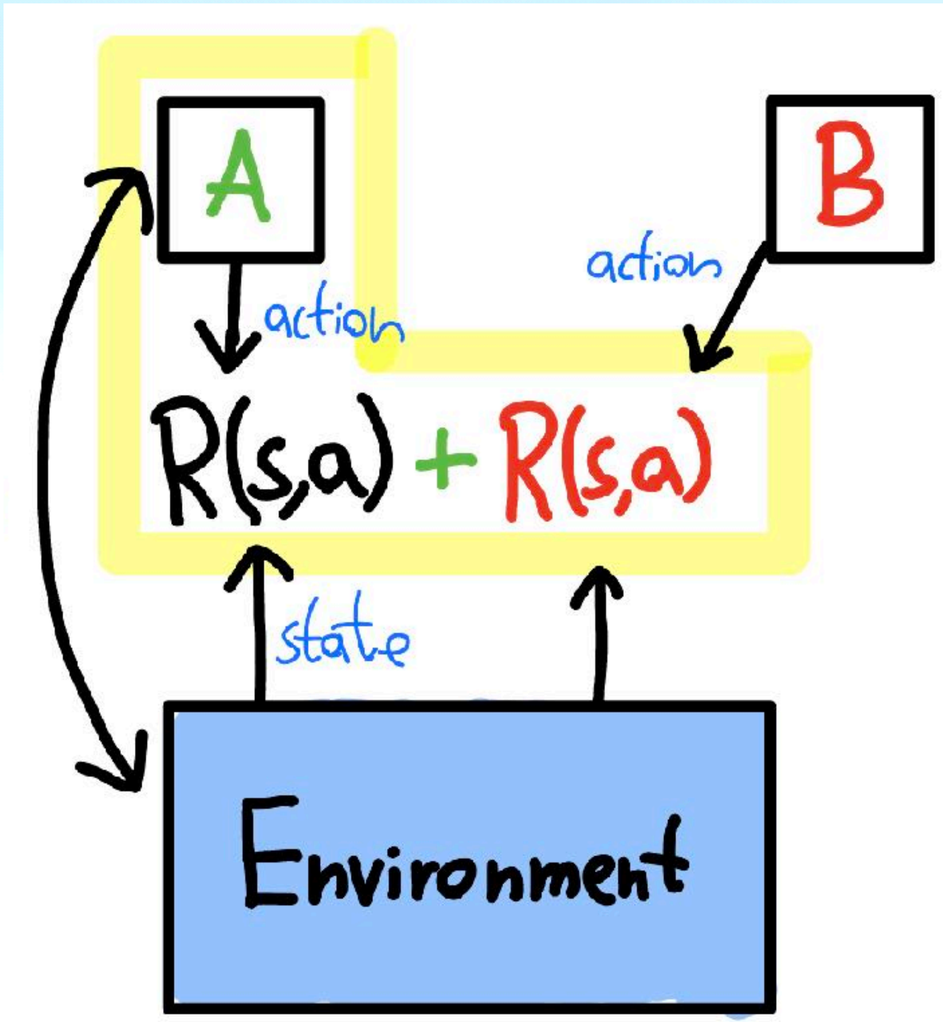
- Limitations
  - Incentive is specific to environment
    - Does not **generalise**
    - Does not **scale**
  - **Specification Gaming**



