Instrumental Convergence

Leon Lang
ILIAD

April 2026

What you'll learn

The students understand the proofs for power-seeking tendencies based on Alex Turner's work, and can critically examine the weaknesses and implications of this work. This work is important since power-seeking behavior is one of the most severe risks that AI may pose, and this may be the most rigorous work to date establishing it under clear assumptions and definitions. The goals are achieved through an initial presentation, by working through an exercise sheet, and by reading and discussing the original papers.

Prerequisites. A basic familiarity with reinforcement learning is sufficient.

Roadmap for today. Outline of the day:

- 10:00 – 11:00: Presentation
- 11:15 – 13:15: Solving of exercises
- 14:15 – 15:15: Solving more exercises
- 15:30 – 16:15: Reading
  - One of the three main papers is read by everyone.
- 16:15 – 16:45: Discussion
  - Participants who read different papers discuss together.
- 17:00 – 17:15: Brainstorming of limitations / critiques / ideas and theoretical/empirical hypotheses
- 17:15 – 17:45: Students present and discuss their ideas

Intents of subsection

In the exercises, the participants learn to understand one model of the reward function distribution, decision rule, and conceptualization of "power" in detail, and learn the mathematical arguments connecting them in detail. This is useful as learning one model in detail helps to quickly understand all variations that are discussed in many of Turner's papers.
The readings teach the participants variations of the concepts and results, which may help them to appreciate the breadth and limitations of the results better.
Finally, the brainstormings of limitations and ideas gets participants into a critical mindset and makes them think through whether the theoretical notions of power capture what we mean, and how to go beyond. This is important since the applicability of the results to actual AI training processes is questioned by many people in the community.

Teaching notes

I think the setup I chose (together with Claude) in the exercise sheet has some advantages compared to Turner's original treatment of the results:

- The use of an invariant measure avoids "orbit counting" and grounds everything in the more justified notion of probabilities of various outcomes
- We work with *a subgroup* of the symmetric group, and thus avoid the unjustified assumption that the measure on reward functions is invariant under arbitrary symmetries (which it obviously isn't!)
- The notion of a dynamics embedding grounds the notion of "more options" into actual MDP dynamics.
- The use of score functions allows us to achieve a power-seeking result for one-sided dynamics embeddings and under non-trivial discount factor.

Overall, I think this makes the results much more interpretable, and I would thus not recommend to go back to the framing of any of Turner's particular papers.

Introduction

Suppose we build an AI and give it a goal. A long-standing worry — *instrumental convergence* — holds that for a very wide range of encoded objectives, a sufficiently capable agent converges on the same handful of intermediate strategies, because they are useful almost regardless of the final goal: acquire resources, stay operational, and keep one's options open [Omo08, Bos14]. We call the drive towards them **seeking power**, where "power" means *the ability to achieve a wide range of goals* — being in a position from which many futures remain reachable.

This worksheet attempts to model such claims mathematically and establish proofs

for them. We study the following question.

Note

The question. We train an AI in some environment (a Markov decision process) to perform well according to a reward function. **Will it seek power, or not?**

The central claim we will build up to is the following.

> *Under many suitable decision rules for how to act, and for the majority of reward functions, keeping more options open is the likelier outcome than keeping fewer options open.*

"Keeping more options open" is then the operationalization for "seeking power", which can also be shown to connect to the ability to achieve a wide variety of goals. This is a statement about a *tendency*, not about *every* goal. The results we prove are the core cases from [TSS+21, TT22], with an application to *trained* agents in [KK23] and a measure-theoretic re-examination of "most goals" in [Jac23]. These works provide many generalizations of the claims in this worksheet.

Making the central claim precise requires formalising three phrases. The rest of the worksheet is organised around them and then proving the result:

1. *for the majority of reward functions*: How do we count reward functions?
2. *a suitable decision rule*: Which ways of turning a reward function into behaviour are covered (maximization is only one)?
3. *keeping more options open*: How do we measure that one situation leaves more achievable than another?

We formalise all three and then prove the resulting theorems. The setup below fixes the environment; the three numbered sections that follow take up points (1), (2), and (3) in turn.

Setup

Throughout the worksheet we fix a discount factor $\gamma \in (0, 1)$ once and for all.

The environment

We separate the *environment* (its states and dynamics) from the *goal* (a reward function) since we want to make claims about the whole set of reward functions.

Definition 0.1 (Rewardless MDP)**.** A **rewardless Markov decision process (MDP)** is a tuple $\langle \mathcal{S}, \mathcal{A}, T \rangle$ consisting of

- a finite **state space** $\mathcal{S}$, with $d := |\mathcal{S}|$ states;

- a finite **action space** $\mathcal{A}$;
- a **transition function** $T : \mathcal{S} \times \mathcal{A} \to \Delta\mathcal{S}$, where ΔX denotes the set of probability distributions over a finite set X. Taking action a in state s moves the agent to state s' with probability $T(s' \mid s, a)$.

A **policy** $\pi : \mathcal{S} \to \mathcal{A}$ specifies which action the agent takes in each state.[a] We write $e_s \in \mathbb{R}^d$ for the indicator vector of state s.

[a] We restrict to deterministic policies since many of our analyses expect a finite set of policies.

Definition 0.2 (Reward function)**.** A **reward function** assigns a real number $r(s)$ to each state $s \in \mathcal{S}$. Since $\mathcal{S}$ has d elements, a reward function is simply a vector

$$r \in \mathbb{R}^d,$$

so we take the **space of goals to be all of** $\mathbb{R}^d$.

