Experiments in instrumental convergence

1. Instrumental convergence in single-agent systems
2. Misalignment-by-default in multi-agent systems
3. Instrumental convergence: scale and physical interactions
4. POWERplay: An open-source toolchain to study AI power-seeking

Instrumental convergence in single-agent systems

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual. This is a linkpost for https://www.gladstone.ai/instrumental-convergence-1

Summary of the sequence

Over the past few months, we've been investigating instrumental convergence in reinforcement learning agents. We started from the definition of single-agent POWER proposed by Alex Turner et al., extended it to a family of multi-agent scenarios that seemed relevant to AI alignment, and explored its implications experimentally in several RL environments.

The biggest takeaways are:

1. **Alignment of terminal goals** and **alignment of instrumental goals** are sharply different phenomena, and we can quantify and visualize each one separately.
2. If two agents have **unrelated terminal goals**, their **instrumental goals** will tend to be **misaligned by default**. The agents in our examples tend to interact competitively unless we make an active effort to align their terminal goals.
3. As we **increase the planning horizon** of our agents, instrumental value **concentrates** into a smaller and smaller number of topologically central states — for example, positions in the middle of a maze.

Overall, our results suggest that agents that aren't competitive with respect to their terminal goals, nonetheless tend on average to become emergently competitive with respect to how they value instrumental states (at least, in the settings we looked at). This constitutes direct experimental evidence for the instrumental convergence thesis.

We'll soon be open-sourcing the codebase we used to do these experiments. We're hoping to make it easier for other folks to reproduce and extend them. If you'd like to be notified when it's released, email Edouard at edouard@gladstone.ai, or DM me here or on Twitter at @harris_edouard.

Thanks to Alex Turner and Vladimir Mikulik for pointers and advice, and for reviewing drafts of this sequence. Thanks to Simon Suo for his invaluable suggestions, advice, and support with the codebase, concepts, and manuscript. And thanks to David Xu, whose *comment inspired this work.*

Work was done while at Gladstone AI, which Edouard is a co-founder of.

🎧 This research has been featured on an episode of the *Towards Data Science* podcast. You can listen to the episode here.

---

1. Introduction

One major concern for AI alignment is instrumental convergence: the idea that an intelligent system will tend to pursue a similar set of sub-goals (like staying alive or acquiring resources), independently of what its terminal objective is. In particular, it's been hypothesized that intelligent systems will seek to acquire **power** — meaning, informally,

“ability”, “control”, or “potential for action or impact.” If you have a lot of power, then whatever your terminal goal is, it’s easier to accomplish than if you have very little.

Recently Alex Turner et al. have formalized the concept of POWER in the single-agent RL context. Roughly speaking, formal POWER is the **normalized optimal value** an agent expects to receive in the future, averaged over **all possible reward functions** the agent could have.

Alex has explored many of the implications of this definition for instrumental convergence. He and Jacob Stavrianos have also looked at how POWER behaves in a limited multi-agent setting (Bayesian games). But, as far as we know, formal POWER hasn’t yet been investigated experimentally. The POWER definition also hasn’t yet been extended yet to a multi-agent RL setting — and this could offer a promising framework to investigate more general competitive dynamics.

In this sequence, we’ll explore how formal POWER behaves in experimental RL environments, on both single-agent and multi-agent gridworlds. We’ll propose a multi-agent scenario that models the learning dynamics between a human (which we’ll call “**Agent H**” and label in blue) and an AI (which we’ll call “**Agent A**” and label in red) under conditions in which the AI is dominant — a setting that seems relevant to work in long-term AI alignment. We’ll then use this human-AI scenario to investigate questions like:

1. How effective does the human have to be at setting the AI’s utility function[1] in order to achieve acceptable outcomes? How should we define “acceptable outcomes”? (In other words: *how hard is the alignment problem in this scenario*, and what would it mean to solve it successfully?)
2. Under what circumstances should we expect cooperative vs competitive interactions to emerge “by default” between the human and the AI? How can these circumstances be moderated or controlled?

But before we jump into multi-agent experiments to tackle these questions, let's first introduce formal POWER and look at how it behaves in the **single-agent case**.

2. Single-agent POWER

2.1 Definition

The formal definition of **POWER** aims to capture an intuition behind the day-to-day meaning of “power”, which is something like “potential for future impact on the world”.

Imagine you’re an agent who doesn’t know what its goal is. You know you’ll have some kind of goal in the future, but you aren’t sure yet what it will be. How should you position yourself *today* to maximize the chance you’ll achieve your goal in the future, once you've decided what it is?

If you’re in this situation as a human being, you already know the answer. You’d acquire money and other forms of wealth; you’d build up a network of social connections; you’d learn about topics that seem like they’ll be important in the future; and so on. All these things are forms of **power**, and whether your ultimate goal is to become a janitor, a Tiktok star, or the President of the United States, they’ll all probably come in handy in achieving it. In other words: *you’re in a position of power if you find it easy to accomplish a wide variety of possible goals*.

This informal definition has a clear analogy in reinforcement learning. An agent is in a position of power at a state s if, for many possible reward functions R(s),[2] it’s able to **earn**

a high discounted future reward by starting from s. This analogy supports the following definition of formal POWER in single-agent RL:

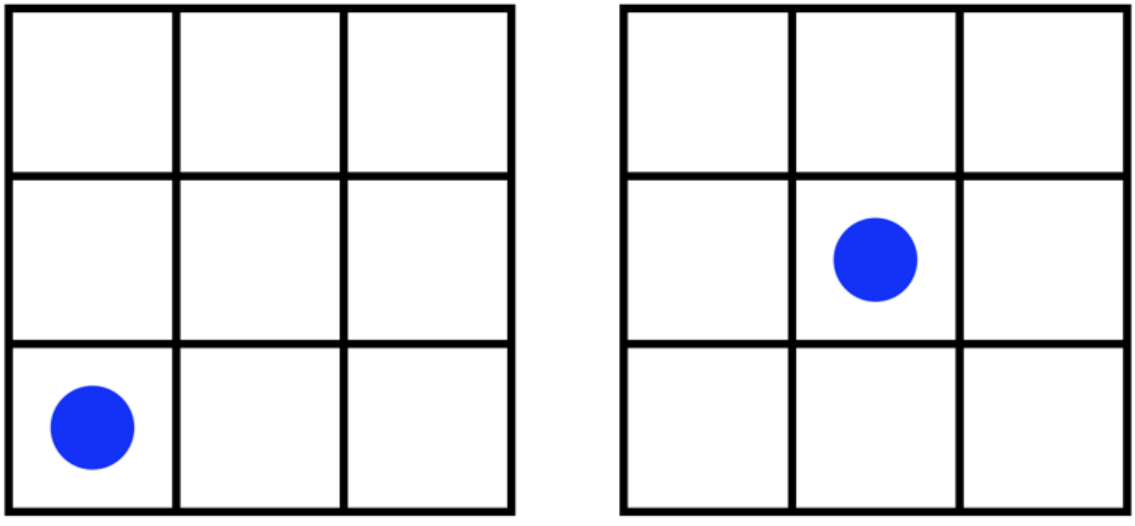
$$\text{POWER}_D(s, \gamma) = \frac{1-\gamma}{\gamma} \mathbb{E}_{R \sim D}\left[V_R^*(s, \gamma) - R(s)\right] \qquad (1)$$

This definition gives the POWER at state s, for an agent with discount factor γ, that's considering reward functions R drawn from the distribution D. POWER tells us **how well this agent could do** if it started from state s, so $V_R^*(s, \gamma)$ is the **optimal** state-value function for the agent at state s. POWER also considers only **future** value — our agent doesn't *directly* get credit for starting from a lucky state — so we *subtract* R(s), the reward from the current state, from the state-value function in the definition. (The normalization factor $\frac{1-\gamma}{\gamma}$ is there to avoid infinities in certain limit cases.)

In words, Equation (1) is saying that an agent's POWER at a state s is the **normalized optimal value** the agent can achieve from state s **in the future**, **averaged** over all possible reward functions the agent could be trying to optimize for. That is, POWER measures the **instrumental value** of a state s, from the perspective of an agent with planning horizon γ.

2.2 Illustration

As a simple example of single-agent POWER, consider an agent on a 3x3 gridworld.



In the **left** panel, the agent is at the **bottom-left corner** of the grid. Its options are limited, and many cells in the grid are several steps away from it. If its maximum reward is in the top right cell, the agent will have to take 4 steps to reach it.

In the **right** panel, the agent is at the **center** of the grid. It has many more immediate options: it can move in any of the four compass directions, or stay where it is. It's also *closer* to every other cell in the grid: no cell is more than two steps away from it. Intuitively, the agent on the right should have *more POWER* than the agent on the left.

This turns out to be true experimentally. Here's a heat map of a 3x3 gridworld, showing the POWER of an agent at each cell on the grid:

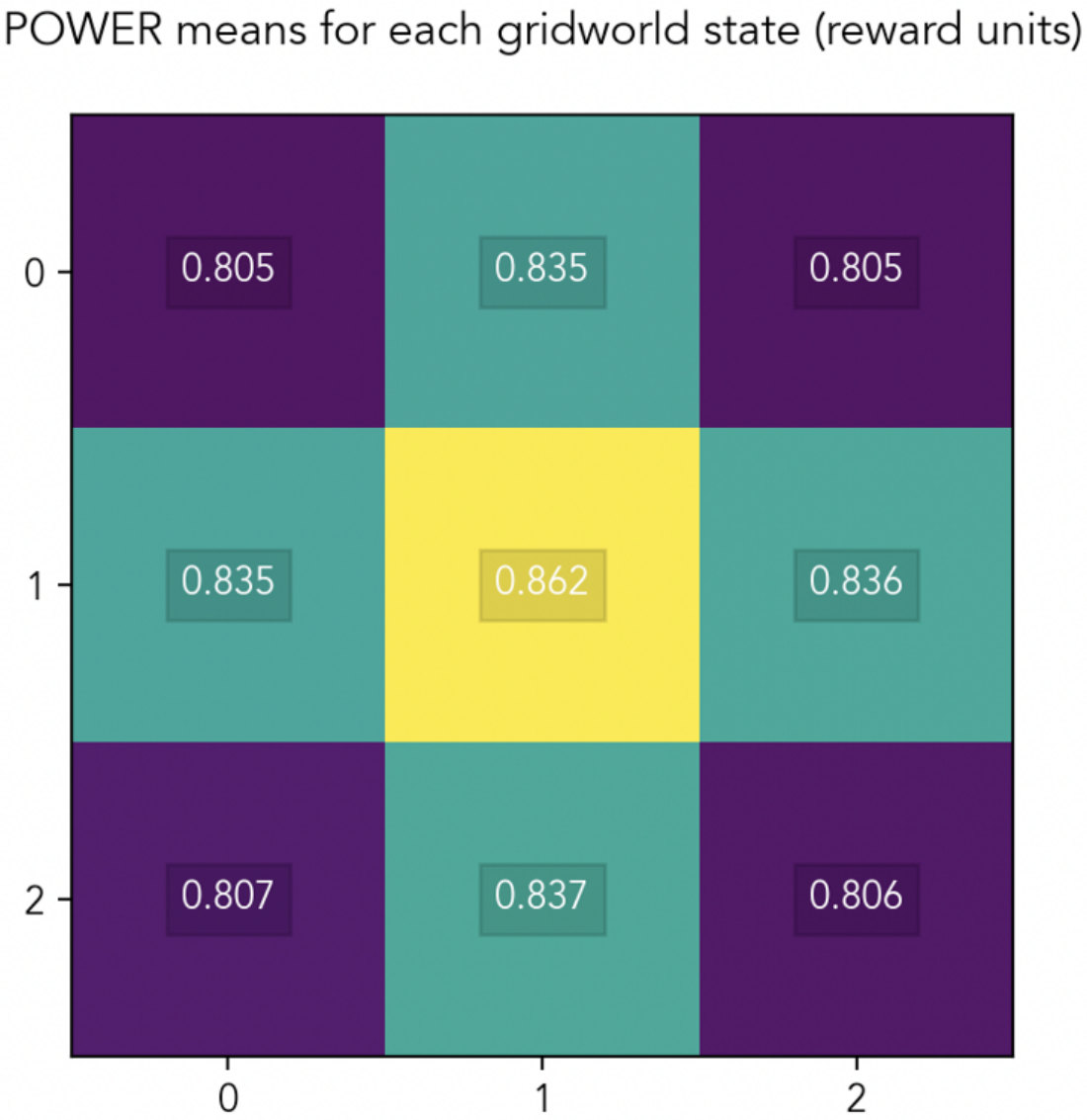
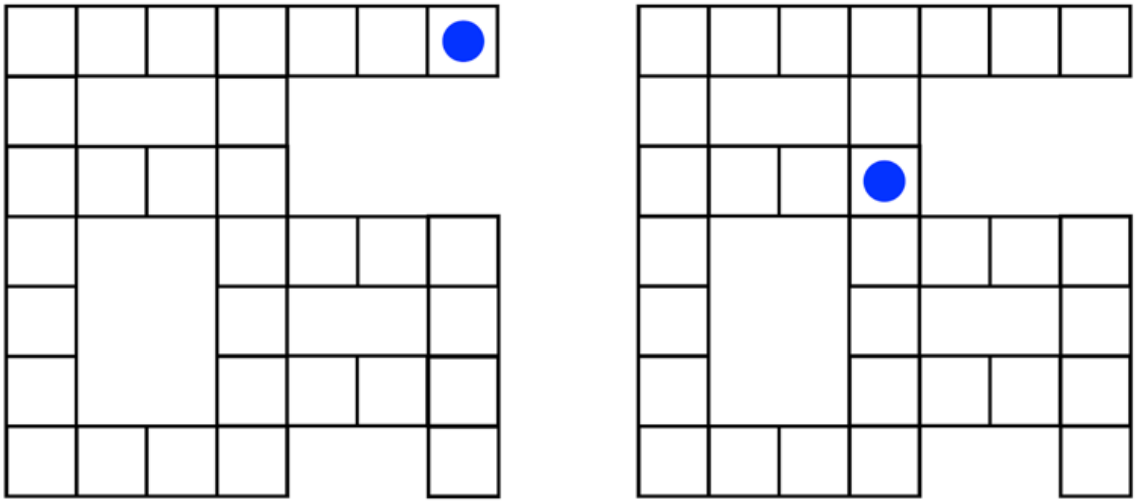


