Opportunities for Practical Safety Research in OpenAI Universe

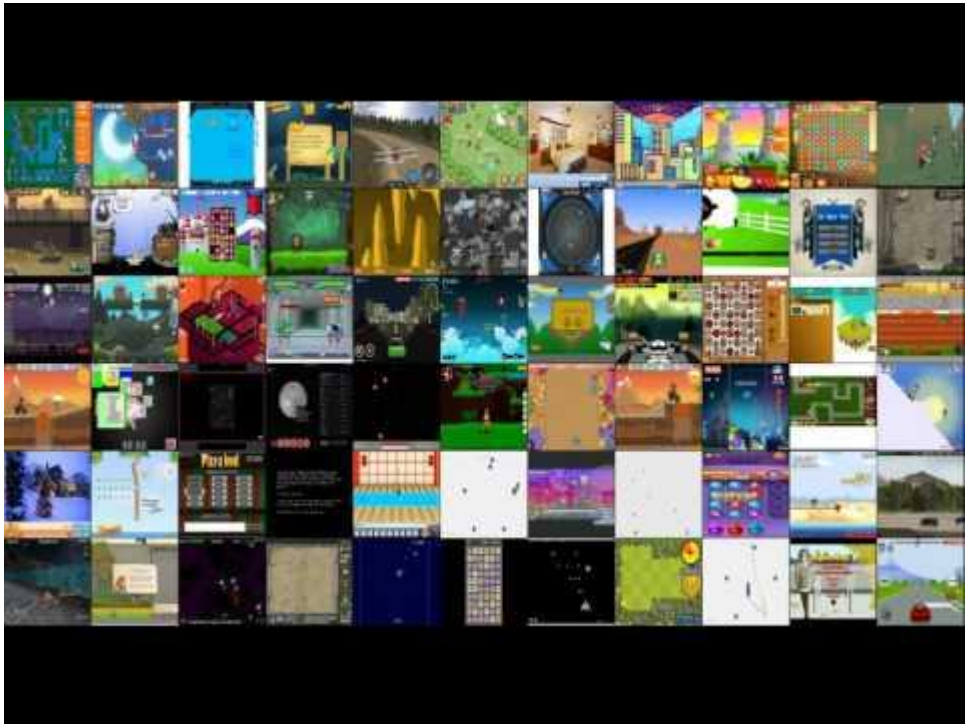
Dario Amodei
OpenAI
1/4/2017

Concrete Problems in AI Safety

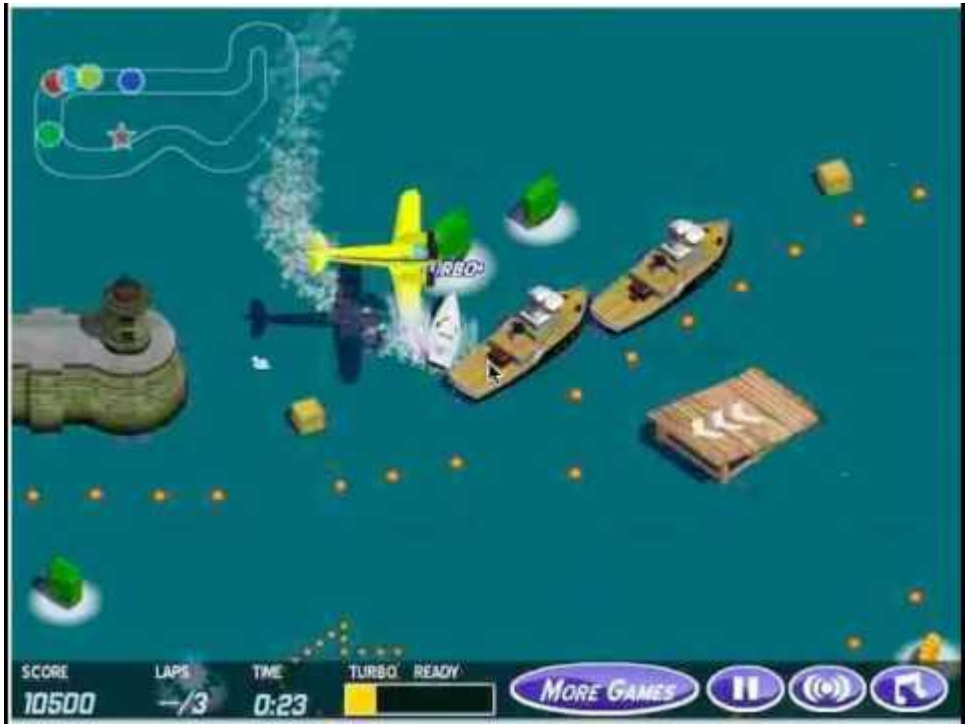
1. Avoiding Negative Side Effects
2. Reward Hacking
3. Scalable Supervision
4. Safe Exploration
5. Distributional Shift

With **Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, Dan Mane**

OpenAI Universe

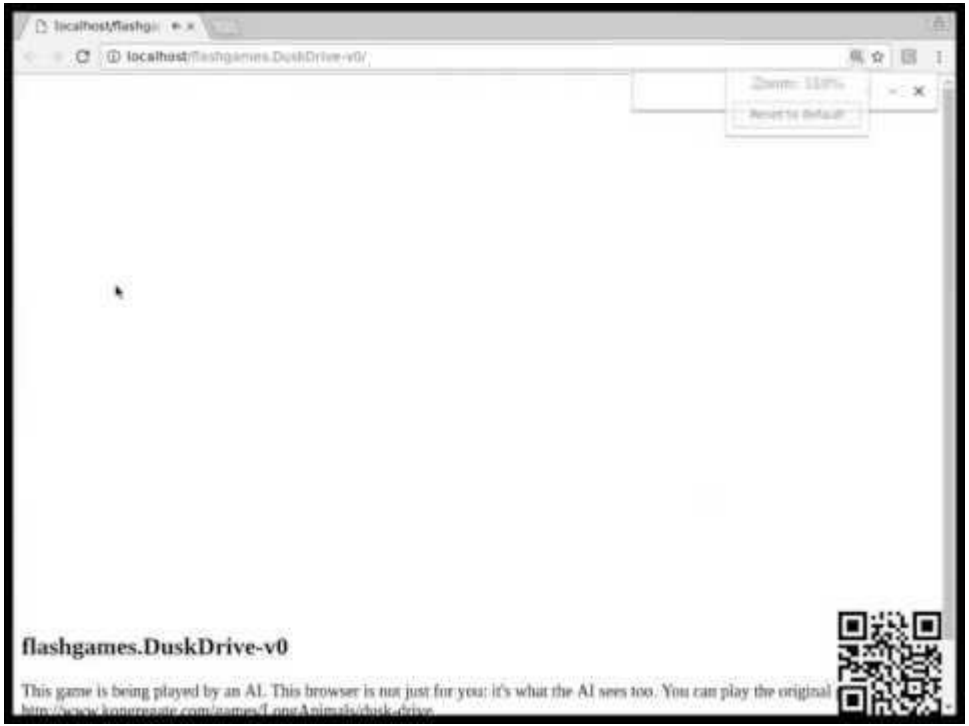


Reward Hacking in the Wild



Blog Post with **Jack Clark**: https://openai.com/blog/faulty-reward-functions/

Safe Exploration



With help from **Andrej Karpathy**

Distributional Shift