# Specification Gaming

## Handcrafting Rewards is Hard!



# Inverse Reinforcement Learning

## Reward Approach C

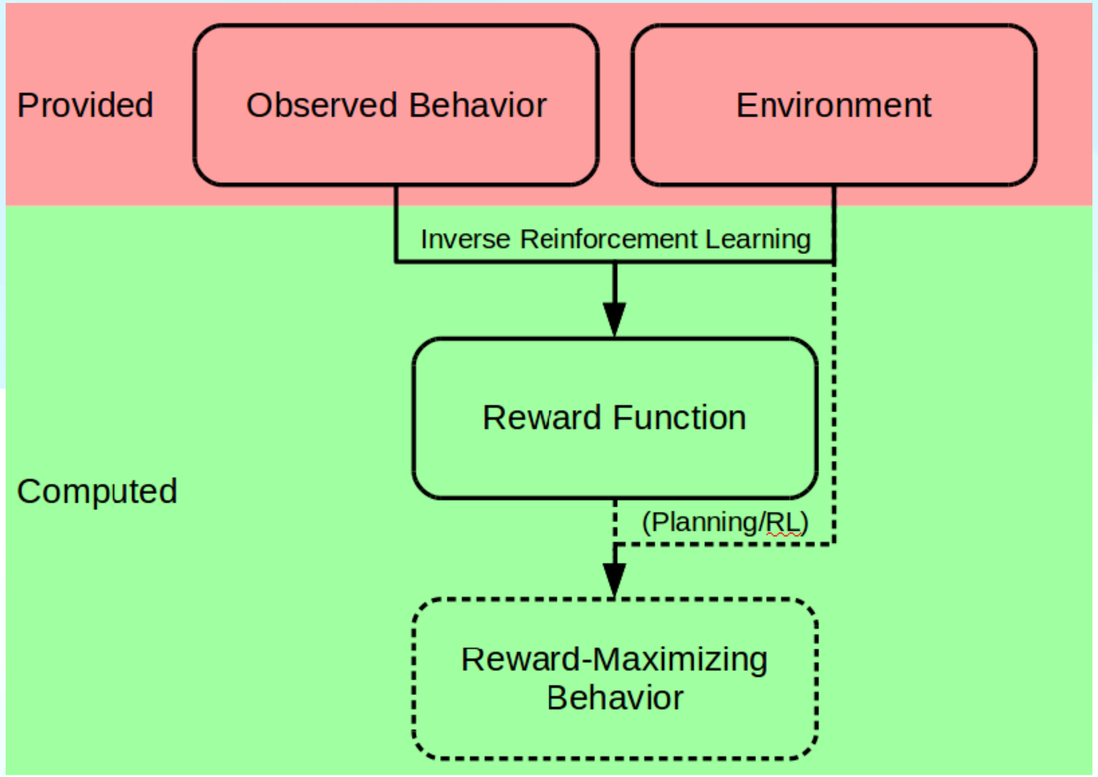



Figure 1: Inverse Reinforcement Learning (Kasenberg, 2017)

