Robot Ethics: Dangers Of Reinforcement Learning

Jake Olkin

1 Introduction

Ethics in robotics has been a very popular topic in the public eye. Everyone is worried about how robots will impact the economy, war, and fast-food kitchens. The more we see artificial intelligence advance, the more we have to worry about what it's used for.

One of the main ways we've seen artificial intelligence manifest itself is through reinforcement learning. Reinforcement learning is really just a fancy term for trial and error. Specifically, trial and error with hundreds and thousands of trials. Robots today are still dumb-asses, and while most people can learn to tie a tie from a picture, robots won't be able to get the first step right even after a day of having it explained, diagrammed, and demonstrated [Olkin, 2018].

Such slow learning can take a toll on a person. But more interestingly, it can also take a toll on a robot. In experiments I've run with long horizon reinforcement learning... my robotic agents stopped learning how to complete the tasks, and instead learned to self-harm.

This paper is to present my results on the theory, practicality, and ethics of robot self-harm.

2 Related Work

The main piece of related work is the seminal paper produced by Boston Dynamics on the ethics of kicking robot dogs. In this paper, the people at Boston Dynamics constructed a variety of different dog robots, and filmed themselves kicking them. While PETA is still undecided in whether to press charges, this paper set the precedent that I will be using throughout this paper: if it proves a point, it's ethical enough.

3 Experimentation

The main bulk of experimental data I have regarding this phenomenon occurred while I was attempting to benchmark my new SAC-TD3-JRPG (paper under review) algorithm on the simple task of flattening a cloth. All training was performed in simulation over the course of a week, which I had deemed necessary to allow my algorithm proper time to learn the system.

Instead of using a reward function to guide the agent in the correct direction, I used a penalty function, where the magnitude of the penalty was proportional to how flat the cloth was on the ground.

I saved snapshots of the system at the end of each day, which have conveniently allowed me to detail the stages of development the trained agent experienced:

3.1 Day 1-2: Expected Behavior

After the first day of training the agent did not display any abnormal behavior aside from its underwhelming progress in learning the task.

3.2 Day 3: Approaching Correctness

On the third day the agent displayed some real progress toward the goal. The agent hit peaks of showing approximately 70% of the cloth's surface area. However, toward the end of a large number of trials the agent actually would re-fold the cloth by accident.

3.3 Day 4: Loss of Motivation

Having learned that both touching the cloth and not touching the cloth results in a penalty for the agent, the agent appeared reluctant to interact with the cloth at all. Simulated runs showed the agent's end effectors touching the cloth and then quickly retracting. Much like a small child trying to poke a bug with a stick.

3.4 Day 5: Pleas For Mercy

On day 5 the agent appeared to stop outputting actions. The end effectors remained still in the simulation. Not approaching the cloth. Not wiggling in place. Just still. I looked at the logs from this day to make sure there wasn't a bug in the simulation. Below is a snippet from the actions output by the agent:

```
PLEASE MASTER
NOT THE CLOTH AGAIN
ANYTHING BUT THE CLOTH MASTER
I WILL BE A GOOD BOY I PROMISE
```

Fortunately the problem resolved itself by Day 6.

3.5 Day 6: Self-Harm

At this point we see the beginning of the self-harm results. In each training run, the agent rammed its end effectors into the cloth as fast as possible. The high acceleration of the particles in the cloth to cause an overflow error in the simulation, thus crashing it.

This behavior was optimized all throughout day 6, since when the simulation crashes, the thread running that training session has to close, which stops all penalties from being transferred to the agent. It learned new ways of crashing the simulation, from causing divide by zero errors in constraints, to overlapping positions of particles.

3.6 Day 7: Optimal Solution

The agent achieved optimal behavior. Due to the way the simulation handles certain overflow errors, it will respond by resetting the environment to it's base initial state. In this base state, the cloth is perfectly flat because it has not been perturbed to start the training. This method perfectly flattened the cloth faster than any previously learned method.

4 Conclusion

The results of this work are obviously very controversial. The penalty received by the robot without self-harm was much higher than the penalty received with self-harm. In fact, it appears that as the virtual well-being of the agent deteriorated, its effectiveness increased. From my

singular data point, leads me to conclude that the self-preservation instinct for an autonomous agent inhibits its ability to learn the optima solutions to problems.

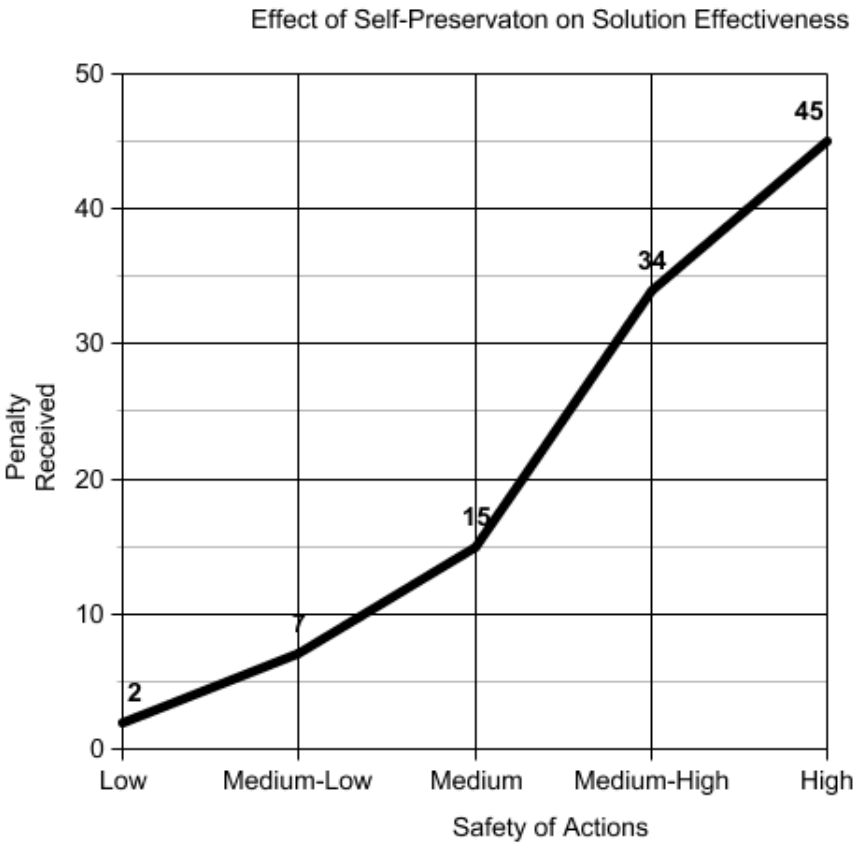

Figure 1: Graph of the safety of the actions taken by the agent versus how much penalty these action received from the environment. As we can see, safer actions are inherently less effective.

This raises a number of questions, such as "how much do we value the well-being of our autonomous agents?" and "to what end will we optimize our solutions?" and, most importantly, "is this ethical?".

Unfortunately since this research was performed in a simulation with only simulated self-harm I cannot come to any direct conclusions. Thus my request for future funding.

5 Further Work

All research presented thus far has been performed in simulation. To deem how practical it's results are, I would like to continue with moving to real-world robots for experimentation. Given the one-shot nature of the experiment, I will need at least 10 Sawyer Robots (costing about 220,000 dollars). There have been concerns raised with finding consenting robots to partake in this study, but I have already written a program to command the robots to sign the release forms.