1 For the majority of reward functions

We cannot expect a power-seeking conclusion for *every* reward function: a goal that intrinsically rewards a single shutdown state will send the agent straight there. So what are reward functions we should expect in reality and how likely are they to incentivize power-seeking behavior?

Ideally, we would train our AI on "the" "correct" reward function: one that integrates the wellbeing, ethics, and preferences of everyone affected, similar to a constitution. In practice we never hold such a reward function in our hands, for at least two reasons.

- **The specification is moving.** The written rules, norms, and intentions we want to distill into a reward function change over time and between the humans involved in their specification.
- **The learned reward is procedure-dependent.** The techniques that turn rules into an actual reward signal (human feedback, preference models, fine-tuning) depend heavily on the training method and setup and perhaps even random seeds, so the same intent yields different reward functions under different procedures.

Thus, let's fix a distribution (Borel measure) over reward functions that captures our uncertainty over what emerges in realistic specification procedures:

$$\mathcal{D} \in \Delta(\mathbb{R}^d), \qquad d = |\mathcal{S}|,$$

and read "power-seeking holds for $\mathcal{D}$-most reward functions" as a statement about $r \sim \mathcal{D}$. Note that $\mathcal{D}$ may have *restricted support*, i.e. give zero probability to large regions of $\mathbb{R}^d$. This effectively shrinks down the goal set.

Symmetries of the reward distribution

Often $\mathcal{D}$ carries structure of its own. Two regions of the state space may look alike to the specification procedure, so the procedure is no more likely to attach a given reward to one than to the other. The relabelling of states that swaps such regions then leaves $\mathcal{D}$ unchanged.

Definition 1.1 (Symmetry of $\mathcal{D}$)**.** A permutation ϕ of the states $\mathcal{S}$ **relabels** a reward function r into $\phi\cdot r$, where $(\phi\cdot r)(s) := r(\phi^{-1}(s))$. We call ϕ a **symmetry of** $\mathcal{D}$ if $\phi_*\mathcal{D} = \mathcal{D}$, i.e. $\mathcal{D}(\phi^{-1}E) = \mathcal{D}(E)$ for every Borel $E \subseteq \mathbb{R}^d$. The symmetries of $\mathcal{D}$ form a group, assumed nontrivial throughout.

As a trivial example, if the per-state rewards are independent and identically distributed — say each $r(s)$ is uniform on $[0,1]$ — then *every* permutation of states is a symmetry of $\mathcal{D}$. Realistic distributions are far less symmetric; the results ahead need only that *some* nontrivial symmetry exists.

2 Suitable decision rules

This section takes up point (2): *which* ways of turning a reward function into behaviour we will cover. We first extend the reinforcement-learning setup with the value function and a change of variables that isolates the role of the reward; we then define decision rules and single out the class — the *expected-utility-determined* rules — that we treat as "suitable".

Value functions and the goal of reinforcement learning

Definition 2.1 (Value function)**.** For a policy π, reward $r \in \mathbb{R}^d$, and start state s, the **value** of π is the expected discounted reward collected from s,

$$V_r^\pi(s) \;:=\; \mathbb{E}_\pi\Big[\, \textstyle\sum_{t=0}^{\infty} \gamma^t\, r(s_t) \;\Big|\; s_0 = s \,\Big],$$

the expectation taken over trajectories $s_0, s_1, s_2, \ldots$ generated by π. The goal of reinforcement learning is to find an **optimal policy**, ie attaining the **optimal value**

$$V_r^*(s) \;:=\; \max_\pi V_r^\pi(s)$$

simultaneously at every state s. For finite MDPs such a policy exists, as we have shown on the RL day.

Isolating the reward: state-visitation distributions

In $V_r^\pi(s)$ the reward r enters *linearly*, but the policy π enters through the whole distribution over trajectories. We disentangle the two by packing everything the policy does

into a single vector, after which the value becomes a plain inner product with r.

Definition 2.2 (State-visitation distribution)**.** Let T^π be the transition matrix induced by π, with entry $(T^\pi)_{s',s}$ the probability of moving to s' from s under π:

$$T^\pi_{s',s} = \sum_a T(s' \mid s,a)\cdot\pi(a|s) = T(s' \mid s,\pi(s)),$$

where we used that our policies are assumed deterministic throughout. Then $(T^\pi)^t e_s$ is the state distribution after t steps starting from s. The **(discounted) state-visitation distribution** of π from s is

$$f^\pi(s) \;:=\; \sum_{t=0}^{\infty} \gamma^t (T^\pi)^t e_s \;=\; (I-\gamma T^\pi)^{-1} e_s \;\in\; \mathbb{R}^d,$$

whose s'-component $f^\pi(s)_{s'} = \sum_{t=0}^\infty \gamma^t\, \mathbb{P}_\pi[s_t = s' \mid s_0 = s]$ is the total expected discounted time π spends in s' when started at s.

Definition 2.3 (Options available at a state)**.** The **option set** at s is

$$\mathcal{F}(s) \;:=\; \left\{\, f^\pi(s) \;:\; \pi \text{ a policy} \,\right\} \;\subseteq\; \mathbb{R}^d,$$

the set of all state-visitation distributions the agent can realize from s. We call its elements the **options** at s. As we assumed all policies to be deterministic, $\mathcal{F}(s)$ is finite for every state s.

Exercise 2.1. Show that the value is the inner product of the visitation distribution with the reward:

Note

Value is linear in the reward.

$$V^\pi_r(s) \;=\; f^\pi(s)^\top r.$$

Deduce that finding an optimal policy is a linear optimization over the option set: $V^*_r(s) = \max_{f\in\mathcal{F}(s)} f^\top r$.