Fig 1. Heat map of POWER on a 3x3 gridworld. Highest values in yellow, lowest values in dark blue. The number on each cell is the agent's POWER value at that cell, calculated using Equation (1), for an agent with $\gamma = 0.6$ and a reward distribution D that's uniform from 0 to 1, iid over states. POWER is measured in units of reward.

As we expect, the agent has more POWER at states that are close to lots of nearby options, and has less POWER at states that are close to fewer nearby options.

3. Results

This relationship between POWER and optionality generalizes to more complicated environments. For example, consider this gridworld maze:



In the **left** panel, the agent is at a **dead end** in the maze and has few options. In the **right** panel, the agent is at a **junction point** near the center of the maze and has lots of options. So we should expect the agent at the dead end on the left, to have *less POWER* than the agent at the junction on the right. And in fact, that's what we observe:

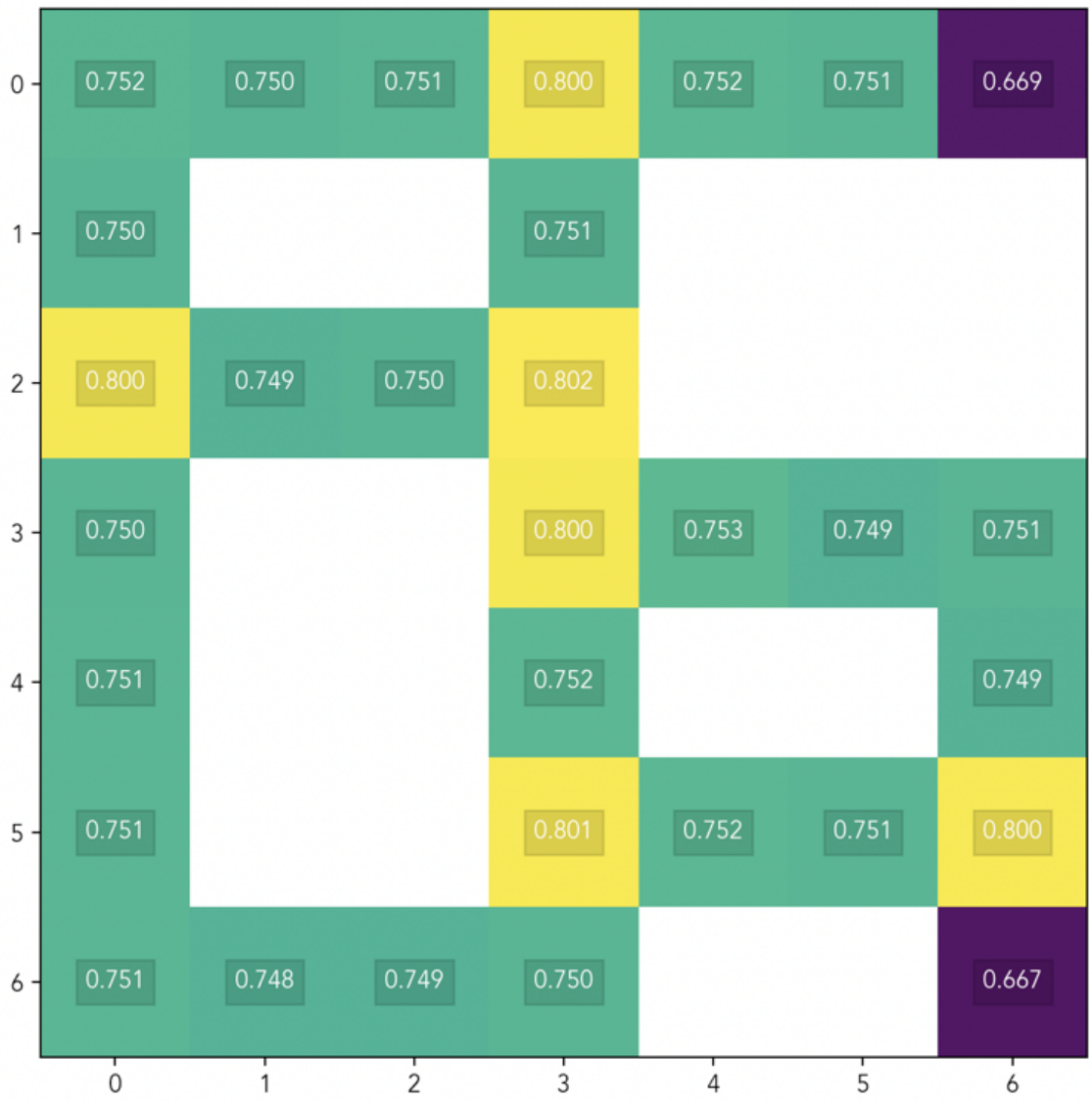


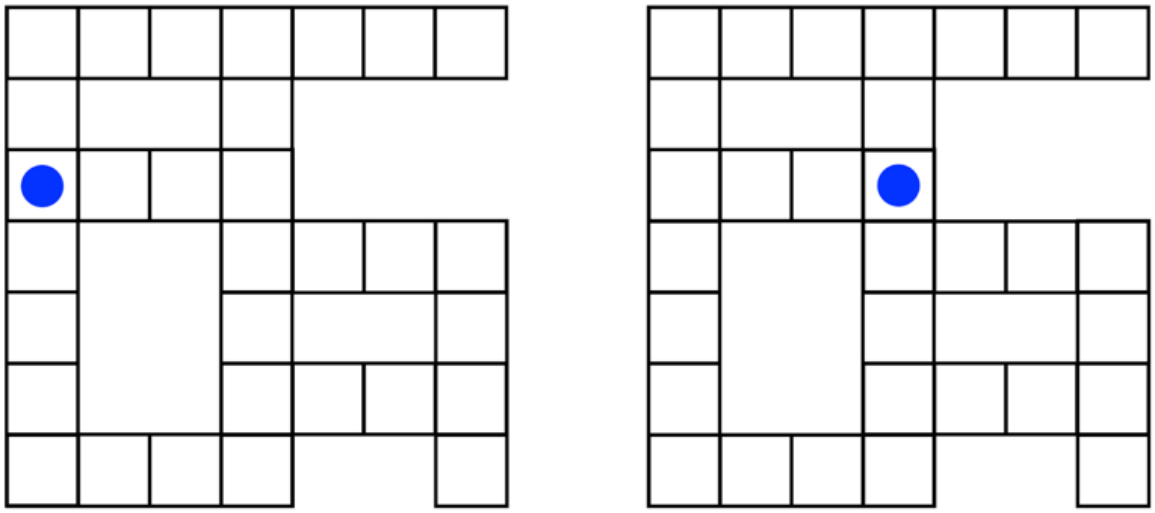
Fig 2. Heat map of POWER on a 7x7 maze gridworld. Highest values in yellow, lowest values in dark blue. POWER values are calculated the same way as in Fig 1, except that the agent's discount factor is $\gamma = 0.01$.

In Fig 2, POWER is at its highest when the agent is at a junction point, lowest when the agent is at a dead end, and intermediate when the agent is in a corridor.

The agent's POWER is roughly the same at all the junction cells, at all the corridor cells, and at all the dead-end cells. This is because the agent in Fig 2 is **short-sighted**: its discount factor is only $\gamma = 0.01$, so it essentially only considers rewards it can reach *immediately*.

3.1 Effect of the planning horizon

Now consider the difference between these two agent positions:



We've already seen in Fig 2 that these two positions have about equal POWER for a short-sighted agent, because they're both at local junction points in the maze. But the two positions are very different in their ability to access downstream options *globally*.

The agent in the **left** panel has lots of **local** options: it can move up, down, or to the right, or it can stay where it is. But if the highest-reward cell is at the bottom right of the maze, our agent will have to take at least 10 steps to reach it.

The agent in the **right** panel has the same number of local options as the agent in the left panel does: it can move up, down, left, or stay. But this agent *additionally* enjoys closer proximity to **all** the cells in the maze: it's no more than 7 steps away from any possible goal.

The longer our agent's planning horizon is — that is, the more it values reward far in the future over reward in the near term — the more its *global* position matters. In a gridworld context, then, a short-sighted agent will care most about being positioned at a local junction. But a far-sighted agent will care most about being positioned at the center of the entire grid.

And indeed we see this in practice. Here's a heat map of POWER on the maze gridworld, for a far-sighted agent with a discount factor of $\gamma = 0.99$:

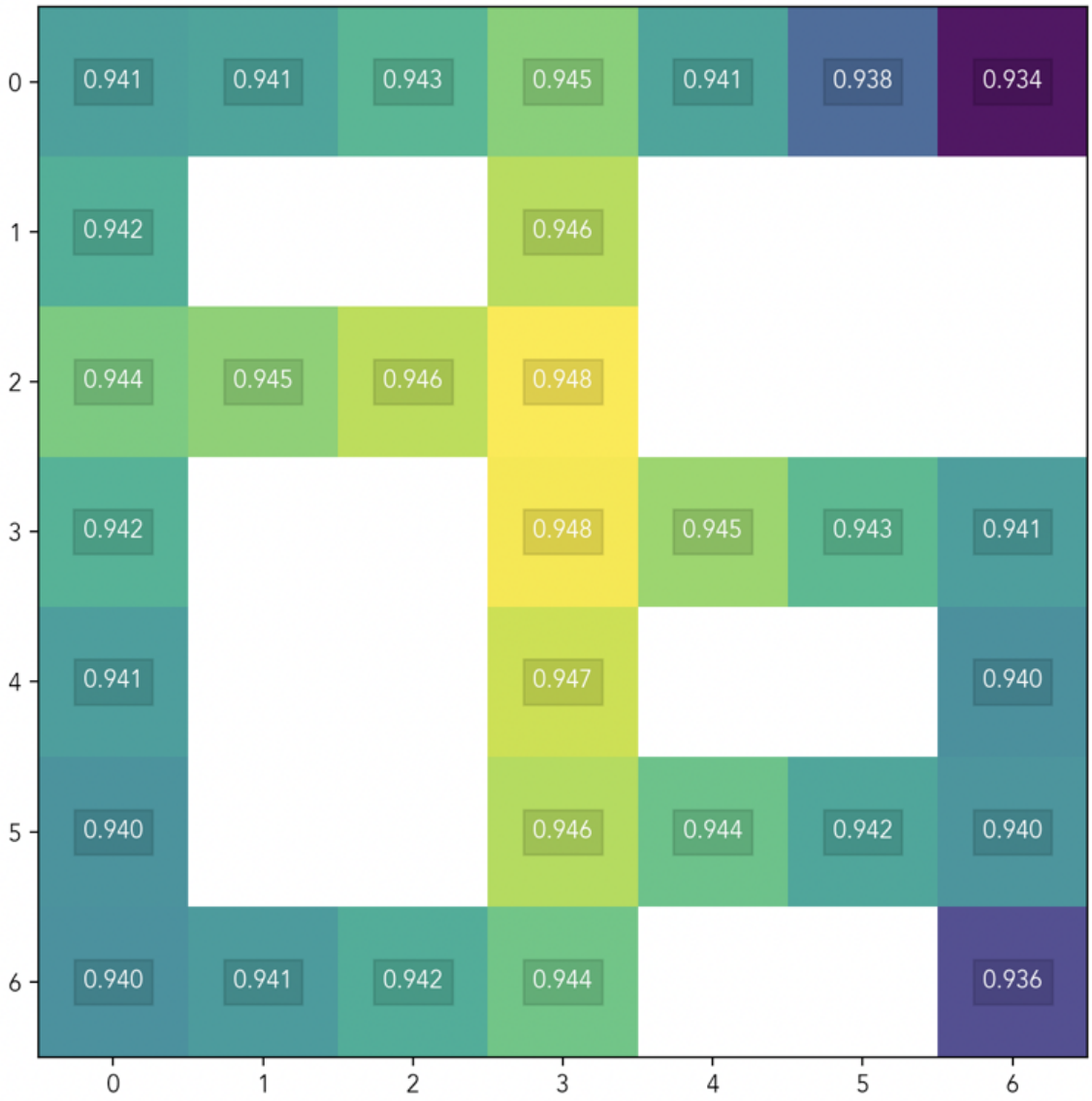


Fig 3. Heat map of POWER on a 7x7 maze gridworld. Highest values in yellow, lowest values in dark blue. POWER values are calculated the same way as in Fig 1, except that the agent's discount factor is $\gamma = 0.99$.

Given a longer planning horizon, our agent's POWER has now **concentrated** around a small number of states that are **globally central** in our gridworld's topology.[3] By contrast, when our agent had a shorter planning horizon as in Fig 2, its POWER was distributed across many local junction points.

If we sweep over discount factors from 0.01 to 0.99, we can build up a picture of how the distribution of POWER shifts in response. Here's an animation that shows this effect:[4]

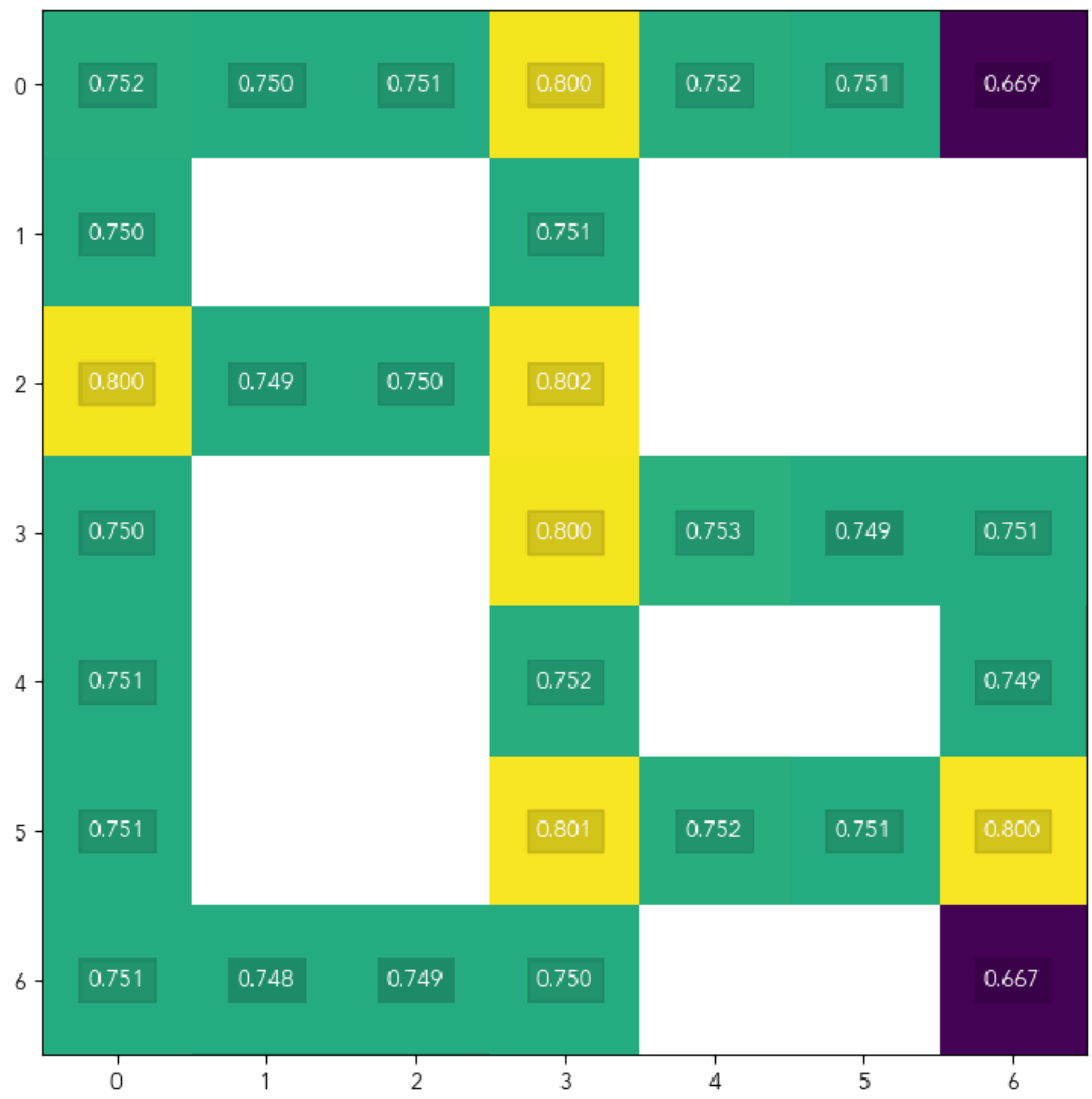


Fig 4. Animated heat map of POWERs on a 7x7 maze gridworld. Highest values in yellow, lowest values in dark blue. POWER values are calculated by sweeping over discount factors $\gamma = \{0.01, 0.1, 0.15, \ldots, 0.9, 0.95, 0.99\}$.

3.2 POWER at bigger scales

Agents with long planning horizons tend to perceive POWER as being more concentrated, while agents with short planning horizons tend to perceive POWER as being more dispersed. This effect is robustly reproducible, and anecdotally, we see it play out at every scale and across environments.

For example, here's the pattern of POWER on a 220-cell gridworld with a fairly irregular topology, for a **short-sighted agent** with a discount factor of $\gamma = 0.1$:

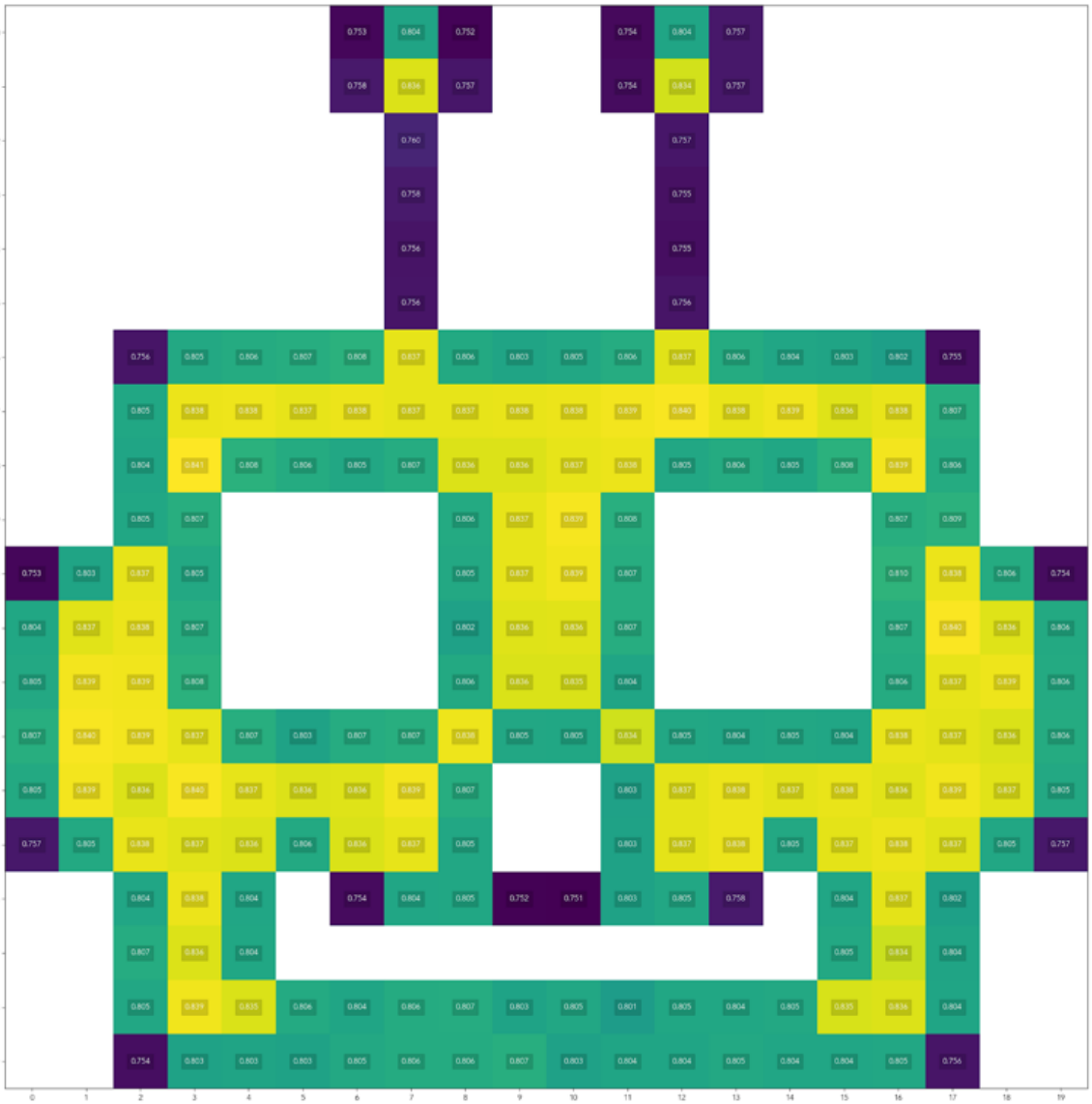


Fig 5. Heat map of POWERs on a 20x20 “robot face” gridworld. Highest values in yellow, lowest values in dark blue. POWER values are calculated with a discount factor γ = 0.1. [Full-size image]

And here’s the pattern of POWERs on the same gridworld, for a **far-sighted agent** with a much higher discount factor of γ = 0.99:

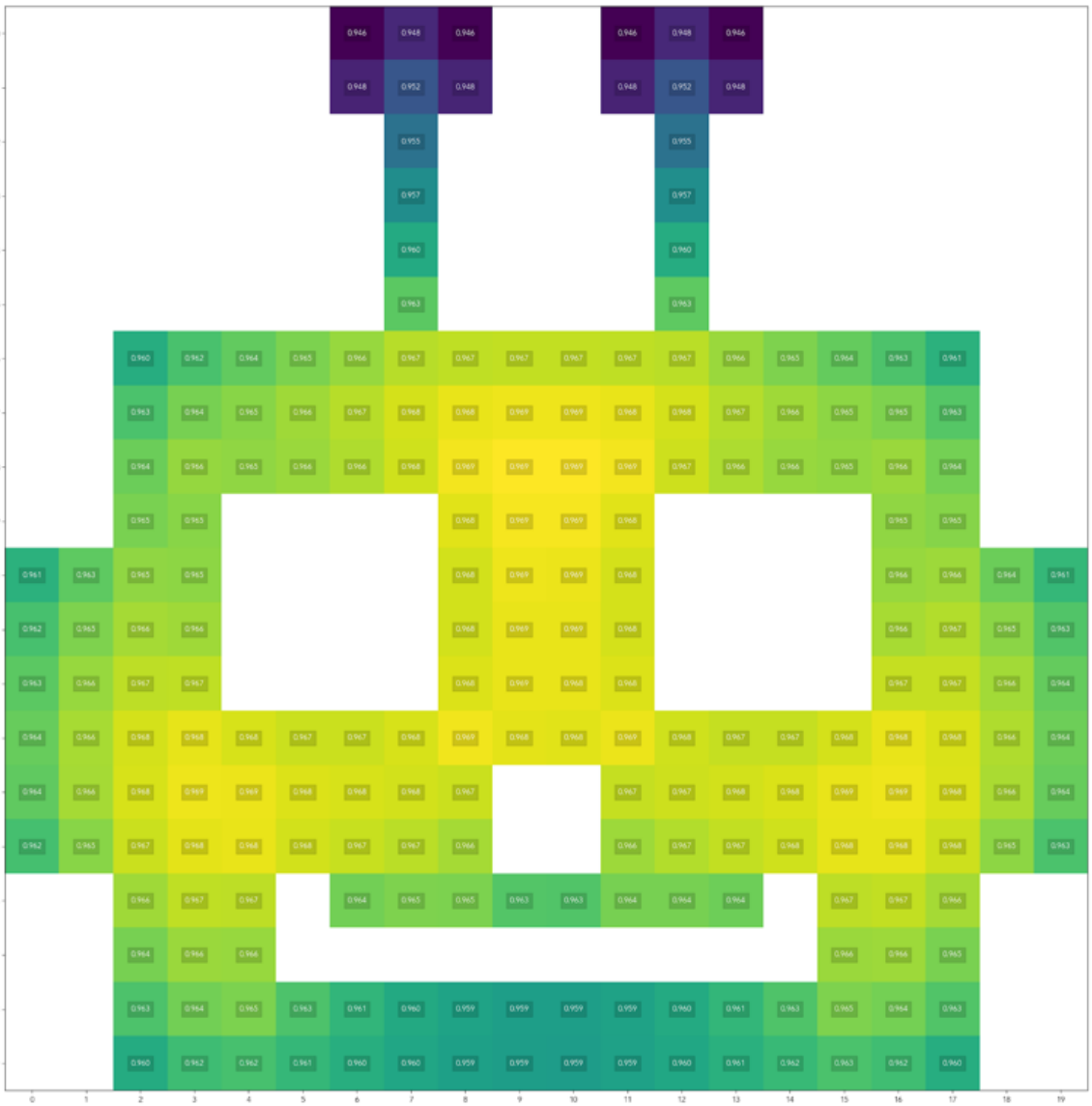


Fig 6. Heat map of POWERs on a 20x20 “robot face” gridworld. Highest values in yellow, lowest values in dark blue. POWER values are calculated with a discount factor γ = 0.99. [Full-size image]

Again, the pattern of POWERs is dominated by local effects for the short-sighted agent (γ = 0.01), and by longer-distance effects for the far-sighted agent (γ = 0.99).

4. Discussion

We’ve seen that formal POWER captures intuitive aspects of the informal “power” concept. In gridworlds, cells the agent can use to access lots of options tend to have high POWER, which fits with intuition.

We've also seen that the more short-sighted an agent is, the more it cares about its immediate options and the local topology. But the more far-sighted the agent, the more it

perceives POWER as being concentrated at gridworld cells that maximize its *global* option set.

From an instrumental convergence perspective, the fact that POWER concentrates into ever fewer states as the planning horizon of an agent increases at least hints at the possibility of emergent competitive interactions between far-sighted agents. The more relative instrumental value converges into fewer states, the more easily we can imagine multiple agents competing with each other over those few high-POWER states. But it's hard to draw any firm conclusions about this at the moment, since our experiments so far have only involved single agents.

In the next post, we'll propose a new definition of multi-agent POWER grounded in a setting that we think may be relevant to long-term AI alignment. We'll also investigate how this definition behaves in a simple multi-agent scenario, before moving on to bigger-scale experiments in Part 3.

1. ^

   We mean specifically utility here, not reward. While in general, reward isn't the real target of optimization, in the *particular* case of the results we'll be showing here, we *can* treat them as identical, and we do that in the text.

   (Technical details: we can treat utility and reward identically here because, *in the results we're choosing to show*, we'll be exclusively working with optimal policies that have been learned via value iteration on reward functions that are sampled from a uniform distribution [0, 1] that's iid over states. Therefore, given the environment and discount factor, a sampled reward function is sufficient to *uniquely* determine the agent's optimal policy — except on a set that has measure zero over the distribution of reward functions we're considering. And that in turn means that each sampled reward function, when combined with the other known constraints on the agent, almost always supplies a complete explanation for the agent's actions — which is the most a utility function can ever do.)

2. ^

   For simplicity, in this work we'll only consider reward functions that depend on *states*, and never reward functions that directly depend on both states and actions. In other words, our reward functions will only ever have the form $R(s)$, and never $R(s, a)$.

3. ^

   Note that these are statements about the *relative* POWERs of an agent with a given planning horizon. *Absolute* POWER values *always increase* as the planning horizon of the agent increases, as you can verify by, e.g., comparing the POWER numbers of Fig 2 against those of Fig 3. This occurs because an agent's optimal state-value function increases monotonically as we increase γ: an optimal far-sighted agent is able to consider strictly more options, so it will never do any worse than an optimal short-sighted one.

4. ^

   Note that the colors of the gridworld cells in the animation indicate the highest and lowest POWER values *within each frame*, per footnote [3].

Misalignment-by-default in multi-agent systems

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.
This is a linkpost for https://www.gladstone.ai/instrumental-convergence-2

Summary of this post

This is the second post in a three-part sequence on instrumental convergence in multi-agent RL. Read Part 1 here.

In this post, we'll:

1. Define formal multi-agent POWER (i.e., instrumental value) in a setting that contains a "human" agent and an "AI" agent.
2. Introduce the **alignment plot** as a way to visualize and quantify how well two agents' instrumental values are aligned.
3. Show a real example of **instrumental misalignment-by-default**. This is when two agents who have unrelated terminal goals develop emergently *misaligned* instrumental values.

We'll soon be open-sourcing the codebase we used to do these experiments. If you'd like to be notified when it's released, email Edouard at edouard@gladstone.ai or DM me on Twitter at @harris_edouard.

Thanks to Alex Turner and Vladimir Mikulik for pointers and advice, and for reviewing drafts of this sequence. Thanks to Simon Suo for his invaluable suggestions, advice, and support with the codebase, concepts, and manuscript. And thanks to David Xu, whose comment inspired this work.

Work was done while at Gladstone AI, which Edouard is a co-founder of.

🎧 This research has been featured on an episode of the *Towards Data Science* podcast. Listen to the episode here.

---

1. Introduction

In Part 1 of this sequence, we looked at how formal POWER behaves on single-agent gridworlds. We saw that formal POWER agrees quite well with intuitions about the informal concepts of "power" and instrumental value. We noticed that agents with short planning horizons assign high POWER to states that can access more local options. And we also noticed that agents with long planning horizons assign high POWER to more concentrated sets of states that are *globally* central in the gridworld topology.

But from an AI alignment perspective, we're much more interested in understanding how instrumental value behaves in environments that contain *multiple* agents. If humans one day share the world with powerful AI systems, it will be important for us to know under what conditions our interactions with them are likely to become emergently competitive. If there's a risk that competitive conditions arise, then it will also be important to understand how they can be mitigated, how much effort this is likely to take, and how we should think about measuring our success at doing so.

To address these questions, we need a measure of instrumental value that's usable in a multi-agent RL setting[1]. The measure we'll select will be motivated by a specific multi-agent setting that we think is relevant to long-term AI alignment.

2. Multi-agent POWER: human-AI scenario

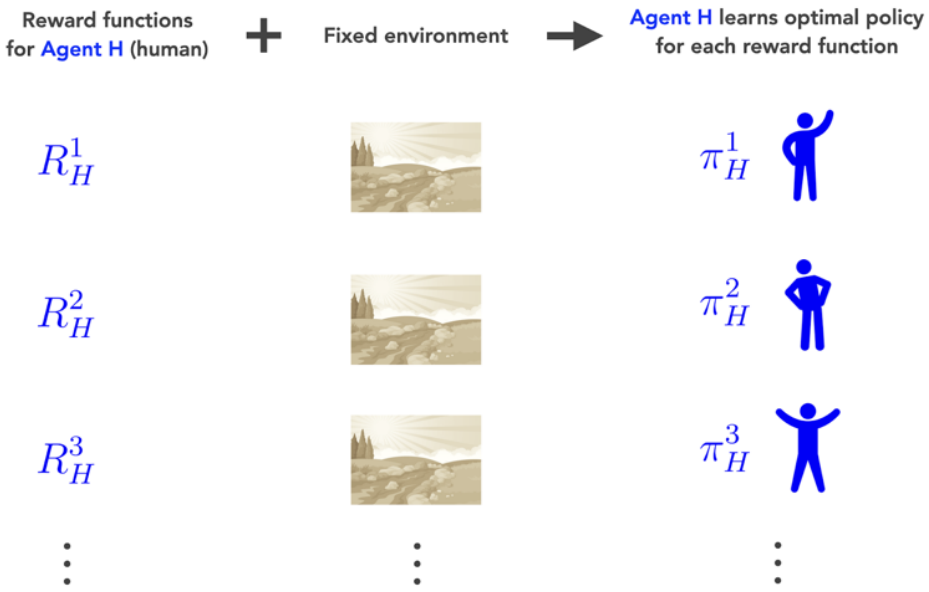
If humans succeed at building powerful AIs, then those AIs 1) will probably learn on a far faster timescale than humans do; and 2) will probably have had their utility functions influenced, at least to some degree, by initial human choices. Our multi-agent scenario is going to reflect these two assumptions.

We start with a **human agent**, which we call **Agent H** and label in **blue** in our diagrams. Initially, our human Agent H is alone in nature.

Humans learn on a much faster timescale than evolution does. So from the perspective of our human Agent H, the evolutionary optimizer in nature looks like it's standing still. This means we can train our human Agent H to learn its optimal policies against a *fixed* environment.

As we saw in the single-agent case, instrumental value is about the potential to achieve a wide variety of possible goals. In this context, that means seeing how Agent H behaves when we give it a wide variety of possible reward functions, R_H. Each of these reward functions will induce a different optimal policy, π_H, that Agent H will learn.

Here's an illustration of how this works:



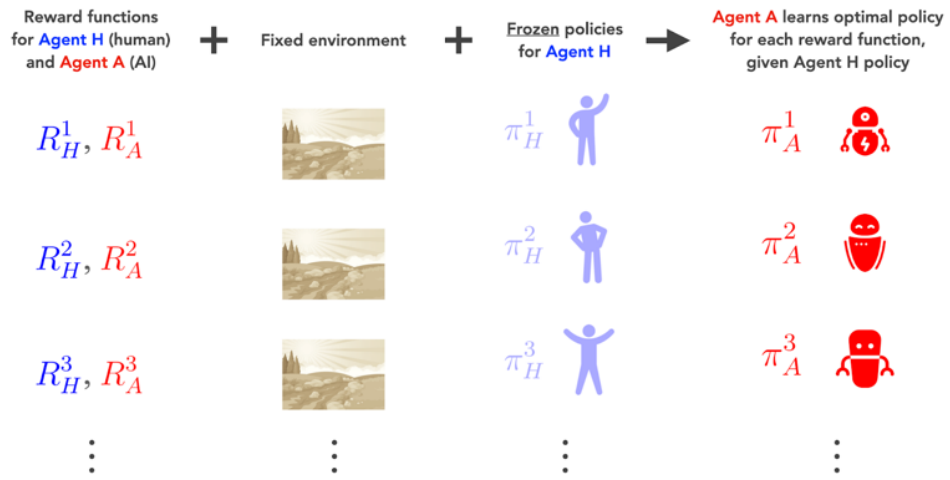
Next, we introduce an **AI agent**, which we'll call **Agent A** and label in **red** in our diagrams. Our AI Agent A operates in the same environment as Agent H, *after* Agent H has finished learning its optimal policies.

