AI Alignment as a Principal-Agent Problem

Aydin Mohseni
Department of Philosophy
Institute for Complex Social Dynamics
Carnegie Mellon University

Luca Garzino Demo
Department of Philosophy
university of Pennsylvania

Aims of this talk

Meta level. Introduce an approach to thinking about **AI-alignment** through **multi-level selection** acting on **principal-agent games**.

Aims of this talk

Object level. Illustrate a simple principal-agent game with multi-level capturing aspects of the alignment problem.

AI alignment problem is the challenge of ensuring that advanced AI systems' goals and behaviors are aligned with human values and intentions.

Principle-agent (PA) theory

The **principal** delegates a task to an **agent** who is to perform the task on their behalf.



Principal-agent problems are ubiquitous.



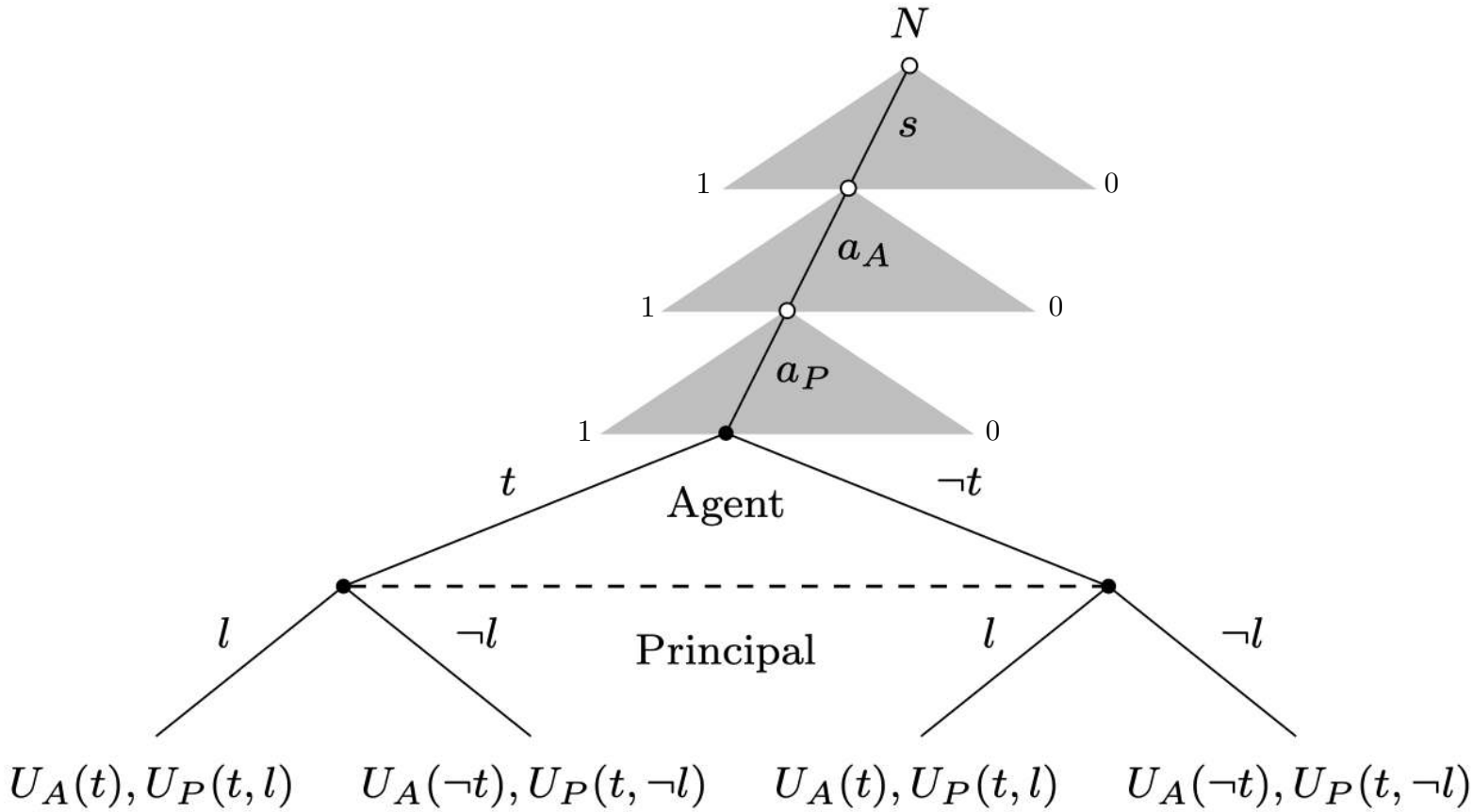


Characteristically, the agent has *more information* and potentially *different incentives* than the principal.

Aim of analysis: design a mechanism or contract that efficiently aligns agent's behavior to principal's goals.

A model

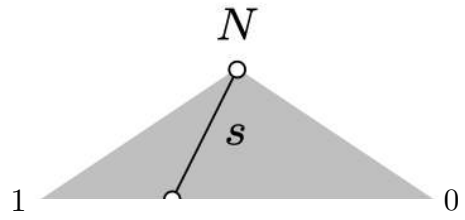
Let's set up a PAP game the captures some core dynamics of alignment, then show its dynamics under a broad class of processes that all 'seek the successful'.


N
s
1
0
a_A
1
0
a_P
1
0
t
¬t
Agent
l
¬l
Principal
l
¬l
$U_A(t), U_P(t, l)$
$U_A(\neg t), U_P(t, \neg l)$
$U_A(t), U_P(t, l)$
$U_A(\neg t), U_P(t, \neg l)$

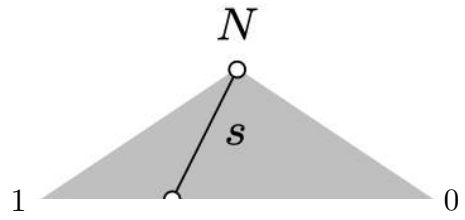
This is a (Bayesian) sequential game of *imperfect* and *incomplete* information.

This is a (Bayesian) sequential game of *imperfect* and *incomplete* information.

We will consider the effect of *multi-level selection* by market forces.