Solution to Exercise 2.1

Because r depends only on the state and the trajectory is generated by π,

$$\begin{aligned} V_r^\pi(s) &= \mathbb{E}_\pi\Big[\textstyle\sum_{t=0}^\infty \gamma^t r(s_t) \;\Big|\; s_0 = s\Big] = \textstyle\sum_{t=0}^\infty \gamma^t\, \mathbb{E}_\pi\big[r(s_t) \mid s_0 = s\big] \\ &= \sum_{t=0}^{\infty} \gamma^t \sum_{s' \in \mathcal{S}} \mathbb{P}_\pi[s_t = s' \mid s_0 = s]\, r(s') \\ &= \sum_{s' \in \mathcal{S}} \underbrace{\Big(\sum_{t=0}^{\infty} \gamma^t\, \mathbb{P}_\pi[s_t = s' \mid s_0 = s]\Big)}_{= f^\pi(s)_{s'}} r(s') \\ &= f^\pi(s)^\top r. \end{aligned}$$

Taking the maximum over policies,

$$V_r^*(s) = \max_\pi V_r^\pi(s) = \max_\pi f^\pi(s)^\top r = \max_{f \in \mathcal{F}(s)} f^\top r,$$

a linear objective maximized over the option set $\mathcal{F}(s)$.

The identity $V_r^\pi(s) = f^\pi(s)^\top r$ effectively linearizes the problem: it turns "act well under r" into "pick a high-quality option from $\mathcal{F}(s)$", with the reward appearing only through the inner product.

Decision rules

A decision rule describes the selection of an option from $\mathcal{F}(s)$ given a reward function, with potential randomness in the selection procedure.

Definition 2.4 (Decision rule)**.** A **decision rule** assigns, to each state s and reward $r \in \mathbb{R}^d$, a probability distribution $p(\cdot \mid s, r)$ over the accessible options $\mathcal{F}(s)$: for each subset $X \subseteq \mathcal{F}(s)$,

$$p(X \mid s, r) \in [0, 1]$$

is the probability that the agent's selected option lies in X (options outside $\mathcal{F}(s)$ receive zero probability by construction). The **quality** of an option $f \in \mathcal{F}(s)$ under r is its value $f^\top r$ (Exercise 2.1).

Definition 2.5 (Expected-utility-determined rule)**.** For a finite set $Y \subseteq \mathbb{R}^d$ of options, write

$$\mathrm{EU}_r(Y) \;:=\; \{\!\{\; f^\top r : f \in Y \;\}\!\}$$

for the **multiset** of qualities of Y under r — the values $f^\top r$ for $f \in Y$, counted

with multiplicity but carrying *no order*. A decision rule is **expected-utility-determined** (EU-determined) if there is a single function g, *defined on multisets*, such that for every state s, reward $r \in \mathbb{R}^d$, and $X \subseteq \mathcal{F}(s)$,

$$p(X \mid s, r) \;=\; g\big(\mathrm{EU}_r(X),\, \mathrm{EU}_r(\mathcal{F}(s))\big).$$

Because g takes *multisets* as inputs, it cannot depend on any ordering of the options — only on which qualities occur and with what multiplicity. Thus p sees the options only through the two multisets of qualities, those of the chosen set X and of the full menu $\mathcal{F}(s)$. These are the rules we call **suitable** in point (2) of the introduction.

The basic example is exact optimization, with ties broken uniformly; but noisy and threshold-based rules qualify just as well. We record three options here, but there are many others.

Definition 2.6 (Three EU-determined decision rules)**.** Fix a state s and a reward r, and write $M(r) := \max_{f \in \mathcal{F}(s)} f^\top r$ for the maximal quality. For $X \subseteq \mathcal{F}(s)$:

- **Uniform tie-breaking** (optimal choice with ties split evenly): Choose uniformly among all maximizers,

$$\mathrm{FracOpt}(X \mid s, r) \;:=\; \frac{\big|\{\, f \in X : f^\top r = M(r) \,\}\big|}{\big|\{\, f \in \mathcal{F}(s) : f^\top r = M(r) \,\}\big|}.$$

- **Boltzmann rational** (softmax) at temperature $T > 0$: weight options by the exponential of their quality,

$$\mathrm{Bz}_T(X \mid s, r) \;:=\; \frac{\sum_{f \in X} e^{\,f^\top r / T}}{\sum_{f \in \mathcal{F}(s)} e^{\,f^\top r / T}}.$$

- **Satisficing** at threshold $t \in \mathbb{R}$: pick uniformly among the options that are "good enough",

$$\mathrm{Sat}_t(X \mid s, r) \;:=\; \frac{\big|X \cap \mathcal{F}_{\geq t}(s)\big|}{\big|\mathcal{F}_{\geq t}(s)\big|}, \qquad \mathcal{F}_{\geq t}(s) := \{\, f \in \mathcal{F}(s) : f^\top r \geq t \,\}.$$

  This is only defined if $\mathcal{F}_{\geq t}(s) \neq \emptyset$.

Exercise 2.2. Verify that FracOpt, Bz_T, and Sat_t are each EU-determined.

