Title: Exploring the utility of happiness for adaptive agents

Happiness is an essential part of the human conscious experience. Yet, the pursuit of happiness is not easy. Boosts in happiness are often fleeting and do not last long. Further, our happiness is not just determined by the absolute level of what we have, but also the *relative* level of what we have, sometimes leaving us miserable even in the midst of favorable circumstances. While a large literature has studied how happiness is shaped by such factors and how people can modulate their happiness, very little is known about why our behaviors are influenced by happiness in the first place. What is the evolutionary advantage of happiness? Why is happiness determined by factors like prior expectations and relative success?

In this talk, I will present our work that explores the value of endowing agents with a subjective utility function in the form of `happiness'. We utilize an intrinsically motivated reinforcement learning framework to explore the settings and environments in which endowing agents with a subjective reward function in the form of happiness is better (or worse) compared to equipping agents with a standard fitness-based reward. Across a variety of setups, we find that happiness-based agents outperform standard fitness-based agents, paving insights into understanding the evolutionary benefits and origins of happiness. At a broader level, this work opens up opportunities for developing principled computational approaches towards improving happiness and subjective well-being as well as developing robust, general-purpose reward mechanisms for intrinsically motivated artificial agents.

Reading List:

Behavioral motivation

- A computational and neural model of momentary subjective well-being
- Tonic dopamine: opportunity costs and the control of response vigor

Computational approaches

- Interactions between learning and evolution
- Where do rewards come from
- Intrinsically motivated reinforcement learning: An evolutionary perspective
- Evolutionary efficiency and happiness
- Inverse reward design
- Near-optimal reinforcement learning in polynomial time

Textbooks

- Reinforcement Learning: An Introduction

Philosophical motivation

- Happiness and thrift: When (spending) less is (hedonically) more
- Will money increase subjective well-being?
- Experienced Utility and Objective Happiness: A Moment-Based Approach