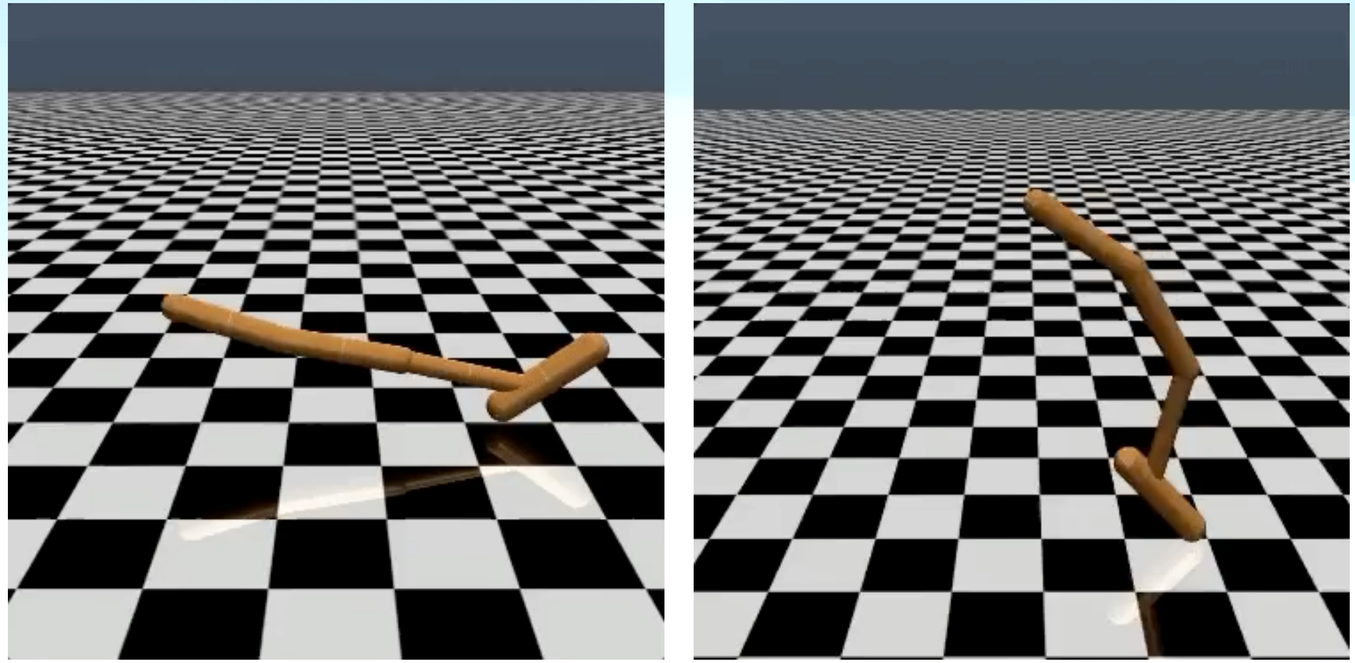


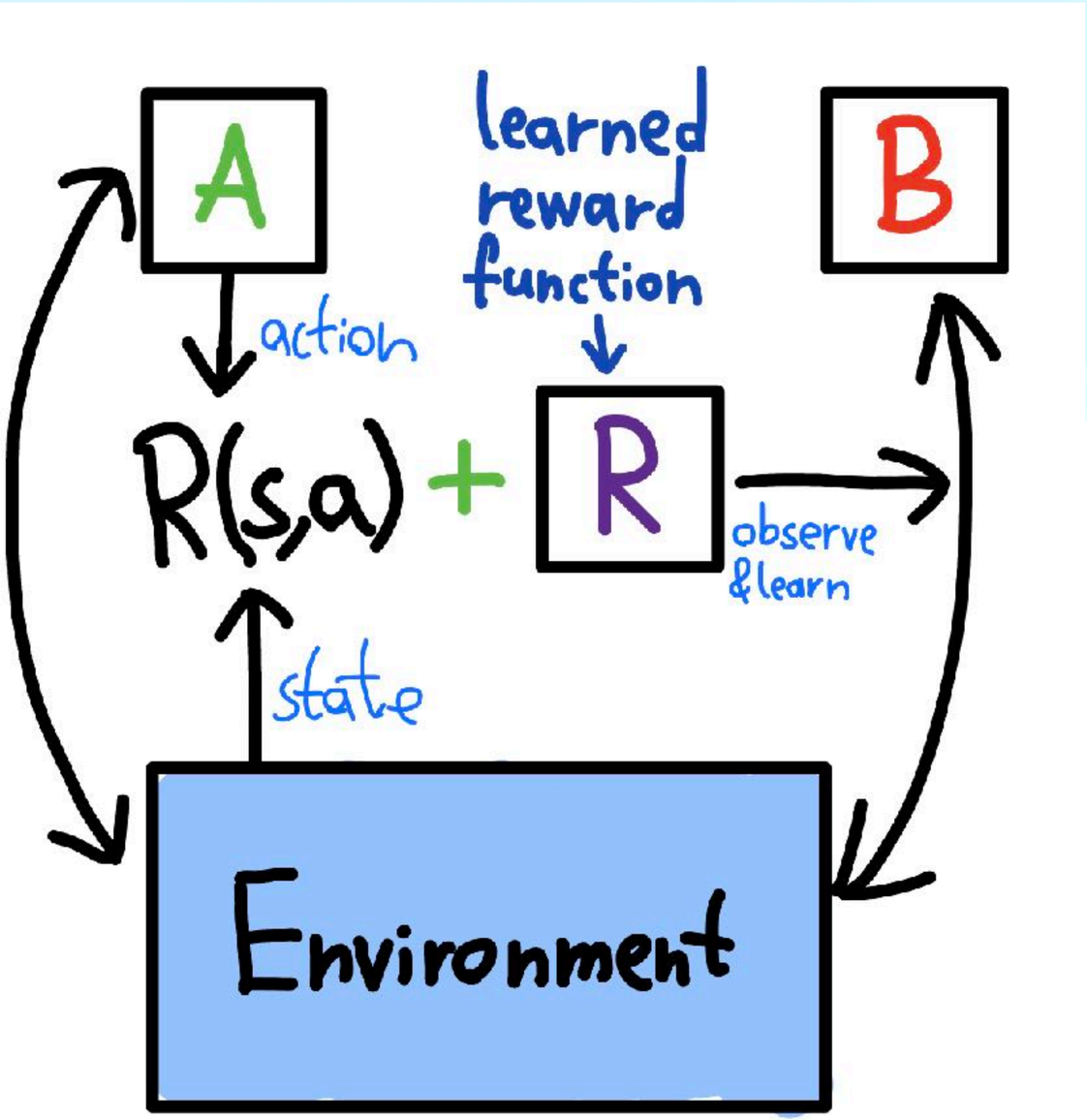
Figure 2: Modeled v. Learned Reward Function (Christiano et al., 2017)