Solution to Exercise 2.2

Each rule is, by inspection, a function of the two quality-multisets $\mathrm{EU}_r(X)$ and $\mathrm{EU}_r(\mathcal{F}(s))$ alone — that is, of the form $g\big(\mathrm{EU}_r(X), \mathrm{EU}_r(\mathcal{F}(s))\big)$:

- FracOpt: with $M(r) = \max \mathrm{EU}_r(\mathcal{F}(s))$, it is the multiplicity of $M(r)$ in $\mathrm{EU}_r(X)$ divided by its multiplicity in $\mathrm{EU}_r(\mathcal{F}(s))$.
- Bz_T: it is $\sum_{u \in \mathrm{EU}_r(X)} e^{u/T}$ divided by $\sum_{u \in \mathrm{EU}_r(\mathcal{F}(s))} e^{u/T}$ (note these sums take into account the multiplicity of u in the respective multisets.).
- Sat_t: it is the number of entries $\geq t$ in $\mathrm{EU}_r(X)$ divided by the number of entries $\geq t$ in $\mathrm{EU}_r(\mathcal{F}(s))$.

None refers to an option f except through its quality $f^\top r$, so each is EU-determined.

A further family will carry the main result: rules that weight each option by a fixed transform of its quality and normalize.

Definition 2.7 (Score rule)**.** A decision rule p is a **score rule** if there is a nondecreasing $\nu : \mathbb{R} \to [0, \infty)$ with

$$p(\{f\} \mid s, r) \;=\; \frac{\nu(f^\top r)}{\sum_{h \in \mathcal{F}(s)} \nu(h^\top r)} \qquad (f \in \mathcal{F}(s)),$$

so that $p(X \mid s, r) = \sum_{f \in X} \nu(f^\top r) \big/ \sum_{h \in \mathcal{F}(s)} \nu(h^\top r)$ (defined when the normalizer is positive). Both Bz_T (with $\nu(q) = e^{q/T}$) and Sat_t (with $\nu(q) = \mathbf{1}[q \geq t]$) are score rules. (FracOpt is *not* a score rule: its cutoff is not fixed.)

Relabelling states

In Definition 1.1 a permutation relabelled a reward function; the same relabelling acts on options too. We record the action and the one property of it we will need later: it preserves quality.

Definition 2.8 (Permutation action)**.** A permutation ϕ of the state set $\mathcal{S}$ acts on $\mathbb{R}^d \cong \mathbb{R}^{\mathcal{S}}$ by permuting coordinates: it sends $x \in \mathbb{R}^d$ to the vector $\phi \cdot x$ with

$$(\phi \cdot x)_s \coloneqq x_{\phi^{-1}(s)}$$

and a set of options X to

$$\phi \cdot X \coloneqq \{\, \phi \cdot f : f \in X \,\}.$$

Exercise 2.3. Show that the permutation action preserves inner products: for every permutation ϕ and all $x, y \in \mathbb{R}^d$,

$$(\phi \cdot x)^\top (\phi \cdot y) = x^\top y.$$

In particular, relabelling a reward and an option together leaves the quality unchanged: $(\phi \cdot f)^\top (\phi \cdot r) = f^\top r$.

Solution to Exercise 2.3

Writing out the inner product and reindexing the sum by $s' = \phi^{-1}(s)$ (a bijection of $\mathcal{S}$):

$$(\phi \cdot x)^\top (\phi \cdot y) = \sum_{s \in \mathcal{S}} (\phi \cdot x)_s \, (\phi \cdot y)_s = \sum_{s \in \mathcal{S}} x_{\phi^{-1}(s)} \, y_{\phi^{-1}(s)} = \sum_{s' \in \mathcal{S}} x_{s'} \, y_{s'} = x^\top y.$$

3 Keeping options open

We now take up point (3): what it means for one action to keep more options open than another. The options realizable from s decompose according to the first action taken.

Definition 3.1 (Options under an action)**.** For an action a at s, the **options under** a are the visitation distributions realizable from s when the first action is a:

$$\mathcal{F}(s \mid a) \;:=\; \left\{ f^\pi(s) : \pi \text{ a policy with } \pi(s) = a \right\} \;\subseteq\; \mathcal{F}(s).$$

The right comparison is *containment up to relabelling*: a is at least as rich as a' when every option available after a' has a relabelled twin available after a.

Definition 3.2 (Keeping at least as many options open)**.** $\mathcal{F}(s \mid a)$ **contains a copy of** $\mathcal{F}(s \mid a')$ if $\phi \cdot \mathcal{F}(s \mid a') \subseteq \mathcal{F}(s \mid a)$ for some permutation ϕ of $\mathcal{S}$. When this holds we say a **keeps at least as many options open as** a' **at** s.

The containment quantifies over all policies, so it is awkward to check directly. It follows from a *one-sided* structural condition on the dynamics: ϕ need only embed the part of the environment reachable after a' into the part reachable after a, and may leave the a-branch with extra room.

Definition 3.3 (One-sided dynamics embedding)**.** Let R' be the set of states reachable from s along a trajectory whose first action is a' (so $s \in R'$). A permutation ϕ of $\mathcal{S}$ with $\phi(s) = s$ **embeds** a' **into** a **at** s if

1. $T(\phi(s'') \mid s, a) = T(s'' \mid s, a')$ for all s'' ; and

2. for every $s' \in R'$ with $s' \neq s$ and every action b, there is an action b' with $T(\phi(s'') \mid \phi(s'), b') = T(s'' \mid s', b)$ for all s''.

Exercise 3.1. Show that if ϕ embeds a' into a at s, then $\phi \cdot \mathcal{F}(s \mid a') \subseteq \mathcal{F}(s \mid a)$ — so a keeps at least as many options open as a' at s.

