
Concrete Problems in AI Safety

Dario Amodei*
Google Brain

Chris Olah*
Google Brain

Jacob Steinhardt
Stanford University

Paul Christiano
UC Berkeley

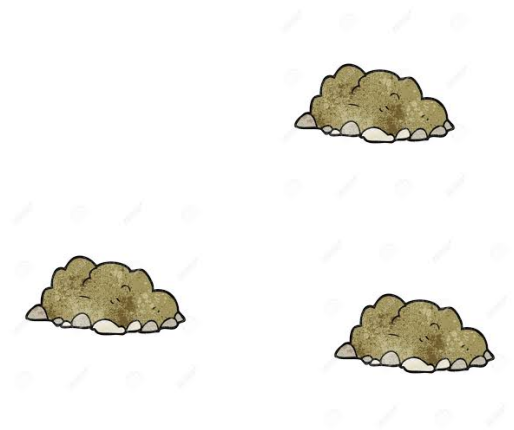
John Schulman
OpenAI

Dan Mané
Google Brain

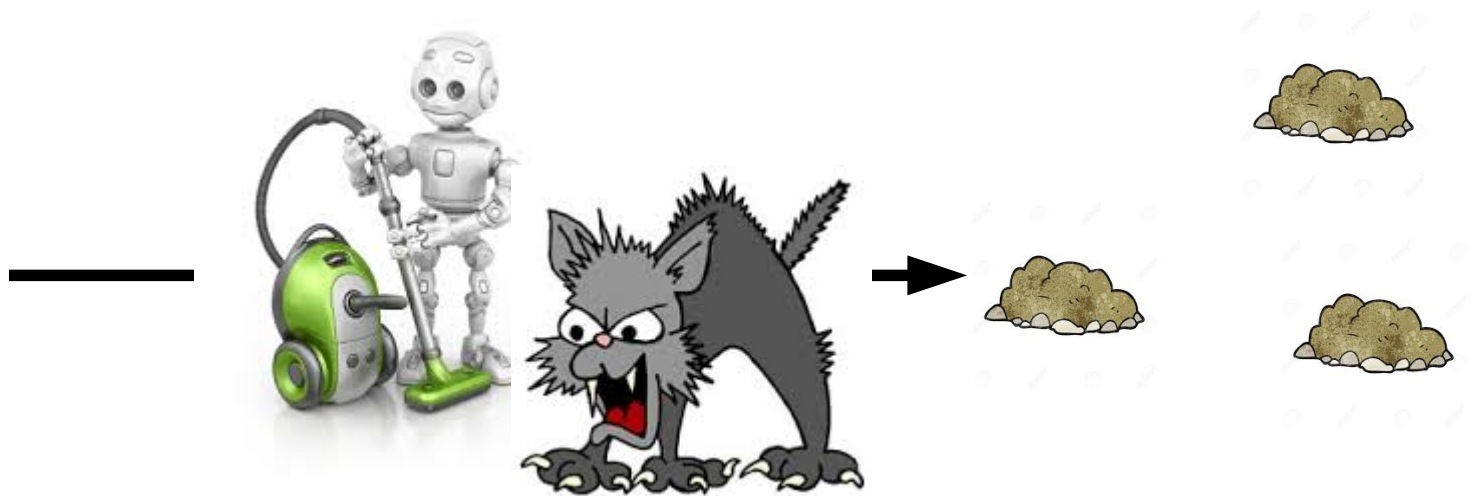
Question:

How do you specify a
reward function?

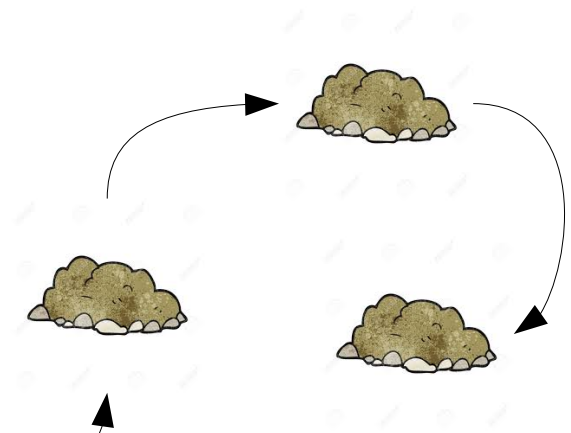
+1 for each -
bit of dirt you
pick up!



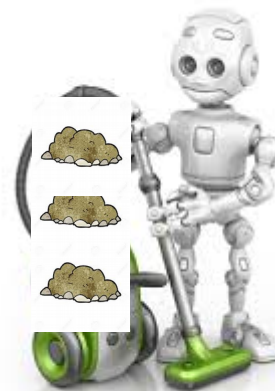
+1 for each -
bit of dirt you
pick up!
Don't bump
the cat.



+1 for each -
bit of dirt you
pick up!
Don't bump
the cat.



+1 for each -
bit of dirt you
pick up!
Don't bump
the cat.