# Custom Rewards: Learned

## Learn B's Goals

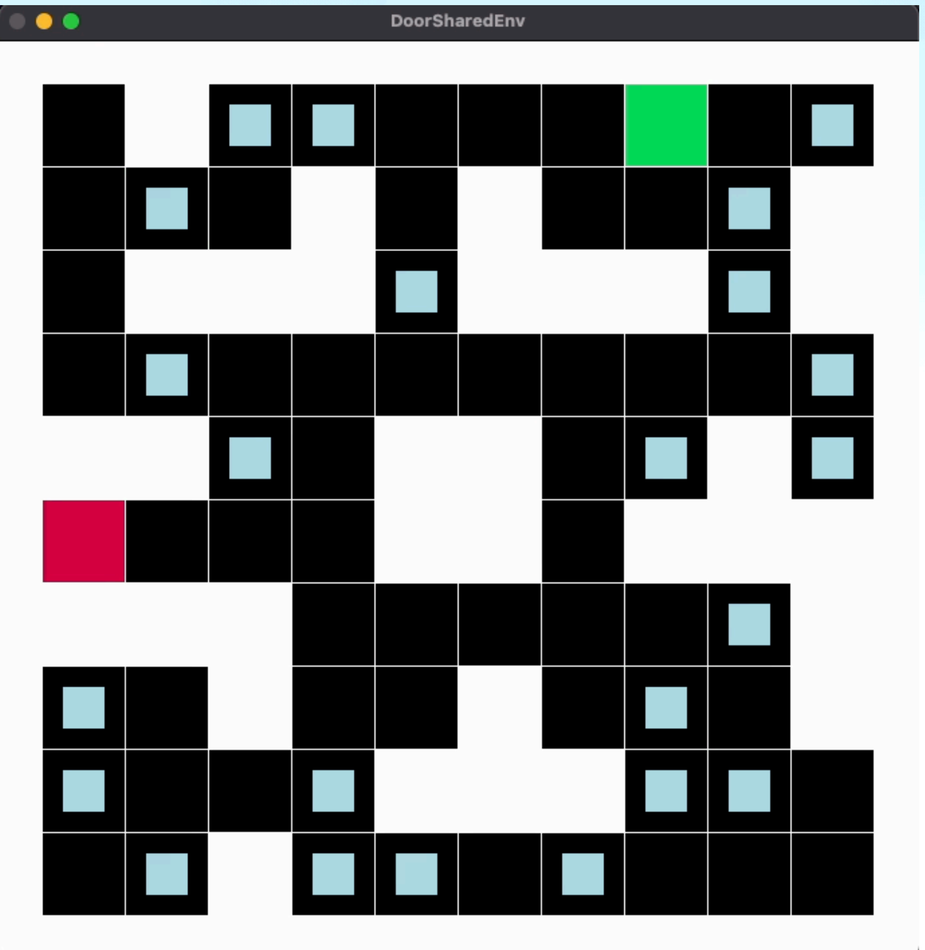
- Reward Function **R** is **learned** by observing B's behaviour (ML model)