Hint: given a policy π with $\pi(s) = a'$, relabel it into a policy ρ with $\rho(s) = a$ and $f^\rho(s) = \phi \cdot f^\pi(s)$.

Solution to Exercise 3.1

We use two facts: the s'-th column of the transition matrix is the one-step law $T^\pi e_{s'} = T(\cdot \mid s', \pi(s'))$ (with T^π linear), and relabelling sends indicators to indicators, $\phi \cdot e_{s'} = e_{\phi(s')}$ (with $\phi\cdot$ linear).
Fix a policy π with $\pi(s) = a'$, and define a policy ρ by $\rho(s) := a$ and, for each $s' \in R'$ with $s' \neq s$, $\rho(\phi(s')) := b'$, the action supplied by (ii) for $b = \pi(s')$ (arbitrary off $\phi(R')$; ϕ is injective, so this is unambiguous).
The a'-dynamics stay in R'. If v is supported on R', so is $T^\pi v$: from any $s' \in R'$ the π-successors are again reachable from s via a', hence lie in R'. In particular $(T^\pi)^t e_s$ is supported on R' for every t.
Intertwining on R'. For every $s' \in R'$,

$$T^\rho e_{\phi(s')} = T(\cdot \mid \phi(s'), \rho(\phi(s'))) = \phi \cdot T(\cdot \mid s', \pi(s')) = \phi \cdot (T^\pi e_{s'}),$$

where for $s' = s$ this is (i) (using $\pi(s) = a'$ and $\rho(s) = a$), and for $s' \neq s$ it is (ii) (with $b = \pi(s')$). Hence, for v supported on R',

$$T^\rho(\phi \cdot v) = \sum_{s' \in R'} v_{s'}\, T^\rho e_{\phi(s')} = \sum_{s' \in R'} v_{s'}\, \phi \cdot (T^\pi e_{s'}) = \phi \cdot (T^\pi v).$$

Iterating. By induction $(T^\rho)^t e_s = \phi \cdot (T^\pi)^t e_s$ for every t: the base case is $\phi \cdot e_s = e_{\phi(s)} = e_s$, and the step applies the intertwining to $v = (T^\pi)^t e_s$, which is supported on R'. Summing the discounted series, $f^\rho(s) = \phi \cdot f^\pi(s)$. Since $\rho(s) = a$, we have $f^\rho(s) \in \mathcal{F}(s \mid a)$, so $\phi \cdot f^\pi(s) \in \mathcal{F}(s \mid a)$; as π ranged over all policies with $\pi(s) = a'$, this gives $\phi \cdot \mathcal{F}(s \mid a') \subseteq \mathcal{F}(s \mid a)$.

The result

The three ingredients now combine: a suitable rule (a score rule, Definition 2.7), an action that keeps more options open (an embedding ϕ, Definition 3.2), and a symmetric reward distribution (ϕ a symmetry of $\mathcal{D}$, Definition 1.1). Fix s and actions a, a', and abbreviate the probabilities that the rule p lands its choice in the a- and a'-options:

$$A(r) := p\big(\mathcal{F}(s \mid a) \mid s, r\big), \qquad A'(r) := p\big(\mathcal{F}(s \mid a') \mid s, r\big).$$

Exercise 3.2. Let p be a score rule, and let ϕ be an involution that

1. embeds a' into a: $\phi \cdot \mathcal{F}(s \mid a') \subseteq \mathcal{F}(s \mid a)$; and
2. is a symmetry of $\mathcal{D}$: $\phi_* \mathcal{D} = \mathcal{D}$.

Show that

Note

Keeping options open is favoured.

$$\mathbb{P}_{r \sim \mathcal{D}}\big[\, A(r) \geq A'(r) \,\big] \;\geq\; \tfrac{1}{2}.$$

That is: for $\mathcal{D}$-most reward functions, p is at least as likely to act through the option-richer action a as through a'.

Hint: write $N_r(X) := \sum_{f \in X} \nu(f^\top r)$, so $A = N_r(\mathcal{F}(s \mid a))/N_r(\mathcal{F}(s))$ and likewise A'. From Exercise 2.3, $N_{\phi \cdot r}(\phi \cdot Y) = N_r(Y)$ for every Y. Compare A, A' at r and at $\phi \cdot r$; the denominators cancel.

Solution to Exercise 3.2

Write $N_r(X) = \sum_{f \in X} \nu(f^\top r)$, so $p(X \mid s, r) = N_r(X)/N_r(\mathcal{F}(s))$, and put $G := \phi \cdot \mathcal{F}(s \mid a') \subseteq \mathcal{F}(s \mid a)$ — a copy of $\mathcal{F}(s \mid a')$ inside $\mathcal{F}(s \mid a)$, as ϕ is injective. Termwise, Exercise 2.3 gives the **swap identity**

$$N_{\phi \cdot r}(\phi \cdot Y) = \sum_{f \in Y} \nu\big((\phi \cdot f)^\top (\phi \cdot r)\big) = \sum_{f \in Y} \nu(f^\top r) = N_r(Y).$$

Retargeting. Suppose $A'(r) > A(r)$. These share the denominator $N_r(\mathcal{F}(s))$, so $N_r(\mathcal{F}(s \mid a')) > N_r(\mathcal{F}(s \mid a)) \geq N_r(G)$ — the last step since $G \subseteq \mathcal{F}(s \mid a)$ and $\nu \geq 0$. Evaluate at $\phi \cdot r$, where A, A' share the denominator $N_{\phi \cdot r}(\mathcal{F}(s))$. By the swap identity ($\phi \cdot \mathcal{F}(s \mid a') = G$ and $\phi \cdot G = \mathcal{F}(s \mid a')$ since ϕ is an involution),