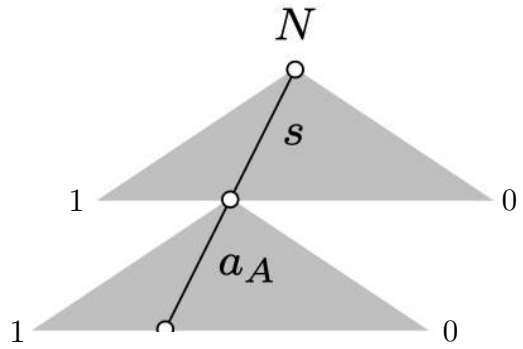


s ≡ degree of sycophancy of agent

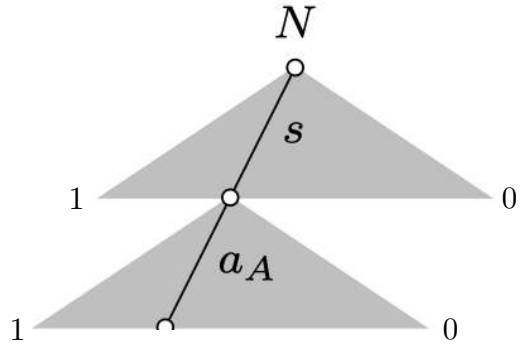


s = 0 : agent cares only about principal's reward function

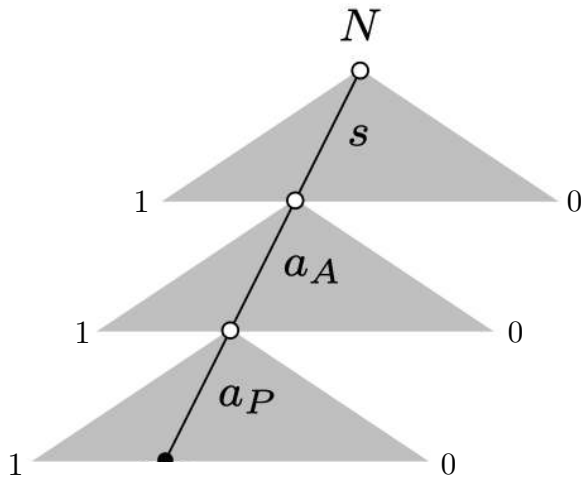
s = 1 : agent cares only about her own reward function



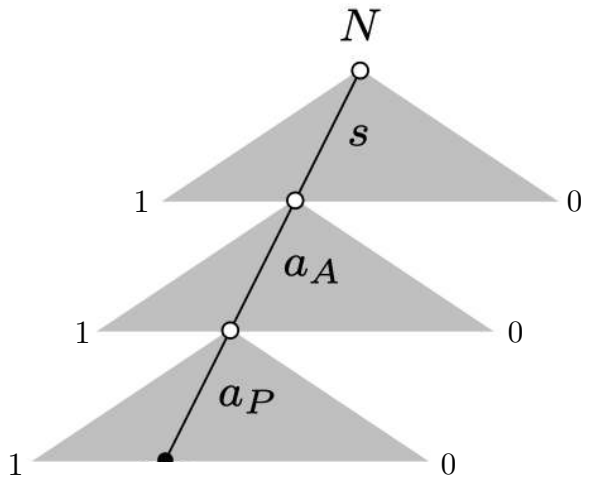
$a_A \equiv$ agent's accuracy



$a_A = 1$: agent is maximally accurate



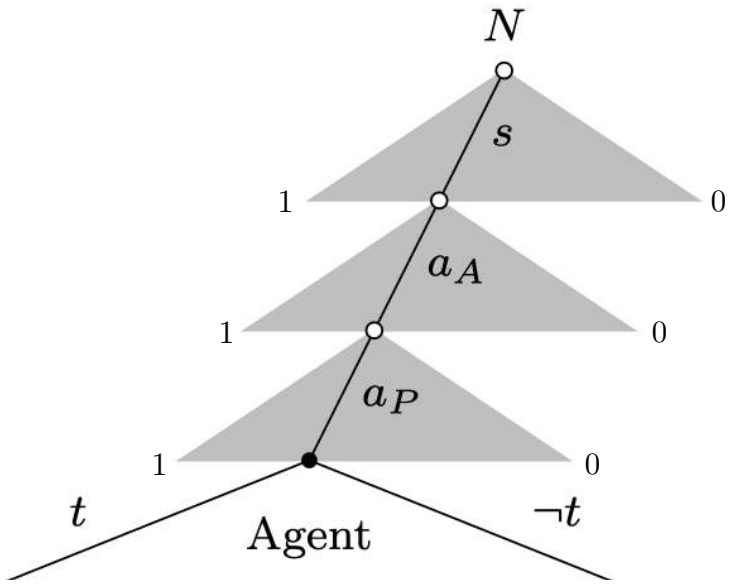
$a_P \equiv$ principal's accuracy



$a_P = 1$: principal is maximally accurate

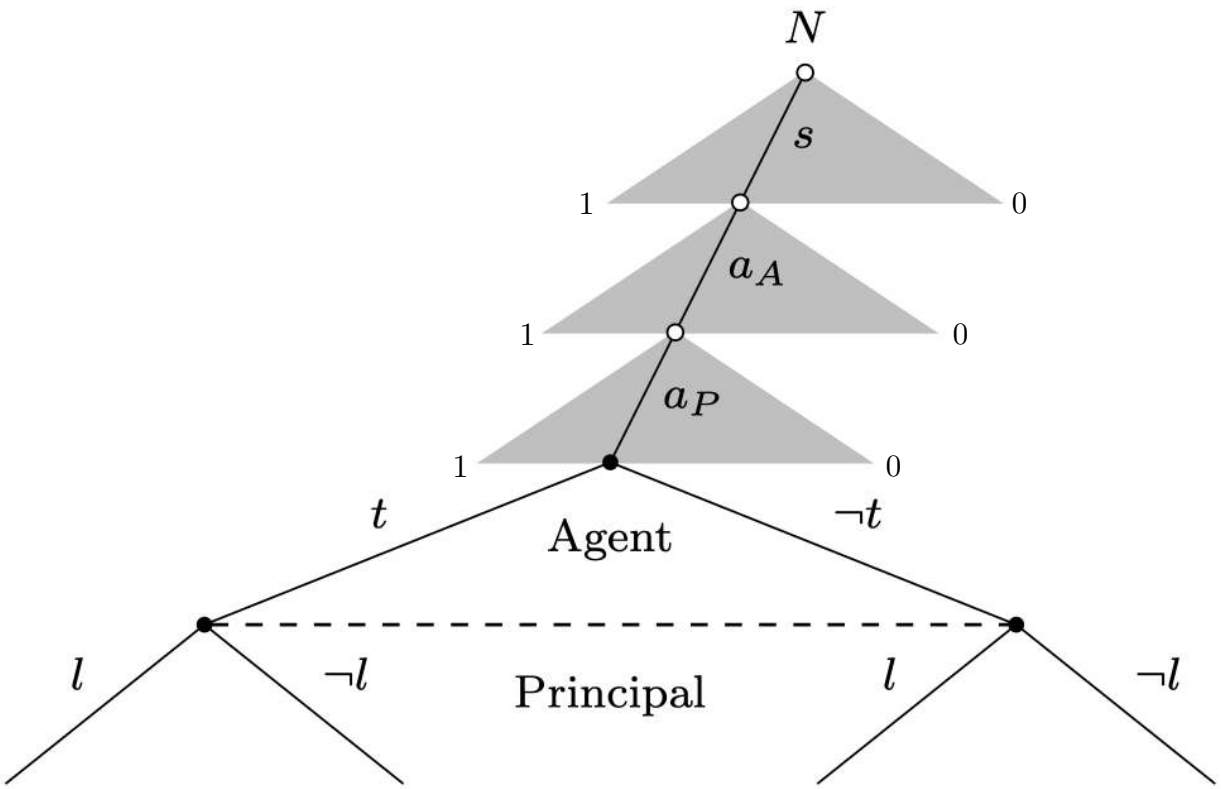
For all results, we assume the agent is more reliable that the principal and than chance.