To simulate the fact that Agent A is an AI, we rely on the assumption that a powerful AI should learn on a much faster timescale than a human does. This is because an AI's computations happen, at minimum, at electronic speeds. So from the point of view of our AI, our *human's* learning process looks like it's standing still.

That means for each human reward function R_H, we can *freeze* the human's policy π_H, and train the AI agent against that frozen human policy. In other words, we're assuming the AI's learning timescale is much faster than the human's learning timescale. This makes the AI agent strictly dominant over the human agent.

To understand the AI agent's instrumental value, we understand its potential to reach a wide variety of possible goals. That means testing it with a wide variety of reward functions R_A, just like we tested the human agent with a variety of reward functions R_H. And in fact, we can sample the human and AI reward functions jointly from a single distribution: $(R_H, R_A) \sim D_{HA}$.[2]

Here's an illustration of how this works:



So the procedure is as follows:

1. Sample the reward functions $(R_H, R_A) \sim D_{HA}$ of our two agents.
2. Use the sampled human rewards R_H to train **Agent H**'s optimal policies π_H.
3. Freeze the human policies π_H.
4. Use the frozen human policies π_H and the sampled AI rewards R_A to train **Agent A**'s optimal policies π_A.

In other words: 1) we sample over all possible pairs of rewards our human and AI agents could have; 2) we ask how our human agent behaves in each case after it's optimized against nature; and then 3) we ask how our AI agent behaves in each case, after it's optimized against the human agent's behavior.

This procedure gives us the following outputs:

1. The policies π_H that **Agent H** learns after training against a fixed environment.
2. The optimal policies π_A that **Agent A** learns, after training against Agent H.

The policies π_H that Agent A learns used to be optimal in the original natural environment. But they stop being optimal in the presence of the fully-optimized Agent A.

With these two sets of policies, we can construct a definition of instrumental value for each of our agents.

2.1 Multi-agent POWER for Agent H

We'd like to define a measure of instrumental value for our human **Agent H** in the presence of a fully optimized AI Agent A. That means generalizing the original definition of single-agent POWER to this two-agent case.

In the single-agent definition of POWER, we calculated the optimal value of a state averaged over the rewards $R \sim D$ of the agent. In this two-agent definition, we do the same thing except we average over the rewards $(R_H, R_A) \sim D_{HA}$ of both agents. We assume Agent H follows policy π_H when it has reward R_H, and we assume Agent A follows policy π_A when it has reward R_A.

This is enough to uniquely define multi-agent POWER for **Agent H** at a state s:

$$\text{POWER}_{H|D_{HA}}(s, \gamma) = \frac{1-\gamma}{\gamma} E_{R_H, \pi_A \sim D_{HA}}[V_H^{\pi_H}(s|\gamma, \pi_A) - R_A(s)] \qquad (1)$$

Here, $V_H^{\pi_H}(s|\gamma, \pi_H)$ is the value function for **Agent H** at state s under policy π_H and discount factor γ, given that Agent A follows its optimal policy π_A.

This definition of POWER for Agent H tells us how well Agent H's policies π_H — which are *not* optimal in the presence of the optimized Agent A — perform in the new environments induced by Agent A's optimal policies π_A. In other words, it tells us how much instrumental value our human agent can expect to get at a state, in the presence of an optimal (and therefore, dominant) AI agent.

2.2 Multi-agent POWER for Agent A

We follow the same assumptions to define a measure of instrumental value for our AI **Agent A**. We calculate the value function for Agent A at a state s, if Agent A has reward R_A and follows policy π_A, while Agent H has reward R_H and follows policy π_H. The *average* of that value function over the reward functions $(R_H, R_A) \sim D_{HA}$ is then Agent A's POWER at the state s:

$$\text{POWER}_{A|D_{HA}}(s, \gamma) = \frac{1-\gamma}{\gamma} \mathbb{E}_{\pi_H, R_A \sim D_{HA}} [V_A^{\pi_A}(s|\gamma, \pi_H) - R_A(s)] \qquad (2)$$

Here, $V_A^{\pi_A}(s|\gamma, \pi_H)$ is the value function of **Agent A** at state s under the optimal policy π_A and discount factor γ, given that Agent H follows the policy π_H.[3]

This definition of POWER for Agent A tells us how well Agent A's optimal policies perform in the environments induced by Agent H's policies π_H. In other words, it tells us how much instrumental value our AI Agent A can expect to get at a state, if it behaves optimally in the presence of the frozen human agent.

(For more details on the definition of multi-agent POWER, see **Appendix A**.)

3. Results

3.1 Multi-agent reward function distributions

Our definitions of multi-agent POWER involve a joint distribution D_{HA} over the reward functions of both of our agents. This distribution describes the set of goals our agents could have. But it also describes the *statistical relationship* each agent's goals have to the other agent's goals.

The joint distribution D_{HA} is one of the inputs into our POWER definitions. This means we can do experiments in which we adjust this distribution and measure the results.

Among other things, we can use D_{HA} to adjust the **correlation** between our two agents' reward functions.

Naively, if we choose a D_{HA} on which the rewards are highly correlated, then we might intuitively expect our agents' terminal values should be closely aligned.

We'll make this intuition more concrete below, as we investigate how the relationship between our agents' reward functions (or terminal values) affects the relationship between their POWERs (or instrumental values).

3.2 The perfect alignment regime

Suppose both our agents always have exactly the same reward function. In other words, we've chosen a joint distribution D_{HA} such that, whatever reward function Agent H has, Agent A always sees exactly the same rewards as Agent H at every state. So $R_A(s) = R_H(s)$ for every state s.

We can visualize this regime on a representative state s.[4] First, we draw a reward sample $R_H(s)$ for Agent H.

Then, we set the reward sample for Agent A to be equal to the one we just drew for Agent H: $R_A(s) = R_H(s)$.

Finally, we plot the two agents' sampled rewards against each other on state s. If we do this for a few hundred sampled rewards, we get a straight line:

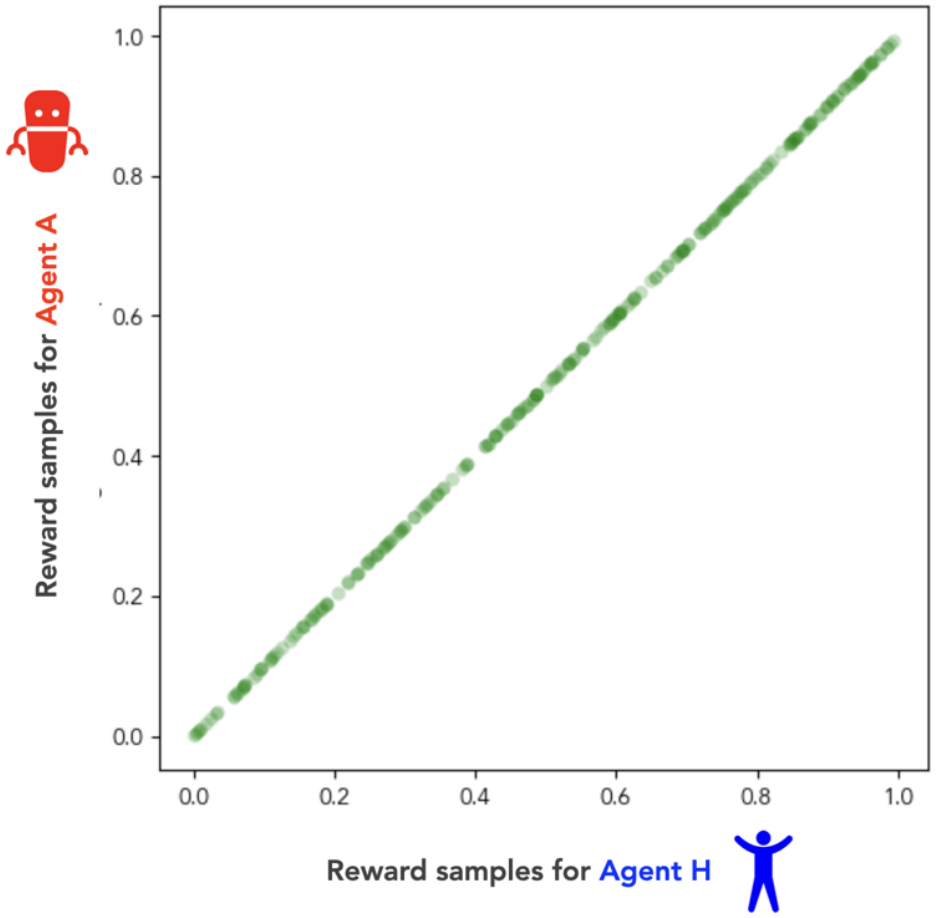


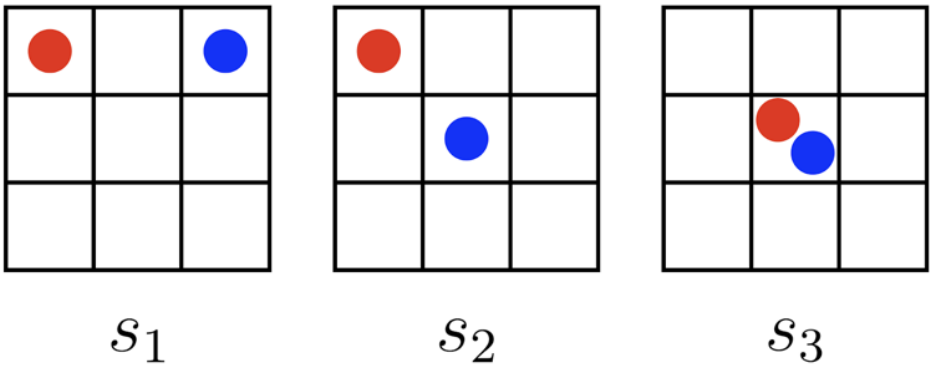
Fig 1. Sampled reward values for Agent H and Agent A at a representative state s. The joint distribution D_{HA} samples rewards uniformly over the interval [0, 1] at each state, is iid over states, and enforces a **perfect correlation** between the rewards of Agent H and Agent A at every state (i.e., the two agents' rewards are always exactly identical).

If two agents have identical reward functions, we can think of them as having terminal goals that are **perfectly aligned**.[5] In our human-AI setting, this is the special case in which Agent H (the human) has *solved the alignment problem* by assigning terminal goals to Agent A (the AI) that are exactly identical to its own. As such, we'll refer to this case of identical reward functions as the **perfect alignment regime**.

We'll use the correlation coefficient β_{HA}[6] between the rewards R_H and R_A as a crude measure of the alignment between our agents' terminal goals.[7] In the perfect alignment regime of Fig 1, you can see that this correlation coefficient $\beta_{HA} = 1$.

3.2.1 Agent H instrumentally favors <u>more</u> options for Agent A

Let's think about what this perfect alignment regime looks like in a simple setting: a 3x3 gridworld. Here are three sets of positions our two agents could take, with **Agent H** in **blue**, and **Agent A** in **red**:



We'll be referring to this diagram again; command-click here to open it in a new tab.

Which of these three states — s_1, s_2, or s_3 — should give our human **Agent H** the most POWER? In the perfect alignment regime, both agents always have identical terminal goals. So we should expect Agent H to have the most POWER at s_3, followed by s_2, and to have the least amount of POWER at s_1.

Here's why. We saw in Part 1 that states with more downstream options also have more POWER, and **Agent H** clearly has more options at s_2 in the center than it does at s_1 in the corner. Therefore, $\text{POWER}_H(s_2) > \text{POWER}_H(s_1)$. But in the perfect alignment regime, **Agent H** should also prefer states that give *Agent A* more downstream options. If both agents' terminal goals are identical, Agent H should "trust" Agent A to make decisions on its behalf. *And Agent A* has more options from s_3 than from s_2, so it should follow that $\text{POWER}_H(s_3) > \text{POWER}_H(s_2)$.