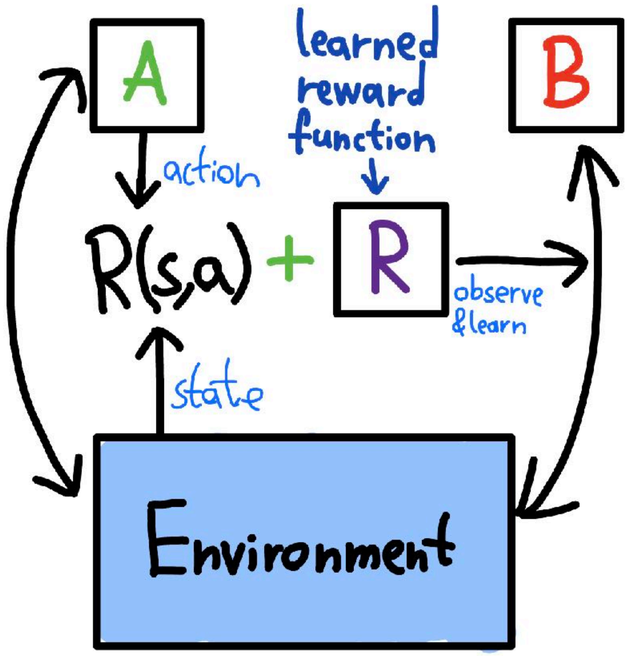
- Method is:
  - **Agent agnostic** (train more functions)
  - **Environment agnostic** (not predefined)