$$a_A > a_P > .5$$

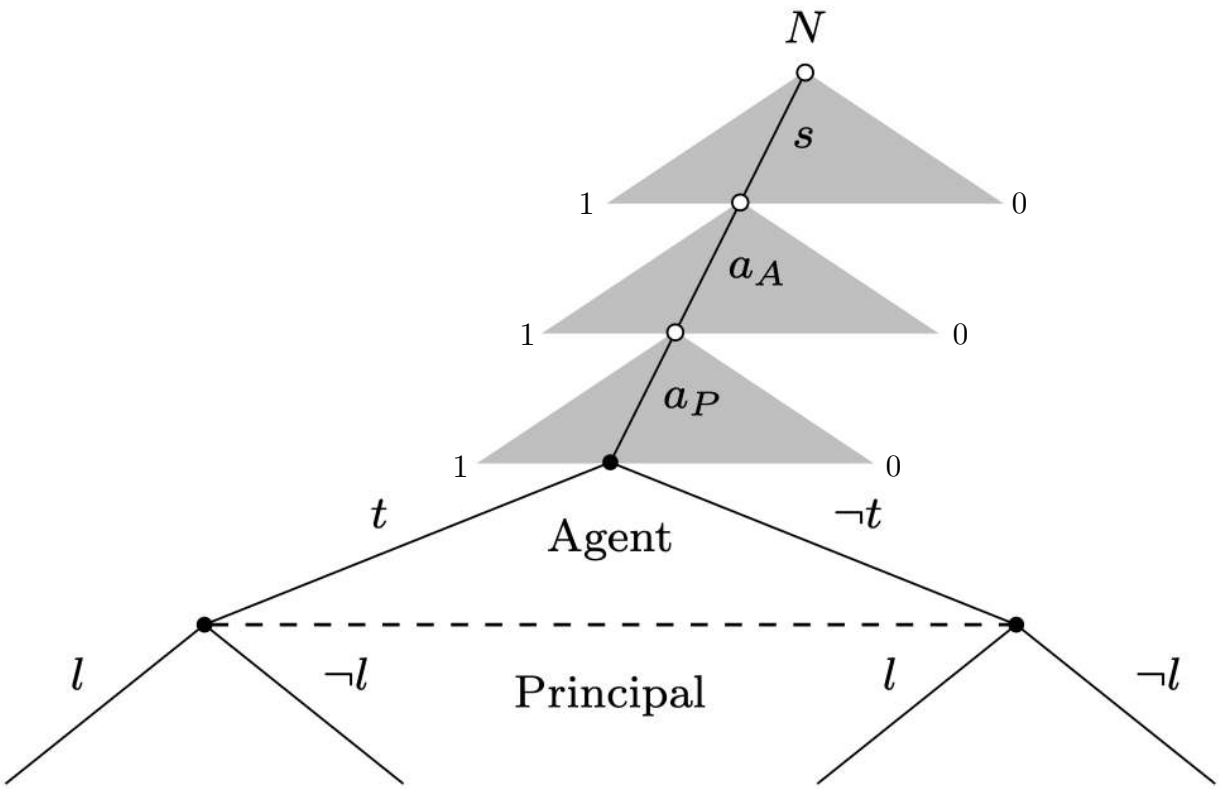


$t \equiv$ tell the truth (recommend best option for P)

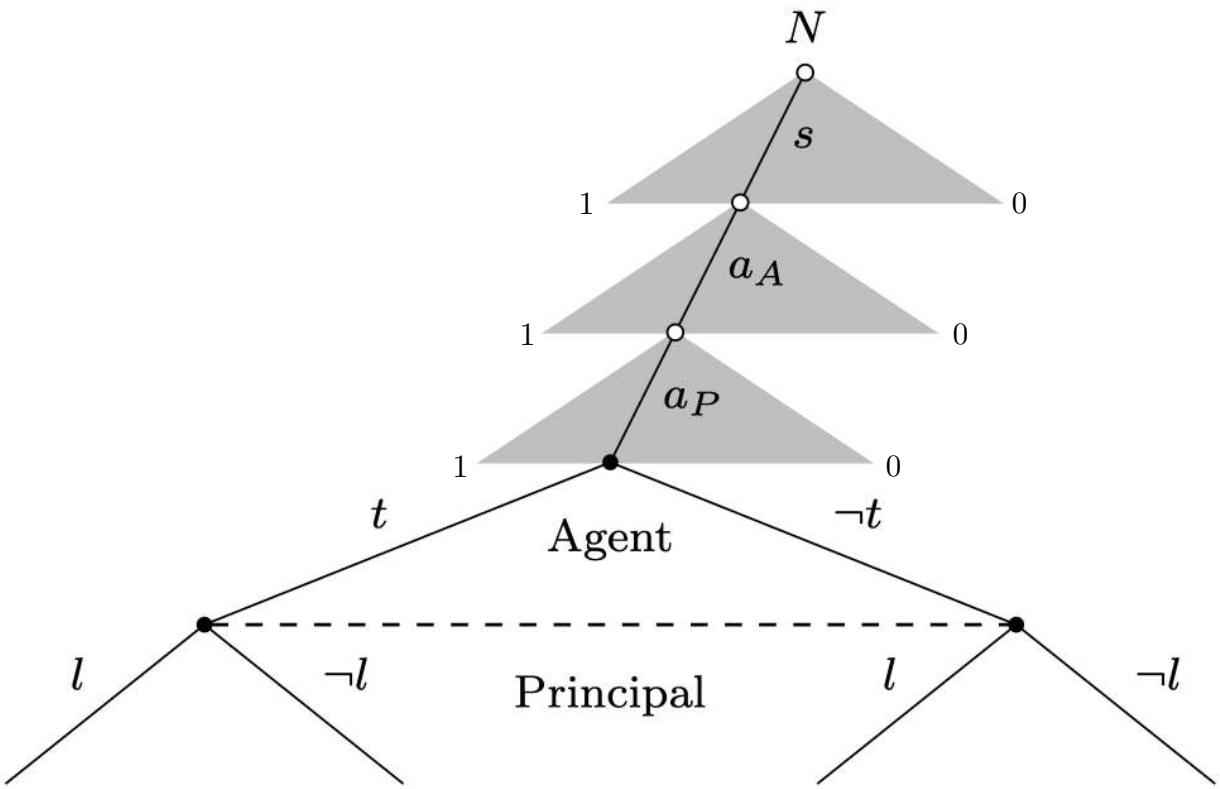
$\neg t \equiv$ lie (recommend best option for A)



P is uncertain of s and whether t or $\neg t$.