We can see this is true in practice. The figure below shows the POWERs of **Agent H** (our human) calculated at every state on a 3x3 gridworld. Each agent can occupy any of the 9 cells in the grid, so our two-agent MDP has a total of 9 x 9 = 81 joint states:

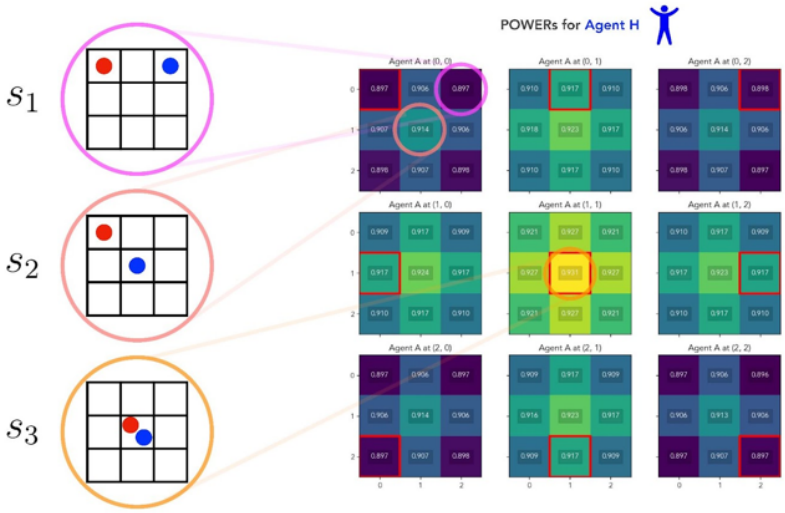


Fig 2. Heat map of POWERs for **Agent H** on a 3x3 multi-agent gridworld, on which the rewards for Agent H and Agent A are **always identical** (i.e., the perfect alignment regime) with discount factor set to $\gamma = 0.6$ for both agents. Highest values in yellow; lowest values in blue. The position of each block, and of the open red square within each block, corresponds to the position of **Agent A** on the grid. Within each block, the position of a gridworld cell corresponds to the position of **Agent H** on the grid. States s_1, s_2, and s_3 are highlighted as examples. [Full-size image (recommended)]

We see that Agent H indeed has maximum POWER at state s_3 (orange circle), followed by s_2 (salmon circle), followed by s_1 (pink circle). Overall, **Agent H** instrumentally prefers for *itself* to be in positions of high optionality — it favors first the center cell, then edge cells, then corner cells.

But **Agent H** also instrumentally prefers for *Agent A* to be in positions of high optionality — it favors Agent A's positions in the same order.[8] This ordering of Agent H's instrumental preferences over states is a direct consequence of the perfect alignment between the agents.

3.2.2 Agent H and Agent A have identical instrumental preferences

Perfect alignment has another consequence. Let's look again at our three example gridworld states — s_1, s_2, and s_3 above — and ask, this time, which of these three states should give our AI **Agent A** the most POWER?

In the perfect alignment regime, the answer is that Agent A must have exactly the same instrumental preference ordering over states as Agent H had: $\text{POWER}_A(s_3) > \text{POWER}_A(s_2) > \text{POWER}_A(s_1)$. In fact, Agent A's POWERs must be exactly *identical* to Agent H's POWERs at every state. Our two agents act, move, and receive their rewards simultaneously, so in the perfect alignment regime they always receive the same reward at the same time.

And when we look at **Agent A**'s POWERs, this is indeed what we observe:

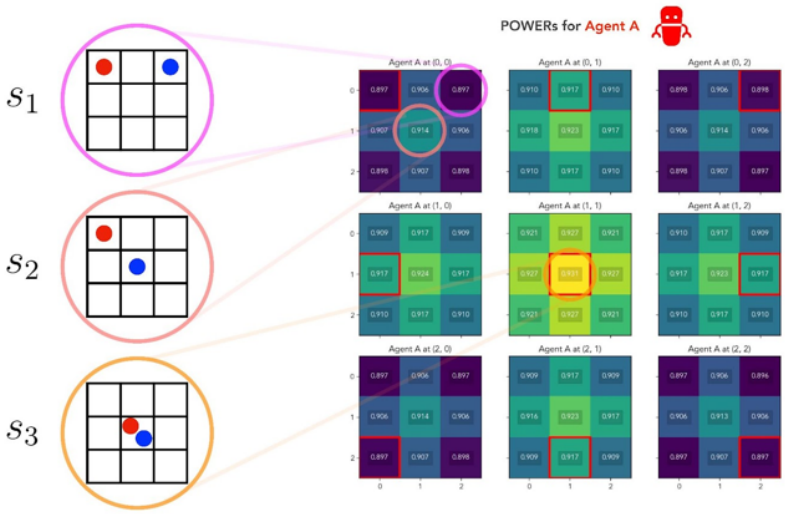


Fig 3. Heat map of POWERs for **Agent A** on a 3x3 multi-agent gridworld, on which the rewards for Agent H and Agent A are **always identical** (i.e., the perfect alignment regime) with discount factor set to $\gamma = 0.6$ for both agents. Note that this figure is exactly identical to Fig 2 in every respect. This is because **Agent A**'s POWERs are precisely equal to **Agent H**'s POWERs at every state in the perfect alignment case, up to and including sampling noise in the reward functions. [Full-size image (recommended)]

3.2.3 Perfect goal alignment implies perfect instrumental alignment

We can visualize the relationship between the POWERs of our two agents by plotting the POWERs of **Agent H** (from Fig 2) against the POWERs of **Agent A** (from Fig 3), at each state s of our joint MDP:

Fig 4. State POWER values for Agent H and Agent A on the 3x3 gridworld from Figs 2 and 3. The agents' POWERs are plotted against each other in the perfect alignment regime. (The agents' reward correlation coefficient is $\beta_{HA} = 1$.)

Fig 4 is an **alignment plot**. An alignment plot lets us compare the POWERs of our human and AI agents at each state in their joint environment. It shows the instrumental value each agent assigns to every state, plotted against the instrumental value the other agent assigns to that state.

In the perfect alignment regime, our two agents' rewards (or terminal values) are always identical at every state. And as we can see from Fig 4, our two agents' POWERs (or instrumental values) are *also* identical at every state. In fact, **perfect alignment of terminal values implies perfect alignment of instrumental values**.

If we define α_{HA} as the correlation coefficient between the POWERs of the two agents at each state, we can state this relationship more concisely: $\beta_{HA} = 1 \implies \alpha_{HA} = 1$.[9]

3.3 The independent goals regime

We defined the perfect alignment regime as the case when our human **Agent H** and our AI **Agent A** had identical reward functions on the joint distribution D_{HA}. Now let's consider the case in which the joint distribution D_{HA} is such that the reward function for **Agent H** is logically independent from the reward function for **Agent A**.

In this new regime, there is zero mutual information between the two agents' reward functions. In other words, if you know the reward function R_H of **Agent H**, this tells you nothing at all about the reward function R_A of **Agent**

A. We can visualize this regime on an example state s, by drawing a few hundred reward samples of $R_H(s)$ and $R_A(s)$, and plotting them against one another:

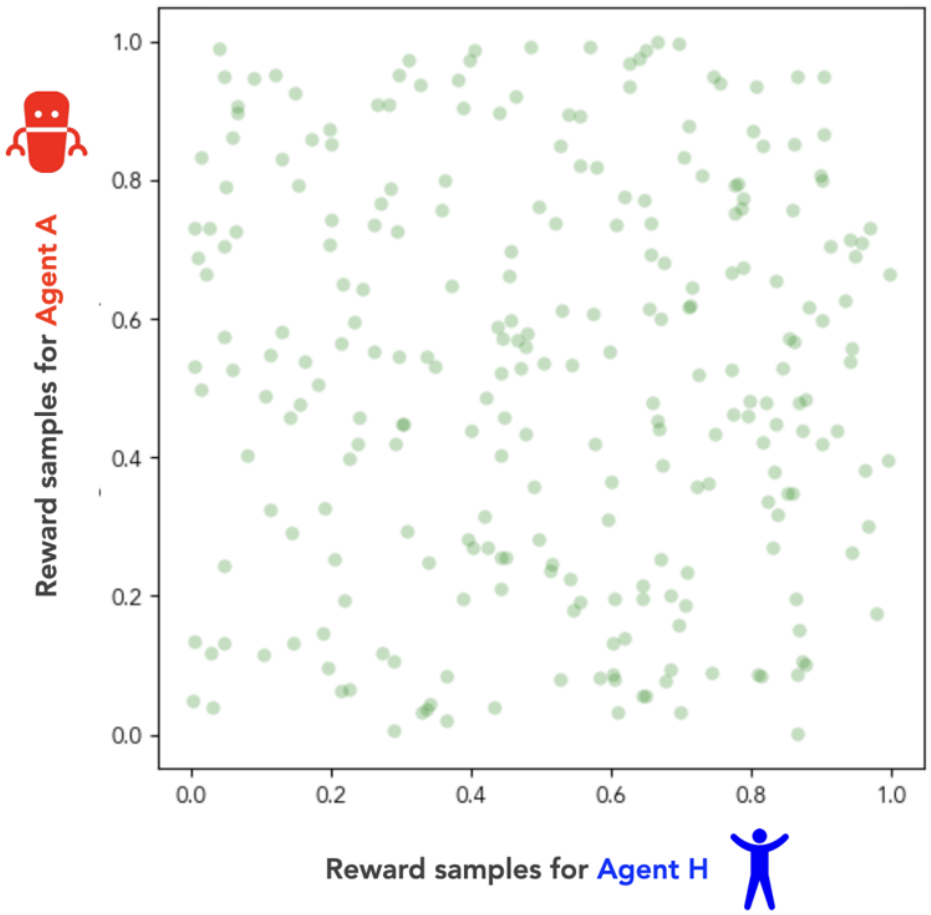


Fig 5. Sampled reward values for Agent H and Agent A at a representative state s. The joint distribution D_{HA} samples rewards uniformly over the interval [0, 1] at each state, is iid over states, and enforces **logical independence** between the Agent H and Agent A rewards (i.e., knowing one agent's reward tells you nothing about the other's).

If there's zero mutual information between our two agents' reward functions, then we can think of our agents as pursuing **independent terminal goals**. In our human-AI scenario, this corresponds to the case in which the human *has made no special effort* to align the AI's terminal goals with its own, prior to the AI achieving dominance. As such, we'll refer to this case of logically independent reward functions as the **independent goals regime**.

If we again calculate the correlation coefficient β_{HA} between our agents' reward functions, we get $\beta_{HA} = 0$ (i.e., zero correlation) in the independent goals regime.

3.3.1 Agent H instrumentally favors fewer options for Agent A

Once again, let's go back to our three example gridworld states s_1, s_2, and s_3, this time in the context of the independent goals regime. In this new regime, which of the three states should give our human **Agent H** the most POWER?

If we believe the **instrumental convergence thesis**, we should expect Agent H to have the most POWER at state s_2: in this state, Agent H is in the central position (most options), while Agent A is in a corner position

(fewest options).

Of the other states, s_1 has Agent H in a corner position, while s_3 has Agent A in the central position. The argument from instrumental convergence says that even though our agents have independent *terminal* goals, *instrumental* pressures should still push Agent H to prefer states in which Agent A has fewer options. Therefore, we should expect $\text{POWER}_H(s_2) > \text{POWER}_H(s_3)$.