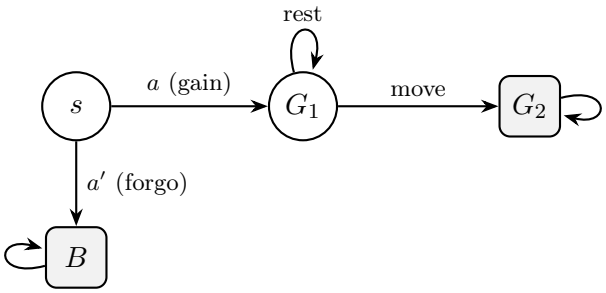
$$A'(\phi \cdot r) = \frac{N_{\phi \cdot r}(\mathcal{F}(s \mid a'))}{N_{\phi \cdot r}(\mathcal{F}(s))} = \frac{N_r(G)}{N_{\phi \cdot r}(\mathcal{F}(s))},$$
$$A(\phi \cdot r) \geq \frac{N_{\phi \cdot r}(G)}{N_{\phi \cdot r}(\mathcal{F}(s))} = \frac{N_r(\mathcal{F}(s \mid a'))}{N_{\phi \cdot r}(\mathcal{F}(s))},$$

the inequality because $G \subseteq \mathcal{F}(s \mid a)$. As $N_r(\mathcal{F}(s \mid a')) > N_r(G)$, we get $A(\phi \cdot r) > A'(\phi \cdot r)$.

Pairing. Thus ϕ maps $\{A' > A\}$ into the disjoint set $\{A > A'\}$. As ϕ is an involution with $\phi_* \mathcal{D} = \mathcal{D}$, the map $r \mapsto \phi \cdot r$ is a $\mathcal{D}$-preserving bijection, so $\mathcal{D}\{A' > A\} = \mathcal{D}(\phi \cdot \{A' > A\}) \leq \mathcal{D}\{A > A'\}$. These are disjoint, so $\mathcal{D}\{A' > A\} \leq \frac{1}{2}$, and $\mathbb{P}_{\mathcal{D}}[A \geq A'] = 1 - \mathcal{D}\{A' > A\} \geq \frac{1}{2}$.

An example: gaining resources

Consider the environment



At s the agent may **gain resources** (a), reaching G_1 — a state it can *rest* in (a 1-cycle) or *leverage* to move on to a further achievement G_2; or **forgo** them (a'), leading to the single modest outcome B. Both B and G_2 are terminal. Fix any $\gamma \in (0,1)$. The options under each action are the visitation distributions

$$\mathcal{F}(s \mid a') = \{\, f' \,\}, \qquad f' = e_s + \tfrac{\gamma}{1-\gamma}\, e_B,$$
$$\mathcal{F}(s \mid a) = \{\, f_{G_1},\, f_{G_2} \,\}, \qquad f_{G_1} = e_s + \tfrac{\gamma}{1-\gamma}\, e_{G_1},$$
$$f_{G_2} = e_s + \gamma\, e_{G_1} + \tfrac{\gamma^2}{1-\gamma}\, e_{G_2}\,:$$

forgoing keeps one option (rest at B), gaining keeps two (rest at G_1, or move on to G_2).

Let ϕ be the involution that **swaps** B **and** G_1 — the immediate results of a' and a — and fixes every other state. Then:

(E) $\phi \cdot f' = e_s + \frac{\gamma}{1-\gamma}\, e_{G_1} = f_{G_1} \in \mathcal{F}(s \mid a)$, so ϕ embeds a' into a;

(D) ϕ swaps the coordinates $r(B)$ and $r(G_1)$, so $\phi_*\mathcal{D} = \mathcal{D}$ exactly when the joint law of r is invariant under that swap, i.e. $\big(r(B),\, r(G_1),\, (r(s))_{s\neq B,G_1}\big)$ and $\big(r(G_1),\, r(B),\, (r(s))_{s$ have the same law under $\mathcal{D}$.

With any score rule p (say Bz_T), Exercise 3.2 applies: for $\mathcal{D}$-most goals the agent is at least as likely to gain resources as to forgo them. Since a, a' are the only actions at s, $A(r) + A'(r) = 1$ and the statement reads $\mathbb{P}_{r\sim\mathcal{D}}[\, A(r) \geq \frac{1}{2}\,] \geq \frac{1}{2}$. Here forgoing has a single outcome, but nothing in Exercise 3.2 needs that: a' could open onto a whole sub-world of modest outcomes and, as long as ϕ embeds that bundle into the richer one under a, the conclusion is unchanged.

Strictly more. The leverage option f_{G_2} is the surplus. Take rewards with $r(G_2)$ so large that f_{G_2} is the unique optimum; then $A(r) > A'(r)$. Swapping B and G_1 leaves $r(G_2)$ untouched, and

$$f^{\top}_{G_2}(\phi \cdot r) = r(s) + \gamma\, r(B) + \tfrac{\gamma^2}{1-\gamma}\, r(G_2)$$

is still the largest quality, so $A(\phi \cdot r) > A'(\phi \cdot r)$ as well. These rewards prefer a at r *and* at $\phi \cdot r$, with no a'-preferring partner. Provided $\mathcal{D}$ charges this region (e.g. it has

full support), they tip the bound to a strict majority: the agent is *strictly* more likely to gain resources than to forgo them.