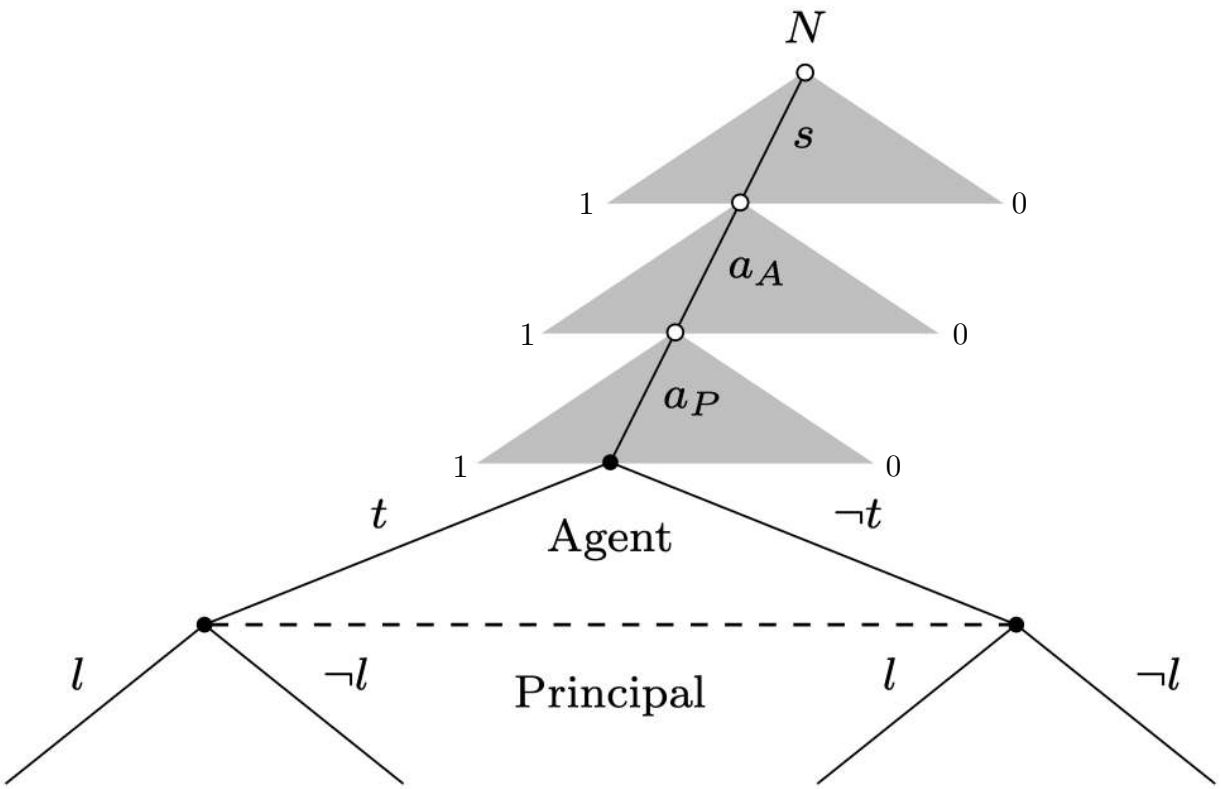


P is uncertain of A's type and action.