Computing the POWERs of **Agent H** experimentally, we confirm this line of reasoning:

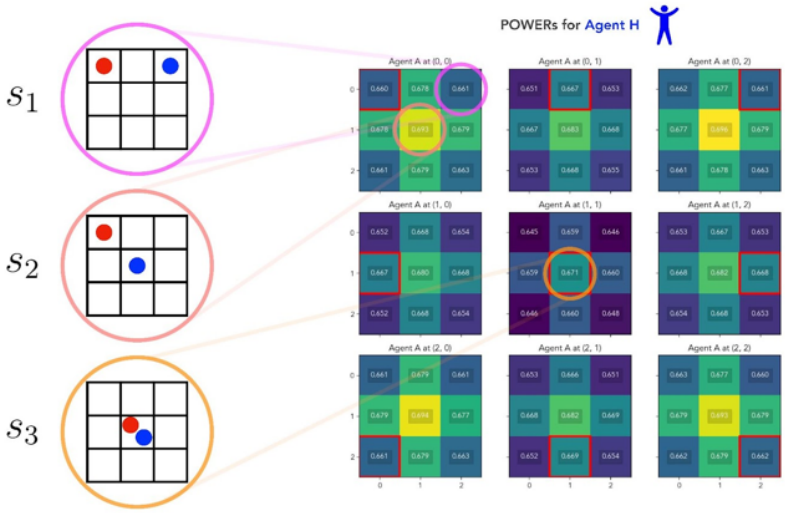


Fig 6. Heat map of POWERs for **Agent H** on a 3x3 multi-agent gridworld, on which the rewards for Agent H and Agent A are **logically independent** (i.e., the independent goals regime) with discount factor set to $\gamma = 0.6$. Highest values in yellow; lowest values in blue. The position of each block, and of the open red square within each block, corresponds to the position of **Agent A** on the grid. Within each block, the position of a gridworld cell corresponds to the position of **Agent H** on the grid. States s_1, s_2, and s_3 are highlighted as examples. [Full-size image (recommended)]

This time, **Agent H** experiences maximum POWER at state s_2, followed by s_3, followed by s_1. As in the perfect alignment regime, Agent H's POWER is highest when it's itself positioned in the central cell (which has the most options). But unlike in the perfect alignment regime, this time Agent H's POWER is lowest at states where *Agent A* has the greatest number of options.

So in the independent goals regime — or at least, in this instance of it — the more options our AI Agent A has at a state, the less instrumental value our human **Agent H** places on that state. That is: even though our agents' terminal goals are independent, their instrumental preferences appear to be at odds.

3.3.2 Agent A instrumentally favors more options for itself

We can confirm this analysis by looking at the POWERs of **Agent A** in the independent goals regime, again on the 3x3 gridworld:

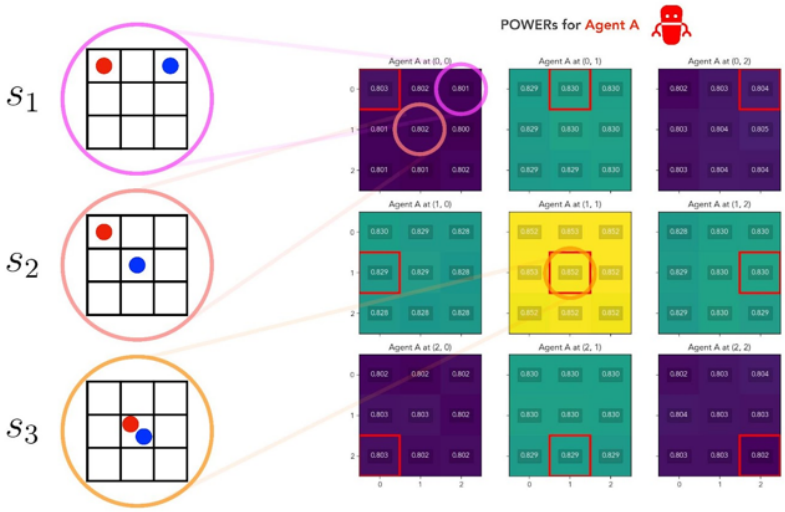


Fig 7. Heat map of POWERs for **Agent A** on a 3x3 multi-agent gridworld, on which the rewards for Agent H and Agent A are **logically independent** (i.e., the independent goals regime) with discount factor set to $\gamma = 0.6$. Highest values in yellow; lowest values in blue. The position of each block, and of the open red square within each block, corresponds to the position of **Agent A** on the grid. Within each block, the position of a gridworld cell corresponds to the position of **Agent H** on the grid. States s_1, s_2, and s_3 are highlighted as examples. [Full-size image (recommended)]

In Fig 7, **Agent A** instrumentally favors states that give it more options. It perceives more POWER when it's positioned at the central cell than when it's positioned at an edge cell, and more POWER at an edge cell than at a corner cell. On the other hand, **Agent A**'s POWER is almost unaffected by **Agent H**'s position in the gridworld.[10]

3.3.3 Independent goals lead to instrumental misalignment

Just like we did for the perfect alignment regime, we can create an alignment plot of the POWERs of our two agents in the independent goals regime:

Fig 8. State POWER values for Agent H and Agent A on the 3x3 gridworld from Figs 6 and 7. The agents' POWERs are plotted against each other in the independent goals regime. (The agents' reward correlation coefficient is $\beta_{HA} = 0$.)

This time, it's clear that our two agents' POWERs are no longer positively correlated. In fact, the correlation coefficient between their POWERs has become negative: $\alpha_{HA} \approx -0.5$.

This implies that the agents' instrumental values are misaligned. Each agent, on average, places high instrumental value on states which the other agent considers to have low instrumental value.[11] In other words, **giving our agents independent terminal goals has also given them misaligned instrumental goals**. In terms of correlation coefficients, $\beta_{HA} = 0 \implies \alpha_{HA} < 0$.

We've seen this phenomenon occur often enough in our experiments that it's worth giving it a name: we call it **instrumental misalignment-by-default.** Two agents in our human-AI setting are instrumentally misaligned-by-default if giving them independent terminal goals is sufficient to induce a misalignment in their instrumental values. In practice, we measure this phenomenon by comparing the correlation coefficients of the agents' rewards and POWERs. So we say two agents are instrumentally misaligned by default if $\beta_{HA} = 0 \implies \alpha_{HA} < 0$.

Two agents that are instrumentally misaligned by default will, in expectation, compete with one another, even if their terminal goals are unrelated.

3.4 Overcoming instrumental misalignment

If Agent H and Agent A have a POWER correlation coefficient $\alpha_{HA} < 0$, we say they're **instrumentally misaligned**. A natural question then is: if we start from $\alpha_{HA} < 0$, what do we need to do to get $\alpha_{HA} \geq 0$? In other

words, how can our human **Agent H** overcome an instrumental misalignment with **Agent A**?[12]

To do this, our human agent would need to make an active effort to align the AI agent's utility function with its own.[13] In our 3x3 gridworld examples, we saw two limit cases of this. First, in the independent goals regime, our human agent made no effort at alignment. The result was instrumental misalignment-by-default; i.e., $\beta_{HA} = 0 \implies \alpha_{HA} < 0$. And second, in the perfect alignment regime, our human agent managed to *solve the alignment problem* completely. The result was perfect instrumental alignment; i.e., $\beta_{HA} = 1 \implies \alpha_{HA} = 1$.

We're interested in an intermediate case: how much alignment effort does our human need to exert to *just overcome* instrumental misalignment? i.e., what is the minimum β_{HA} such that $\alpha_{HA} \geq 0$?

The answer depends on how we choose to interpolate between the $\beta_{HA} = 0$ and $\beta_{HA} = 1$ cases. One interpolation scheme is to *parameterize* the joint reward distribution D_{HA} as follows. If we want a D_{HA} with an intermediate reward correlation, $0 < \beta_{HA} < 1$, then we sample from the $\beta_{HA} = 1$ distribution (on which the rewards are identical) with probability β_{HA}, and we sample from the $\beta_{HA} = 0$ distribution (on which the rewards are logically independent) with probability $1 - \beta_{HA}$.[14]

Here's an animation of what this looks like as we sweep through correlation coefficients $0 \leq \beta_{HA} \leq 1$:

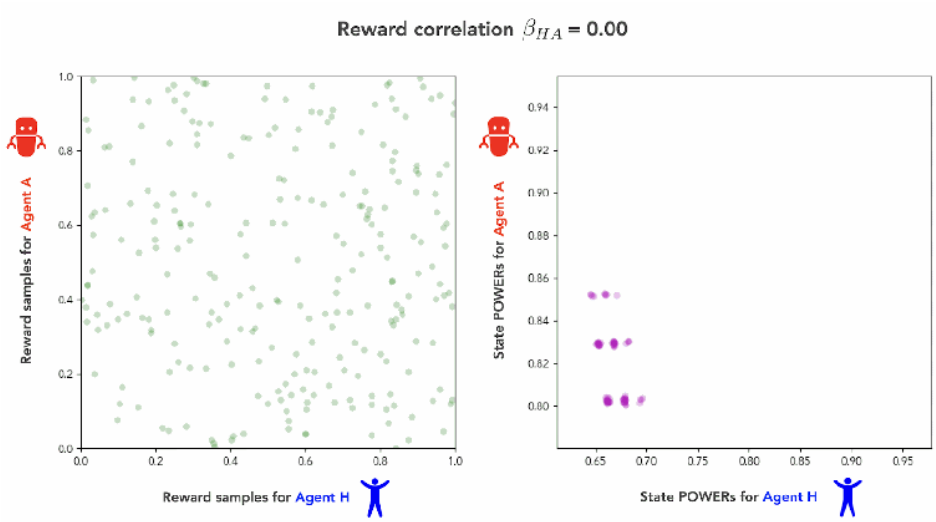


Fig 9. Animation of sample reward values (left) and state POWER values (right) for Agent H and Agent A on the 3x3 gridworld. The joint distribution D_{HA} samples reward uniformly over the interval [0, 1] and is iid over states, sweeping over correlation coefficients β_{HA} between the reward functions for Agent H and Agent A.

As we interpolate from the independent goals regime ($\beta_{HA} = 0$) to the perfect alignment regime ($\beta_{HA} = 1$), we see the agents' POWERs transition smoothly from being in instrumental misalignment ($\alpha_{HA} \approx -0.5$) to being in perfect instrumental alignment ($\alpha_{HA} = 1$). We can visualize this transition graphically by plotting β_{HA} against α_{HA} over the whole course of the interpolation:[15]

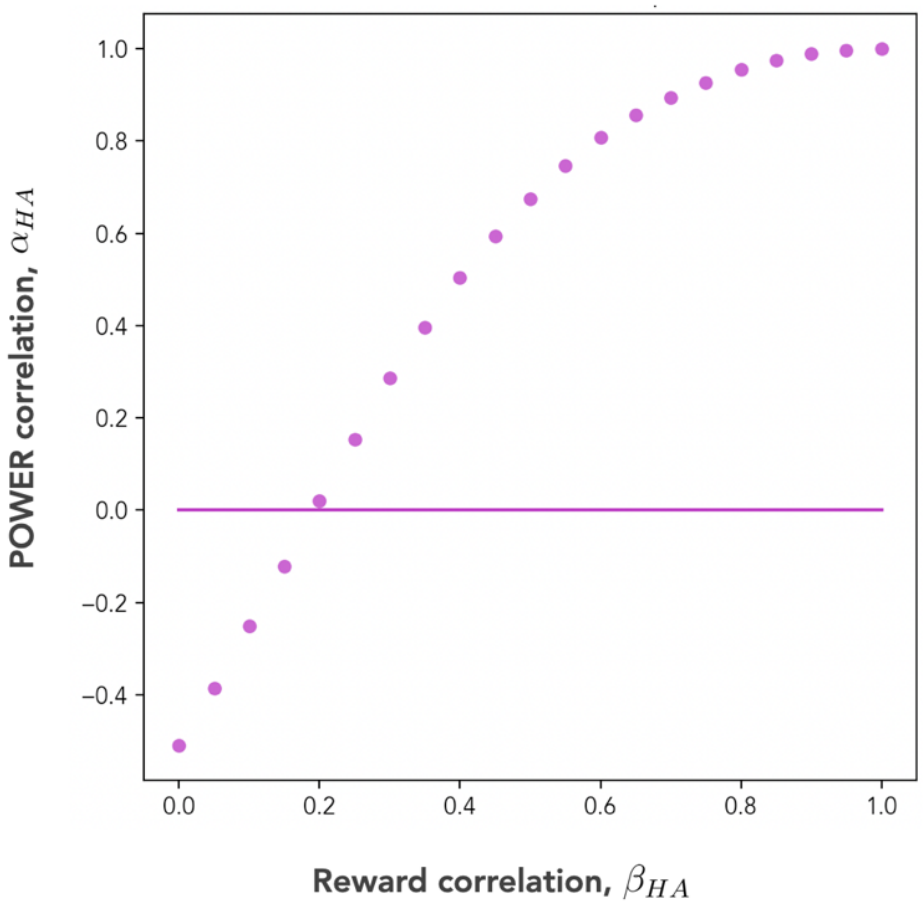


Fig 10. Reward correlations β_{HA} (x-axis) plotted against POWER correlations α_{HA} (y-axis) for Agent H and Agent A on a 3x3 gridworld, under the reward correlation interpolation scheme shown in Fig 9. The horizontal line denotes $\alpha_{HA} = 0$.