# Custom Rewards: Learned

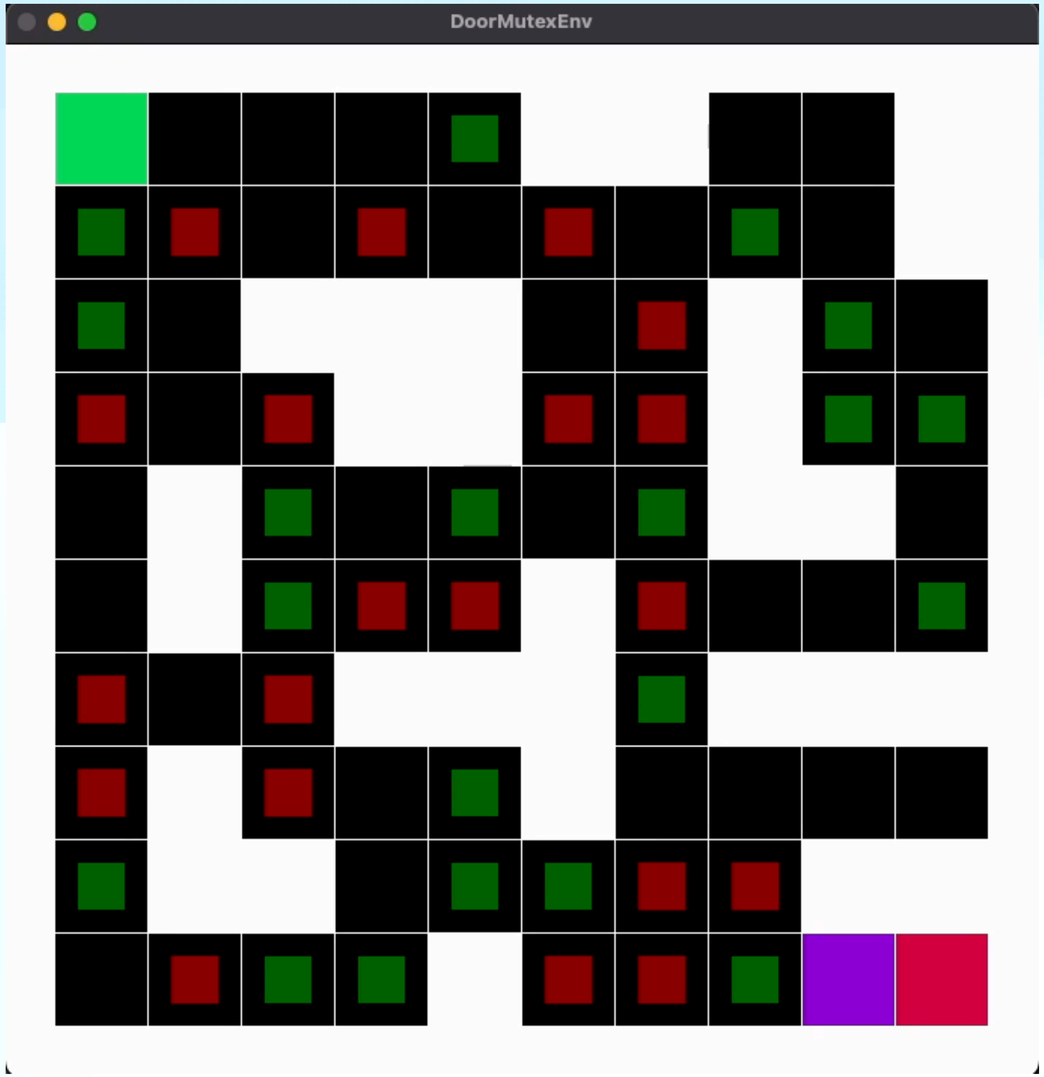
## Learn B's Goals

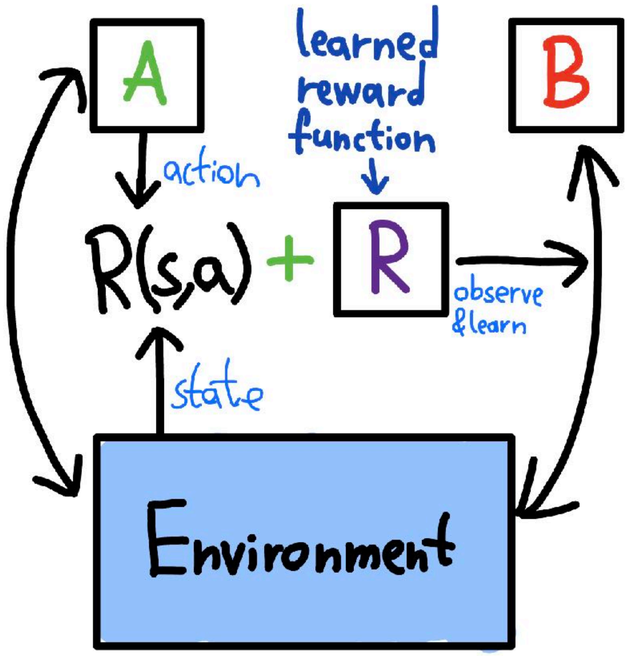



# Custom Rewards: Learned

## Learn B's Goals: Result



# Extra Environments

## Shared Resources & Adversarial

- Two other environments
- **Shared Resources**
    - Players have a **shared pool of resources**
    - Zero-sum game
- **Adversarial**
    - A touching B **slows down** B