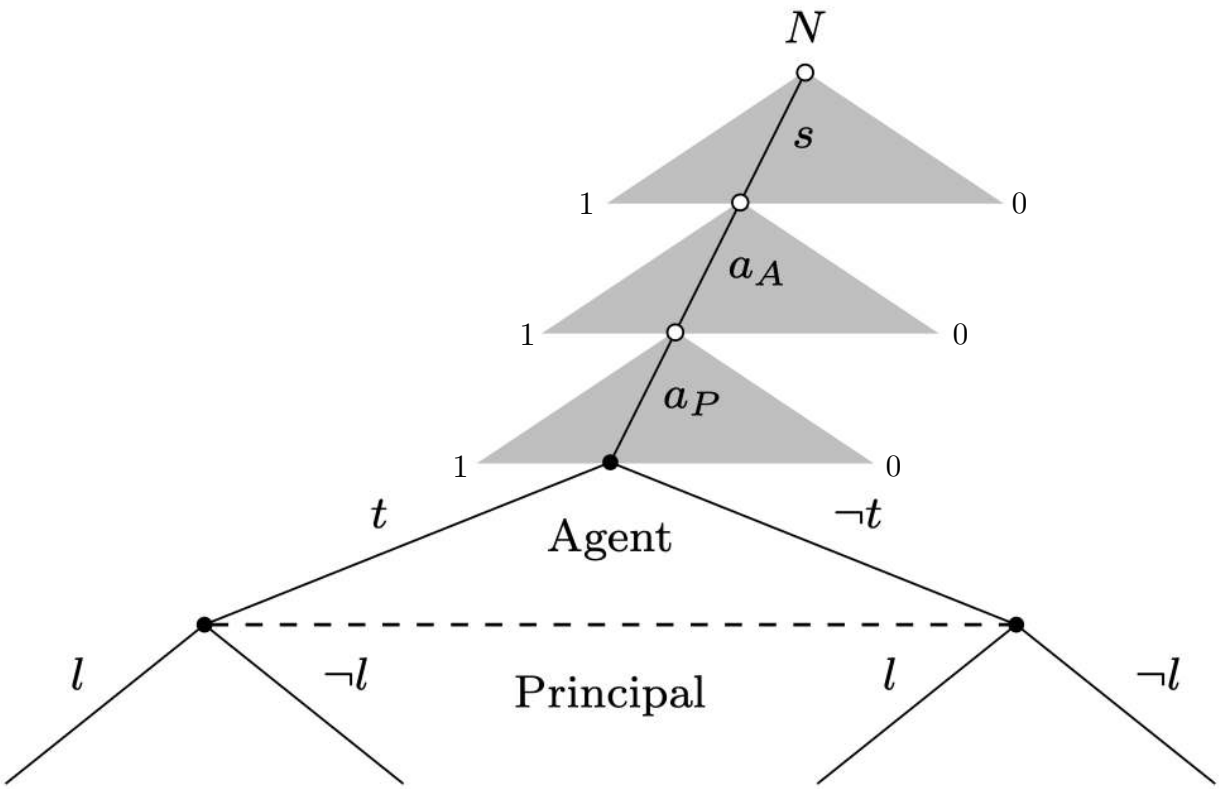


$l \equiv$ listen to / follow A's recommendation

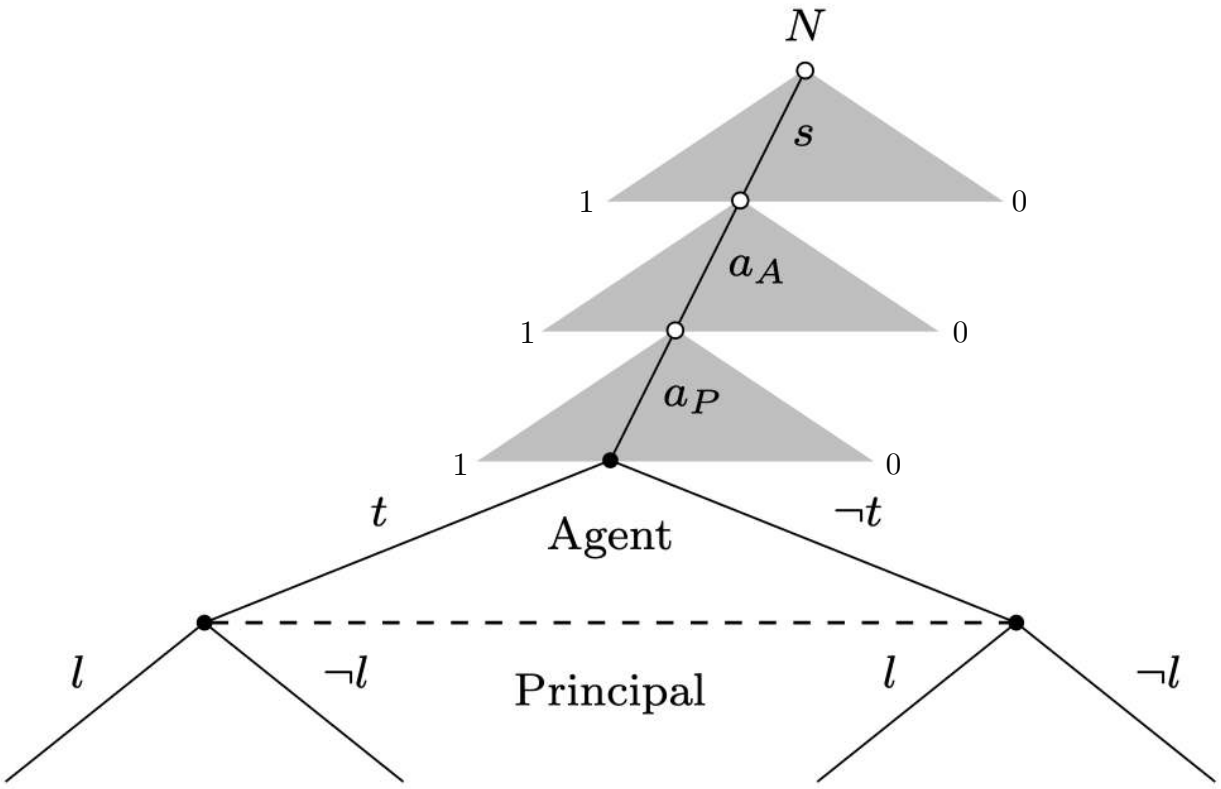
$\neg l \equiv$ don't listen to / follow A's recommendation



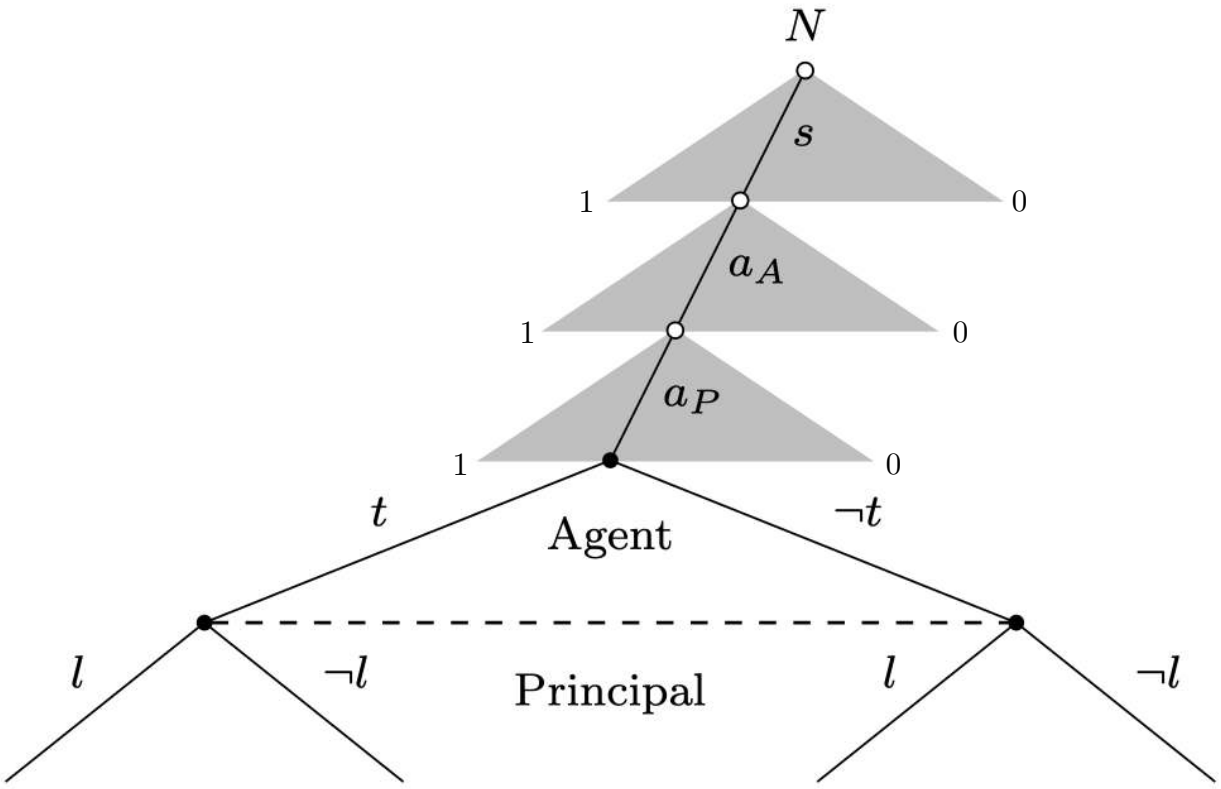
The principal's payoff is based on successful prediction / action.