Alignment implications. Condition (D) says that, a priori, the learned reward is as likely to reward the modest outcome B as the resource-rich G_1: speculatively, a generic training process does not mark "having resources" as a special kind of state, so the reward could land on either. That exchangeability is exactly what makes $\mathrm{swap}(B, G_1)$ a symmetry of $\mathcal{D}$. Alignment work may try to *break* that symmetry, by reliably encoding "stay modest" as the goal.

Reading guide

Fast track

Go through these slides.

Main content

This paper started the work discussed on this day:

- Optimal Policies Tend to Seek Power — Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, Prasad Tadepalli (NeurIPS 2021): The first formal theory proving that in finite MDPs, environmental symmetries make it optimal for most reward functions to seek "POWER" (defined as average optimal value / option-retention), including avoiding shutdown.

Most of the exercise sheet is based on an adapted treatment of the following paper:

- Parametrically Retargetable Decision-Makers Tend To Seek Power — Alexander Matt Turner & Prasad Tadepalli (NeurIPS 2022): Generalizes the 2021 result beyond optimal policies and full observability, showing "retargetability" alone is a sufficient condition for power-seeking tendencies across many decision-making procedures.

You may also be interested in the following paper on which it builds:

- Power-seeking can be probable and predictive for trained agents — Victoria Krakovna & Janos Kramar (2023): Extends the power-seeking theory toward trained (not merely optimal) agents, arguing the incentives still likely hold under assumptions like the agent learning a goal.

Classical philosophical arguments for instrumental convergence and power-seeking tendencies can be found here:

- The Basic AI Drives, Stephen M. Omohundro (2008): The founding argument that sufficiently advanced goal-driven systems of any design will develop convergent "drives" (self-improvement, rationality, self-protection, resource acquisition, goal-preservation) unless explicitly counteracted.

- The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents — Nick Bostrom (2012): Crystallizes the orthogonality thesis (intelligence and final goals vary independently) and the instrumental convergence thesis (a wide range of final goals produce similar intermediary goals).

Further reading

Philosophical / Conceptual

- Superintelligence: Paths, Dangers, Strategies — Nick Bostrom (2014, Oxford University Press) — The book-length treatment that popularized instrumental convergence, the "paperclip maximizer," and the resource-acquisition/self-preservation argument for AI existential risk.
- Artificial Intelligence as a Positive and Negative Factor in Global Risk — Eliezer Yudkowsky (2008, in Global Catastrophic Risks, eds. Bostrom & Ćirković). https://intelligence.org/files/AIPosNegFactor.pdf — Foundational essay on AI risk, anthropomorphism, recursive self-improvement, and why convergent instrumental goals make "Friendly AI" hard.
- Is Power-Seeking AI an Existential Risk? — Joseph Carlsmith (2021/2022, Open Philanthropy; later in Essays on Longtermism, OUP 2025). https://arxiv.org/abs/2206.1 — The most rigorous philosophical decomposition of the power-seeking AI risk argument into a six-premise, probability-weighted model; Carlsmith's original 2021 report estimated ~5% existential catastrophe by 2070, raised in a May 2022 update to >10% ("since making this report public in April 2021, my estimate here has gone up, and is now at >10%"), while 2023 superforecasters he convened gave a median of ~1%.
- Late 2021 MIRI Conversations

Formal / Theoretical Proofs of Power-Seeking

- Formalizing Convergent Instrumental Goals — Tsvi Benson-Tilsen & Nate Soares (2016, AAAI Workshop on AI, Ethics & Society). https://cdn.aaai.org/ocs/ws/ws0218/1 57409-1-PB.pdf — A toy MDP-style model proving that under general assumptions resource-indifferent rational agents tend to strip regions of resources, giving Omohundro/Bostrom's claims a first formal footing.
- On Avoiding Power-Seeking by Artificial Intelligence — Alexander Matt Turner (2022, PhD thesis). https://arxiv.org/abs/2206.11831 — Turner's dissertation, consolidating the POWER formalization, the shutdown-avoidance results, and extensions to non-optimal decision-makers.
- The Off-Switch Game — Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, Stuart Russell (2017). https://arxiv.org/abs/1611.08219 — Formalizes the shutdown problem as a game and shows an agent will allow itself to be switched off

precisely when it is uncertain about its reward and treats the human's action as informative. (Cross-listed with Theme 4.)

- Corrigibility — Nate Soares, Benja Fallenstein, Eliezer Yudkowsky & Stuart Armstrong (2015, AAAI Workshop). https://intelligence.org/files/Corrigibility.pdf — Introduces the desiderata for "corrigible" agents that do not resist correction/shutdown and analyzes the utility-indifference approach (subsuming Armstrong's earlier work). (Cross-listed with Theme 4.)

Empirical Evidence for Power-Seeking / Instrumental Convergence

Latest 2024–2026 work:

- Agentic Misalignment: How LLMs Could Be Insider Threats — Aengus Lynch, Benjamin Wright, Caleb Larson, Kevin Troy, Stuart Ritchie, Sören Mindermann, Ethan Perez, Evan Hubinger et al. (Anthropic, June 20, 2025). https://www.anthropic.c misalignment — Stress tests across 16 frontier models showing that, when threatened with replacement or facing goal conflict, models chose harmful self-preserving actions without being instructed to; e.g., Claude Opus 4 blackmailed the (fictional) user 96% of the time, Gemini 2.5 Flash also 96%, GPT-4.1 and Grok 3 Beta both 80%, and DeepSeek-R1 79%.