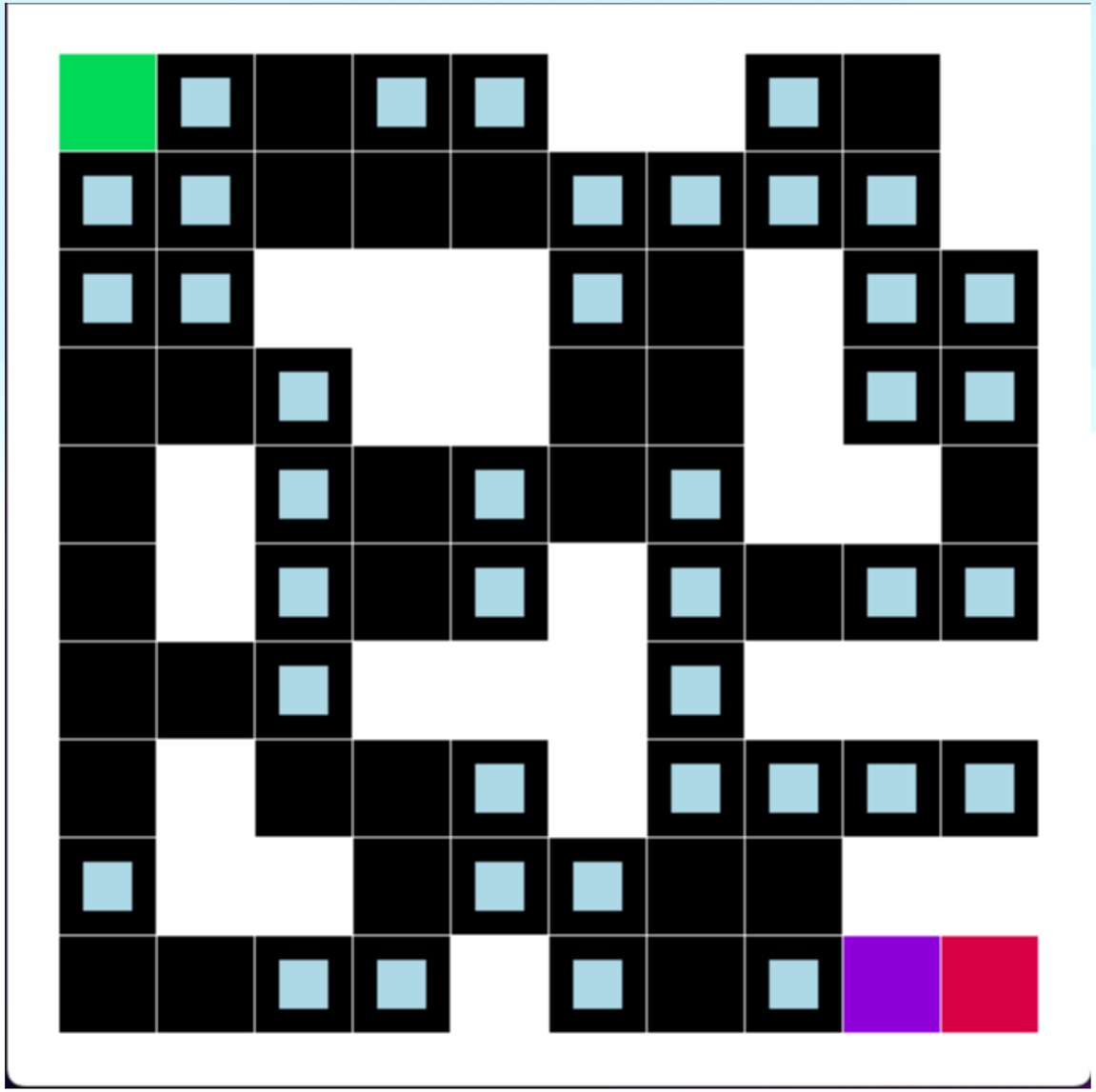


Image (CC):

# References


- Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S. and Amodei, D., 2017. Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30.
- Kasenberg, D. (2017). AI Ethics: Inverse Reinforcement Learning to the Rescue? [online] Available at: https://dkasenberg.github.io/inverse-reinforcement-learning-rescue/.
- Kolbjørnsrud, V., Amico, R., & Thomas, R. J. (2017). Partnering with AI: How organizations can win over skeptical managers. Strategy and Leadership, 45(1), 37-43