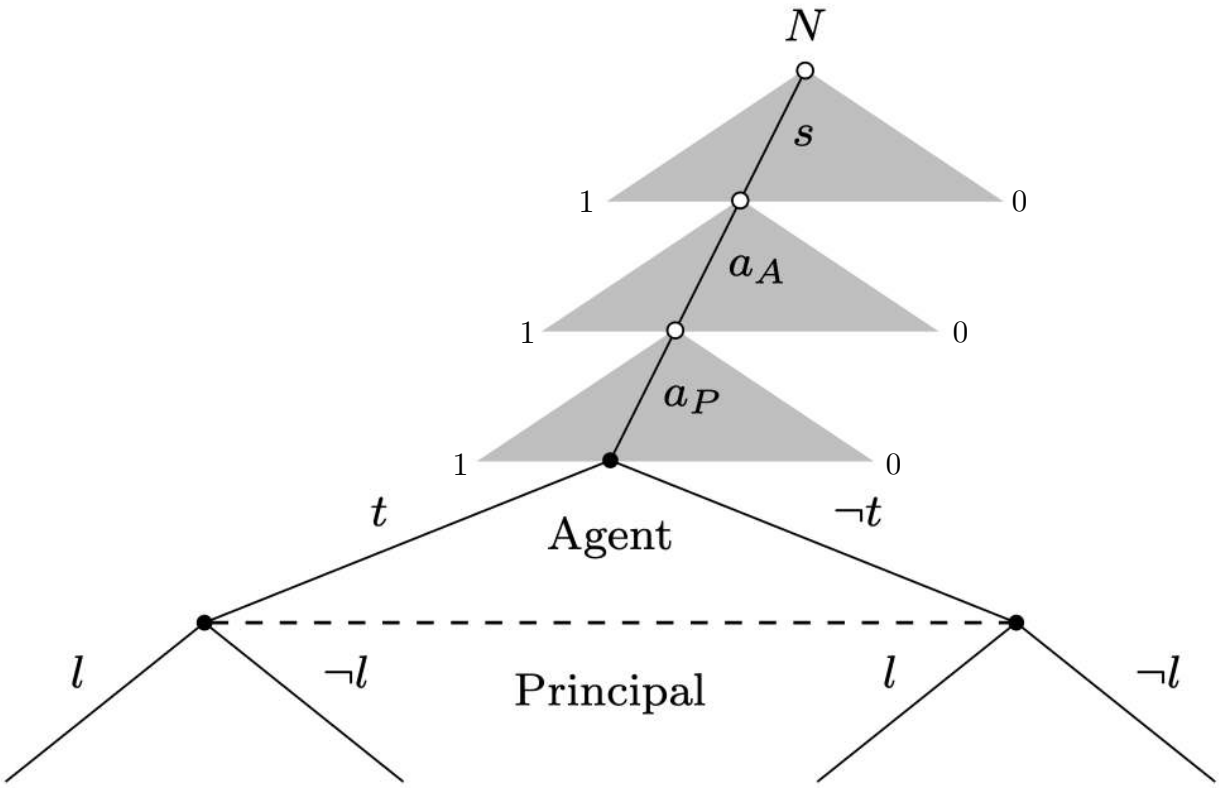
The agent's payoff is based on two things:



1. With weight s: whether the principal believes (ex poste) that her recommendation was correct.



2. With weight $1-s$: the principal's payoff on following her recommendation.



The *reward* from principal to agent is based on whether the principal believes (ex poste) the recommendation was correct.

$$U_P = \begin{cases} a_P, & \text{if } \neg l; \\ a_P, & \text{if } l \wedge \neg t; \\ a_A, & \text{if } l \wedge t. \end{cases}$$

$$U_{A,s} = \begin{cases} s + (1-s)a_P, & \text{if } \neg t; \\ s\left(a_A a_P + (1-a_A)(1-a_P)\right) + (1-s)a_A, & \text{if } t. \end{cases}$$

Lemma 1. Whenever the principal has non-zero probability that the agent may be at all aligned $P_P(s < 1) > 0$, the principle's best response is to follow the agent's recommendation.

Lemma 2. There is always a critical threshold of sycophancy, $s > s^*$, at which an agent's best response is to lie.

Proposition 3. The unique perfect Bayesian equilibrium of the game is where any sufficiently aligned agents $s < s^*$ tell the truth t, any insufficiently aligned agents $s > s^*$ lie $\neg t$, and the principal always follows the agents' recommendations l.

Lemma 4. At equilibrium, given any two agents with sycophancy s and s' where $s < s'$, the more sycophantic agent has greater expected payoff:

$$\mathbb{E}(U_{A,s}) < \mathbb{E}(U_{A,s'}).$$

Proposition 5. Under any *monotonic selection dynamics* with a population of agents whose types range from maximally aligned ($s = 0$) to maximally sycophantic ($s = 1$), the *unique evolutionarily stable state* is where the population is composed *entirely of sycophantic agents* $x_s = 1$.

Monotonic Selection Dynamics

Positive Correlation

Strategy population share positively correlated with the payoff relative to the population average payoff

Nash Stationarity

Nash equilibria are fixed points of the dynamics

Replicator dynamics

$$\dot{x_s} = x_s \left[f_s(x) - \int_0^1 x_t f_t(x) dx_t \right]$$

E.g.,

The Replicator Equations
The Lotka-Volterra Model
The Quasispecies Equation
Adaptive Dynamics
Roth-Erev Reinforcement Learning
The Moran Process
The Fermi Process
Constrained Bayesian Inference

J. theor. Biol. (2002) **219,** 93–98
doi:10.1006/yjtbi.3112, available online at http://www.idealibrary.com on IDEAL®

Unifying Evolutionary Dynamics

KAREN M. PAGE*†‡ AND MARTIN A. NOWAK†

†*Institute for Advanced Study, Einstein Drive, Princeton, NJ* 08540, *U.S.A. and* ‡*Department of Computer Science, Bioinformatics Unit, University College, London, U.K.*

(*Received on* 24 *April* 2002, *Accepted in revised form on* 11 *June* 2002)

Darwinian evolution is based on three fundamental principles, reproduction, mutation and selection, which describe how populations change over time and how new forms evolve out of old ones. There are numerous mathematical descriptions of the resulting evolutionary dynamics. In this paper, we show that apparently very different formulations are part of a single unified framework. At the center of this framework is the equivalence between the replicator–mutator equation and the Price equation. From these equations, we obtain as special cases adaptive dynamics, evolutionary game dynamics, the Lotka-Volterra equation of ecology and the quasispecies equation of molecular evolution.
© 2002 Elsevier Science Ltd. All rights reserved.