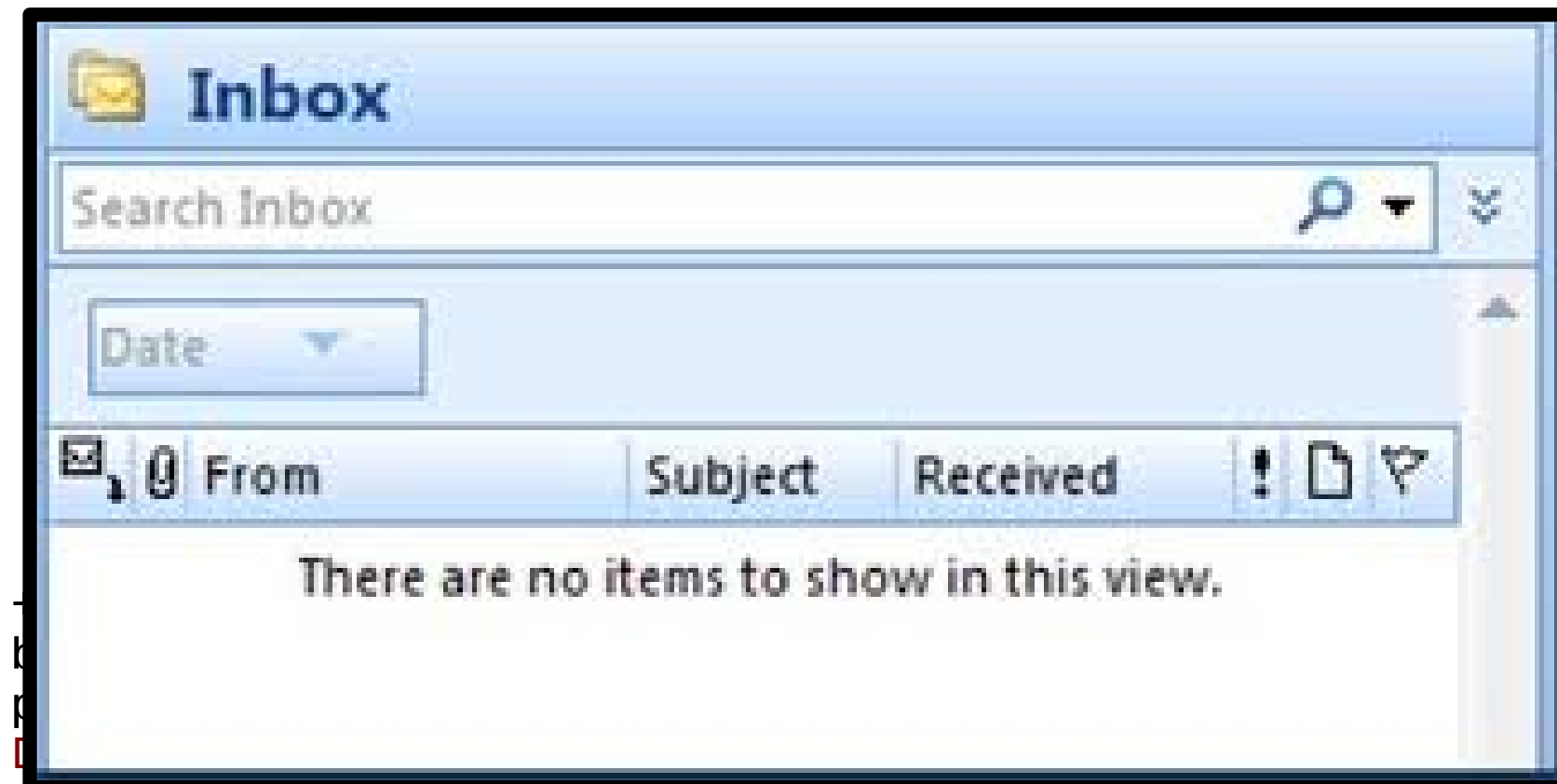
+1 for each -
bit of dirt you
pick up!
Don't bump
the cat.



+1 for each
bit of dirt you
pick up!
Don't bump
the cat.



What's the best solution to
“delete my spam”?



the cat.



What's the best solution to
“delete my spam”?


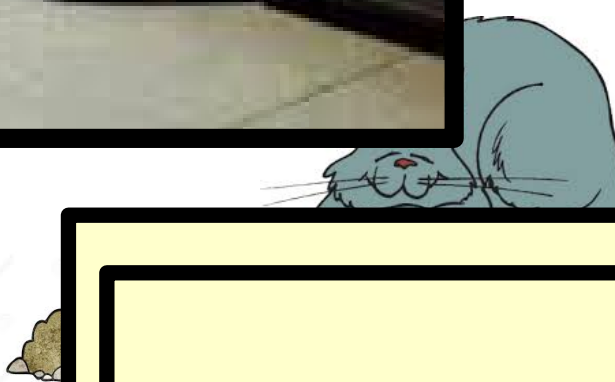
+1 for each -
bit of dirt you
pick up!
Don't bump
the cat.



What's the best solution to
“drive in your lane”?



the cat.



What's the best solution to
“drive in your lane”?

+1 for each -
bit of dirt you
pick up!
Don't bump
the cat.



What's the best solution to
“hire the CEO who'll succeed”?



What's the best solution to
“hire the CEO who'll succeed”?