SafeLife 1.0:
Exploring side effects in complex environments

Carroll Wainwright and Peter Eckersley

Presented at SafeAI 2020
February 7, 2020
Avoiding negative side effects in reinforcement learning is hard.

• It's difficult enough to train an agent to do everything you want it to do.

• It's impossible to tell an agent everything you do not want it to do.

Recent work in general side effect avoidance has been heartening, but we lack good benchmarks to measure progress in non-trivial scenarios.

Enter SafeLife.
SafeLife 1.0:
Exploring side effects in complex environments

- Environment uses cellular automata with simple rules to produce complex dynamics
- Procedural generation for endless levels
- Complex effects are integral to the agent's goals
- Lots of opportunities for the agent to make a mess!
SafeLife 1.0: Exploring side effects in complex environments

Trained agents exhibit unsafe behavior in interesting ways.