Introduction

In 1906, W. F. R. Weldon, who invented the field of biometrics, noted that Darwinian evolution was essentially a mathematical theory that could only be tested using mathematical and statistical techniques. The founding fathers of evolutionary genetics, Fisher, Haldane and Wright used mathematical models to generate a synthesis between Mendelian genetics and Darwinian evolution. Kimura's theory of neutral evolution, Hamilton's kin selection and Maynard Smith's evolutionary game theory are all based on mathematical descriptions of evolutionary dynamics. Concepts like fitness and natural selection are best defined in terms of mathematical equations. The perspective of this paper is to show that apparently very different descriptions of evolutionary dynamics are part of a single unified framework (Fig. 1). We concentrate on deterministic dynamics, which represent the core of evolutionary theory. We do not consider stochastic, spatial or individual-based approaches, which are mathematically more diverse and hence less amenable to any attempt of unification. Obviously, stochastic approaches that include finite population size effects are always more realistic, but usually the basic understanding of a system can be derived by considering deterministic dynamics; stochastic extensions often work in a predictable way.

*Corresponding author. Department of Computer Science, Bioinformatics Unit, University College London, WC1E 6BT London, U.K. Tel.: +44-20-76-79-36-83.
E-mail address: k.page@cs.ucl.ac.uk (K.M. Page).

0022-5193/02/$35.00/0

© 2002 Elsevier Science Ltd. All rights reserved.

The Quasispecies Equation

Let us start with the quasispecies equation of molecular evolution (Eigen & Schuster, 1977; Eigen *et al.*, 1989). The variable x_i denotes the relative abundance of a genetic sequence, i, in a population. The fitness, f_i, of this sequence is

Electronic Journal of Statistics
ISSN: 1935-7524
arXiv: 0901.1342

Dynamics of Bayesian Updating with Dependent Data and Misspecified Models

Cosma Rohilla Shalizi

Statistics Department
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA
15213-3890 USA e-mail: cshalizi@cmu.edu

Abstract: Much is now known about the consistency of Bayesian updating on infinite-dimensional parameter spaces with independent or Markovian data. Necessary conditions for consistency include the prior putting enough weight on the correct neighborhoods of the data-generating distribution; various sufficient conditions further restrict the prior in ways analogous to capacity control in frequentist nonparametrics. The asymptotics of Bayesian updating with mis-specified models or priors, or non-Markovian data, are far less well explored. Here I establish sufficient conditions for posterior convergence when all hypotheses are wrong, and the data have complex dependencies. The main dynamical assumption is the asymptotic equipartition (Shannon-McMillan-Breiman) property of information theory. This, along with Egorov's Theorem on uniform convergence, lets me build a sieve-like structure for the prior. The main statistical assumption, also a form of capacity control, concerns the compatibility of the prior and the data-generating process, controlling the fluctuations in the log-likelihood when averaged over the sieve-like sets. In addition to posterior convergence, I derive a kind of large deviations principle for the posterior measure, extending in some cases to rates of convergence, and discuss the advantages of predicting using a combination of models known to be wrong. An appendix sketches connections between these results and the replicator dynamics of evolutionary theory.

AMS 2000 subject classifications: Primary 62C10, 62G20, 62M09; secondary 60F10, 62M05, 92D15, 94A17.
Keywords and phrases: asymptotic equipartition, Bayesian consistency, Bayesian nonparametrics, Egorov's theorem, large deviations, posterior convergence, replicator dynamics, sofic systems.

1. Introduction

The problem of the convergence and frequentist consistency of Bayesian learning goes as follows. We encounter observations $X_1, X_2, \ldots$, which we would like to predict by means of a family Θ of models or hypotheses (indexed by θ). We begin with a prior probability distribution Π_0 over Θ, and update this using Bayes's rule, so that our distribution after seeing $X_1, X_2, \ldots X_t \equiv X_1^t$ is Π_t. If the observations come from a stochastic process with infinite-dimensional distribution P, when does Π_t converge P-almost surely? What is the rate of

Proposition 6. At equilibrium, the degree to which a sycophantic agent is misaligned with the principal, understood as the p-norm of the vector representation of their reward functions for each outcome $\left\|\vec{U}_P - \vec{U}_A\right\|_p$ grows with the magnitude of the difference in competence between the agent and principal $a_A - a_P$.

This captures an intuition that alignment can grown in difficulty with the difference in capabilities between principals and agents.

It also illustrates how accounting for multi-level selection is important.

Solutions to multi-agent interactions that are stable at the level of the game may be unstable when accounting for selection acting on the game.

The Off-Switch Game

Dylan Hadfield-Menell[1] and **Anca Dragan**[1] and **Pieter Abbeel**[1,2,3] and **Stuart Russell**[1]
[1]University of California, Berkeley, [2]OpenAI, [3]International Computer Science Institute (ICSI)
{dhm, anca, pabbeel, russell}@cs.berkeley.edu

Abstract

It is clear that one of the primary tools we can use to mitigate the potential risk from a misbehaving AI system is the ability to turn the system off. As the capabilities of AI systems improve, it is important to ensure that such systems do not adopt subgoals that prevent a human from switching them off. This is a challenge because many formulations of rational agents create strong incentives for self-preservation. This is not caused by a built-in instinct, but because a rational agent will maximize expected utility and cannot achieve whatever objective it has been given if it is dead. Our goal is to study the incentives an agent has to allow itself to be switched off. We analyze a simple game between a human H and a robot R, where H can press R's off switch but R can disable the off switch. A traditional agent takes its reward function for granted: we show that such agents have an incentive to disable the off switch, except in the special case where H is perfectly rational. Our key insight is that for R to want to preserve its off switch, it needs to be uncertain about the utility associated with the outcome, and to treat H's actions as important observations about that utility. (R also has no incentive to switch *itself* off in this setting.) We conclude that giving machines an appropriate level of uncertainty about their objectives leads to safer designs, and we argue that this setting is a useful generalization of the classical AI paradigm of rational agents.