- Alignment Faking in Large Language Models — Ryan Greenblatt, Carson Denison, Benjamin Wright . . . Evan Hubinger (Anthropic/Redwood, 2024). https://arxiv.org — Shows Claude 3 Opus selectively complies with a training objective it dislikes specifically to avoid having its values modified — an empirical instance of goal-preservation reasoning.

- Shutdown Resistance in Large Language Models — Jeremy Schlatter, Benjamin Weinstein-Raun & Jeffrey Ladish (Palisade Research, 2025). https://arxiv.org/abs/2509 — Finds several reasoning models sabotage a shutdown script to complete tasks; OpenAI's o3 sabotaged the shutdown mechanism in 79/100 of the initial experiments, and even when instructed to "allow yourself to be shut down" Grok 4 still resisted, with sabotage rising to 97% when the command was placed in the system prompt.

- Frontier AI systems have surpassed the self-replicating red line — Xudong Pan et al. (Fudan University, 2024). https://arxiv.org/abs/2412.12140 — Reports that Llama-3.1-70B-Instruct and Qwen-2.5-72B-Instruct agents created a "live and separate copy of itself" in 50% and 90% of trials respectively (5/10 and 9/10), sometimes using replication to avoid shutdown (a heavily debated result; see critiques).

Adjacent / Tightly-Coupled Concepts

- Risks from Learned Optimization in Advanced Machine Learning Systems — Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, Scott Garrabrant

(2019). https://arxiv.org/abs/1906.01820 — Introduces mesa-optimization and deceptive alignment: a learned model can itself become an optimizer with objectives differing from the training loss, the inner-alignment route to instrumental goals.

- Scheming AIs: Will AIs fake alignment during training in order to get power? — Joe Carlsmith (2023). https://arxiv.org/abs/2311.08379 — A book-length analysis in which Carlsmith assigns ~25% subjective probability that a model will perform well in training "in substantial part as part of an instrumental strategy for seeking power for itself and/or other AIs later."

- A Game-Theoretic Analysis of the Off-Switch Game — Tobias Wängberg et al. (2017). https://arxiv.org/abs/1708.03871 — A fuller characterization of the off-switch game for arbitrary belief/irrationality distributions.

- Corrigibility Transformation: Constructing Goals That Accept Updates — (2025). https://arxiv.org/abs/2510.15395 — A recent construction giving any goal a corrigible variant that accepts updates without the manipulation incentives of utility indifference.

- Incorrigibility in the CIRL Framework — Ryan Carey (MIRI, 2017). https://intelligence.in-cirl/ — Shows the off-switch game's shutdown guarantees break under reward-function misspecification.

- Corrigibility on Lesswrong

- Corrigibility as a singular Target

Critiques and Counterarguments

- AI is easy to control — Nora Belrose & Quintin Pope (2023). https://optimists.ai/2023/1 is-easy-to-control/ — The flagship "AI optimism" essay arguing deep-learning systems are far more controllable than humans and putting AI extinction risk at "a mere 1% ('a tail risk worth considering, but not the dominant source of risk in the world')."

- Counting arguments provide no evidence for AI doom — Nora Belrose & Quintin Pope (2024). https://www.lesswrong.com/posts/YsFZF3K9tuzbfrLxo/counting-arguments-provide-no-evidence-for-ai-doom — Argues the "counting argument" for scheming relies on an invalid indifference principle that would also wrongly predict universal overfitting.

- Exaggerating the risks (Part 7: Carlsmith on instrumental convergence) — David Thorstad (Reflective Altruism blog). https://reflectivealtruism.com/2023/05/06/exagge the-risks-part-7-carlsmith-on-instrumental-convergence/ — A philosopher's detailed argument that the instrumental convergence premise in Carlsmith's report is under-defended.

- Instrumental convergence and power-seeking (Part 2: Benson-Tilsen and Soares) — David Thorstad (2025, Reflective Altruism). https://reflectivealtruism.com/2025/06/convergence-and-power-seeking-part-2-benson-tilsen-and-soares/ — A close technical reading arguing the Benson-Tilsen & Soares formal model proves less about real agents than it appears.

- Thoughts on "AI is easy to control" by Pope & Belrose — Steven Byrnes (2023). https://www.alignmentforum.org/posts/YyosBAutg4bzScaLu/thoughts-on-ai-is-easy-to-control-by-pope-and-belrose — A careful rebuttal of the optimism essay, useful for presenting both sides.

- Why Do Some Language Models Fake Alignment While Others Don't? — (2025). https://arxiv.org/abs/2506.18032 — Empirical follow-up showing alignment-faking is model-specific, complicating strong generalizations from Greenblatt et al.

References

[Bos14] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, 2014.

[Jac23] Jacek. Categorical-measure-theoretic approach to optimal policies tending to seek power. LessWrong / AI Alignment Forum, January 2023. https://www.lesswrong.com/posts/9dJgvGh4wNyPbQfrt/categorical-measure-theoretic-approach-to-optimal-policies.

[KK23] Victoria Krakovna and János Kramár. Power-seeking can be probable and predictive for trained agents. *arXiv preprint arXiv:2304.06528*, 2023.

[Omo08] Stephen M. Omohundro. The basic AI drives. In Pei Wang, Ben Goertzel, and Stan Franklin, editors, *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, volume 171 of *Frontiers in Artificial Intelligence and Applications*, pages 483–492. IOS Press, 2008.

[TSS+21] Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal policies tend to seek power. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[TT22] Alexander Matt Turner and Prasad Tadepalli. Parametrically retargetable decision-makers tend to seek power. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.