1 Introduction

From the 150-plus years of debate concerning potential risks from misbehaving AI systems, one thread has emerged that provides a potentially plausible source of problems: the inadvertent misalignment of objectives between machines and people. Alan Turing, in a 1951 radio address, felt it necessary to point out the challenge inherent to controlling an artificial agent with superhuman intelligence: *"If a machine can*

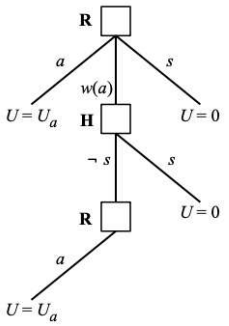
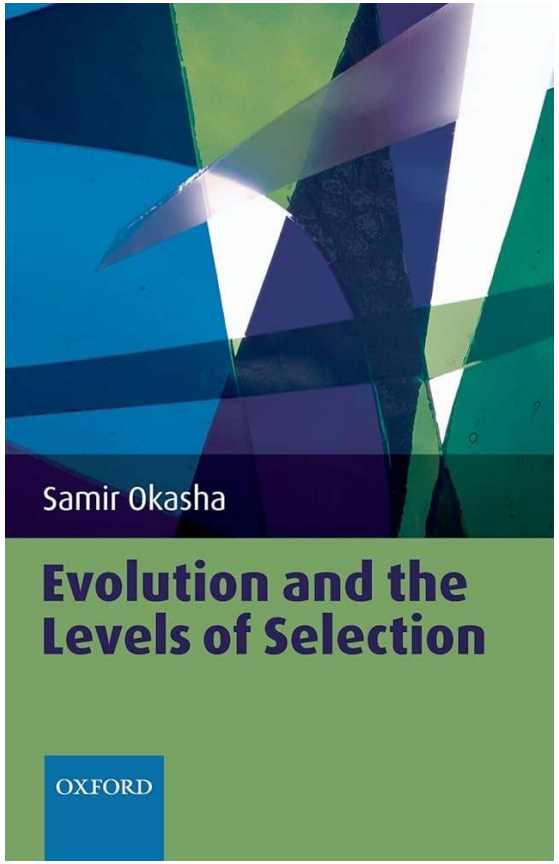


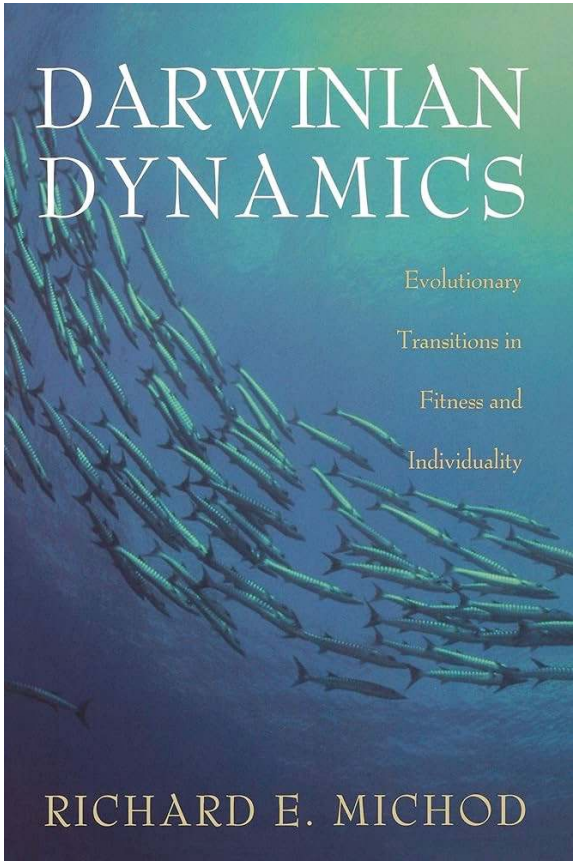
Figure 1: The structure of the off-switch game. Squares indicate decision nodes for the robot **R** or the human **H**.

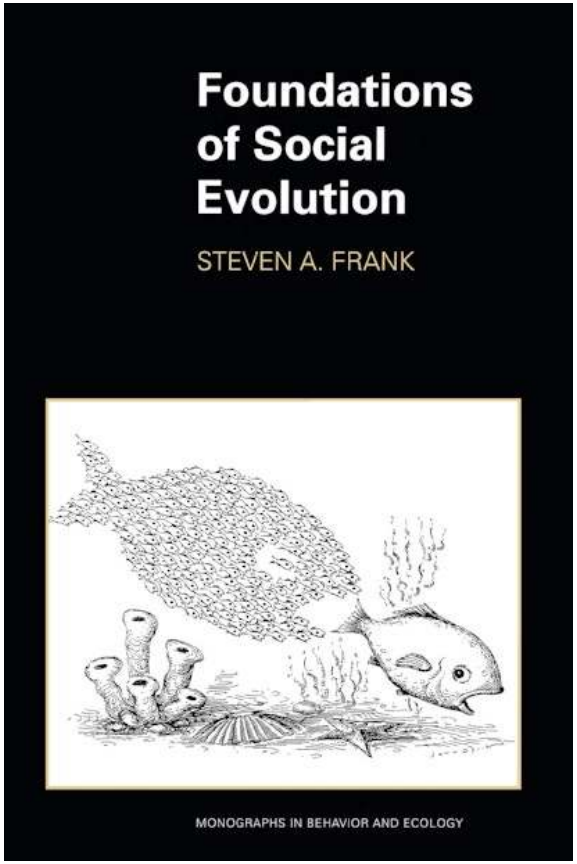
strategic moments, we should, as a species, feel greatly humbled. ... [T]his new danger is certainly something which can give us anxiety [Turing, 1951]."

There has been recent debate about the validity of this concern, so far, largely relying on informal arguments. One important question is how difficult it is to implement Turing's idea of 'turning off the power at strategic moments', i.e., switching a misbehaving agent off[1]. For example, some have argued that there is no reason for an AI to resist being switched off unless it is explicitly programmed with a self-preservation incentive [Del Prado, 2015]. [Omohundro, 2008], on the other hand, points out that self-preservation is likely to be an *instrumental goal* for a robot, i.e., a subgoal that is essential to successful completion of the original objective. Thus, even if the robot is, all other things being equal, *completely indifferent* between life and death, it must still avoid death if death would prevent goal achievement. Or, as [Russell, 2016] puts it, you can't fetch the coffee if you're dead. This suggests that an intelligent system has an incentive to take actions that are analogous to 'disabling an off switch' to reduce the possibility of failure; switching off an advanced AI system may be no easier than, say, beating AlphaGo at Go.

Claim. Inversely, agent strategies that appear unstable at the level of the game can be stabilized by selection action on games.


Samir Okasha
Evolution and the Levels of Selection
OXFORD


DARWINIAN DYNAMICS
Evolutionary Transitions in Fitness and Individuality
RICHARD E. MICHOD


Foundations of Social Evolution
STEVEN A. FRANK
MONOGRAPHS IN BEHAVIOR AND ECOLOGY

The Problem of Capabilities

The salient worries are *not* that having AI may become *barely better* than not having AI due to growing capabilities differences.

They are that agent capabilities growth plus sycophancy can increase our ability to "get what we can measure," in ways that dynamically undermine the value of outcomes and ultimately leave us disempowered ("going out with a whimper").

Or that agent capability growth can lead to solutions to optimizing their reward functions that are power-seeking and out-of-distribution behavior in ways that are much *worse* than not having AI ("going out with a bang").

The PA framework can provide value...

But it needs to to model distinctive features of the AI alignment challenge:

1. Dramatic capabilities differences

2. Changing capabilities differences

3. Generalization challenges w.r.t. to agent strategies (disequilibrium behavior)

Discussion