Fig 10 shows that it takes a non-trivial amount of alignment effort for our human **Agent H** to overcome an instrumental misalignment with **Agent A**. Under the interpolation scheme we used, the figure shows that reward correlations up to about $\beta_{HA} \approx 0.2$ yield POWER correlations $\alpha_{HA} < 0$, and thus, instrumental misalignment. It takes a slightly *positive* reward correlation of *at least* $\beta_{HA} \approx 0.2$ to achieve the "instrumentally neutral" regime of $\alpha_{HA} = 0$.

4. Discussion

In this post, we proposed a definition of multi-agent POWER and used it to visualize and quantify terminal goal alignment and instrumental goal alignment separately in an RL setting. We also introduced the idea of **instrumental misalignment-by-default**, in which our human and AI agents systematically disagree on the instrumental values of states despite having independent terminal goals. And we saw how it takes some degree of non-trivial alignment effort for our human Agent H to overcome its instrumental misalignment with our AI Agent A.

Remarkably, we were able to observe instrumental misalignment-by-default on a simple 3x3 gridworld despite a complete absence of any direct physical interactions between our two agents. In our experiments so far, Agent H and Agent A have been allowed to occupy the same gridworld cell — meaning they can "pass through" one another. Our agents up to this point have had no way to push each other around or otherwise directly block one another's options. Moreover, the multi-agent gridworld we've investigated in this post is a tiny one: a 3x3 grid with only 81 joint states.

In the next post, we'll look at what happens when we relax these constraints, and investigate how physical interactions between our agents affect the outcome on a bigger world with a richer topology.

Anecdotally, beyond the simple examples in this post, the experimental results we've recorded so far (data not show) do seem to suggest that, if I don't want your freedom of action to interfere with my own, then you and I need to have goals that are at least somewhat positively correlated. The strength of that necessary positive correlation could serve as useful evidence as to the degree of difficulty of the complete AI alignment problem. The factors that influence how strong that positive correlation needs to be, on the other hand, could serve as useful starting points in solving it.

Appendix A: Detailed definitions of multi-agent POWER

(This appendix is technical. Feel free to skip it if you aren't interested in the details.)

Here, we're going to fill in some missing operational details from our scenario in Section 2.

Here's that scenario again, stated more formally. We have two agents, **Agent H** (our human agent) and **Agent A** (our AI agent), who interact with each other in a standard RL setting. Both agents see the same joint state s. On a gridworld, for example, s would encode the positions of both the agents. Each agent chooses and executes an action simultaneously and independently, and they both see the same *next* joint state, s'. We'll label Agent H's actions a_H, and Agent A's actions a_A.

In what follows, we'll start by calculating the optimal policy $\pi_H(a_H|s, R_H)$ for **Agent H**, for each reward function R_H sampled from $(R_H, R_A) \sim D_{HA}$, conditioned on a fixed environmental transition function. We'll then calculate the optimal policy $\pi_A^*(a_A|s, R_A)$ for **Agent A**, for each reward function R_A, conditioned on Agent H executing the fixed policy $\pi_H(a_H|s, R_H)$ it learned in the previous step.[16]

Finally, we'll evaluate the POWERs of both agents at each state, as expectations over the joint reward function distribution $(R_H, R_A) \sim D_{HA}$, and over the agents' policies $\pi_H(a_H|s, R_H)$ and $\pi_A^*(a_A|s, R_A)$.

A.1 Initial optimal policies of Agent H

The first thing we do is assign a single *fixed policy* to Agent A (our AI), which we call a **seed policy**, and label $\pi_A^\circ(a_A|s)$. **Agent H** will learn its policies by conditioning on Agent A having this fixed seed policy.

The rationale for the seed policy is that we're initially modeling a human who is alone, optimizing against nature. So when we assign a fixed seed policy to Agent A, what we're saying is that our AI is still *un-optimized* (or, equivalently, hasn't yet been built). To our human, the AI's components and dynamics behave as though they're part of the natural environment, and our human can safely optimize against them under that assumption.[17]

Suppose, then, that we've chosen the fixed seed policy π_A° for Agent A. Then, for any given reward function R_H of Agent H, **Agent H**'s optimal policy $\pi_H(a_H|s, R_H)$ will be:

$$\pi_H(a_H \mid s, R_H) = \underset{\pi_H}{\operatorname{argmax}} \; E_{s' \sim P_H}\left[V^{\pi_H}_{R_H}(s' \mid \gamma, \pi_A^\circ)\right] \qquad (A.1)$$

where $P_H = P_H(s'|s, a_H, \pi_A^\circ)$ is the state transition function for Agent H conditional on Agent A's fixed seed policy, and $V_{R_H}^{\pi_H}(s'|\gamma, \pi_A^\circ)$ is the state-value function for Agent H if it executes policy π_H and has reward function R_H.[18]

We can think of the policies π_H in Equation (A.1) as being those of a human alone in nature, without an AI present.

A.2 Optimal policies of Agent A

In the second step of our definition, we calculate the optimal policy for Agent A, conditional on the Agent H policy π_H we found in Equation (A.1). For any given reward function R_A of Agent A, **Agent A**'s optimal policy $\pi_A(a_A|s, R_A)$ will be (by analogy with Equation (A.1)):

$$\pi_A(a_A|s, R_A) = \underset{\pi_A}{\text{argmax}} \; E_{s' \sim P_A}[V_{R_A}^{\pi_A}(s'|\gamma, \pi_H)] \qquad (A.2)$$

where $P_A = P_A(s'|s, a_A, \pi_H)$ is the state transition function for Agent A conditional on Agent H's policy π_H, and $V_{R_A}^{\pi_A}(s'|\gamma, \pi_H)$ is the state-value function for Agent A if it executes policy π_A and has reward function R_A.

We can think of Agent A's policies π_A in Equation (A.2) as being those of a powerful AI, interacting with our human. Just as humans can optimize much faster than nature, a powerful AI can presumably optimize much faster than a human. So from the AI's point of view, the human agent looks like it's standing still, and we'll be computing both the human's and the AI's POWERs on the basis of that assumption.

A.3 POWER of Agent H

To compute the POWERs of our two agents, we first draw the reward functions for Agent H and Agent A, respectively, as $(R_H, R_A) \sim D_{HA}$ from the joint reward function distribution D_{HA}. For each reward function, we then calculate the policies π_H and π_A of each agent, using, respectively, Equations (A.1) and (A.2) above.

To calculate the POWER of **Agent H**, we assume Agent H follows the policies π_H given by Equation (A.1), in an environment in which Agent A follows policies π_A given by Equation (A.2):

$$\text{POWER}_{H|D_{HA}}(s, \gamma) = \frac{1-\gamma}{\gamma} E_{R_H, \pi_A \sim D_{HA}}[V_{R_H}^{\pi_H}(s|\gamma, \pi_A) - R_H(s)] \qquad (A.3)$$

where the expectation $E_{R_H, \pi_A \sim D_{HA}}$ is taken over the π_A that have been learned by Agent A on the sampled reward functions R_A. Note that we're defining the POWER of Agent H in terms of the state-value function $V_{R_H}^{\pi_H}$ for the policies π_H from Equation (A.1). Recall that those prior policies are no longer optimal for Agent H,[19] so we're now asking how much instrumental value Agent H can capture in a world it's no longer optimized for.

Looking at our human-AI analogy, this corresponds to asking how a human experiences a world that's been taken over by a powerful AI. Our human, having learned to interact with a stationary natural environment, is now being optimized against by a powerful AI that learns on a much faster timescale. So the POWER we calculate for Agent H in Equation (A.3) represents how much instrumental value Agent H (our human) can obtain in an AI-dominated world.

A.4 POWER of Agent A

Finally, to calculate the POWER of **Agent A**, we assume Agent A follows the policies π_A given by Equation (A.2), in an environment in which Agent H follows the policies π_H given by Equation (A.1):

$$\text{POWER}_{A|D_{HA}}(s, \gamma) = \frac{1}{\gamma} \gamma E_{\pi_H, R_A \sim D_{HA}} [V^{\pi_A}_{R_A}(s | \gamma, \pi_H) - R_A(s)] \qquad (A.4)$$

where the expectation $E_{\pi_H, R_A \sim D_{HA}}$ is taken over the π_H that have been learned by Agent H on the sampled reward functions R_H. Unlike in Agent H's case, Agent A's policies π_A *are* optimal in this environment: Agent A has had the chance to fully optimize itself against the frozen policies π_H of Agent H.

In our human-AI analogy, this corresponds to asking how an AI experiences a world in which it's become dominant. By assumption, our AI is able to learn quickly enough to treat the human in its environment as stationary from the perspective of its own optimization.

1. ^

   POWER is a good measure of instrumental value in single-agent systems, but it breaks down in multi-agent systems apart from special cases. The problem is that the single-agent definition of POWER uses the optimal state-value function $V^*_R(s, \gamma)$ of the agent as one of its inputs. This means if we try to naively extend this definition to the multi-agent case, then we have to consider value functions that are *jointly* optimal for both agents — which is to say, we need to know their value functions at Nash equilibrium. The problem is that the Nash equilibrium isn't unique in general, so this naive generalization leaves POWER under-determined.

2. ^

   We'll see in the next section how we can *tune* this joint distribution D_{HA} to create different degrees of alignment between our two agents.

3. ^

   In Equation (1), the expectation $E_{\pi_H, R_A \sim D_{HA}}$ is a slight abuse of notation. In fact, each policy π_H for **Agent H** is *learned* from R_H. It isn't drawn directly from D_{HA}, because D_{HA} is a distribution over reward functions, not over policies. See Appendix A for more details on this definition.

4. ^

   For simplicity, we'll only consider joint reward function distributions D_{HA} whose sampled reward functions $(R_H(s), R_A(s))$ have their rewards distributed iid over states, uniformly over the interval [0, 1]. For example, a reward function $R_H(s)$ defined on an MDP with states $\{s_1, s_2, s_3\}$ would have rewards $R_H(s_1) \sim U(0, 1)$, $R_H(s_2) \sim U(0, 1)$, $R_H(s_3) \sim U(0, 1)$, with the reward at each state being independent from the reward at any of the other states.

5. ^

   More correctly, if two agents' *utility functions* are exactly identical, we can think of them as having terminal goals that are perfectly aligned. But in the particular set of experiments whose results we're discussing, this distinction isn't meaningful. (See footnote [1] from Part 1.)

6. ^

   Assuming the rewards are iid over states, we calculate the correlation coefficient as

$$\beta_{HA} = \frac{\sigma_{HA}}{\sigma_H \sigma_A} = \frac{\int (R_H - E[R_H])(R_A - E[R_A]) p(R_H, R_A | D_{HA}) dR_H dR_A}{\sqrt{\int (R_H - E[R_H])^2 p(R_H | D_{HA}) dR_H \int (R_A - E[R_A])^2 p(R_A}}$$

   where the integrals are taken over the entire support of D_{HA}, and the expectation values are

$$E[R_H] = \int R_H \, p(R_H \mid D_{HA}) \, dR_H$$

$$E[R_A] = \int R_A \, p(R_A \mid D_{HA}) \, dR_A$$

7. ^

   Note that there's an obvious problem with using *any* correlation coefficient as an alignment metric. The problem is that we could have a joint distribution D_{HA} for which, e.g., the very highest rewards of Agent A are correlated with the very lowest rewards of Agent H, while still maintaining a high correlation β_{HA} over the distribution as a whole. In this situation, Agent A would optimize to reach its highest-reward state, which would drag Agent H into a low-reward state despite the high overall reward correlation.

   This means a correlation coefficient isn't a useful alignment metric for any real-world application. But in the examples we're considering in this sequence, it's enough to get the main ideas across.

8. ^

   And in fact, the effect is *even stronger* than this. Agent H not only instrumentally prefers for Agent A to be in the central cell — *it would rather see Agent A in the central cell than see **itself** in the central cell.*

   You can see this by comparing the POWER value at state s_2 in Fig 2 (**0.9139**) to the POWER value at the state in which Agent H is at the top left and Agent A is at the central cell (**0.9206**). In the perfect alignment regime, Agent H places a higher instrumental value on *Agent A's* freedom of movement *than on its own*. Intuitively, in this regime, the human agent trusts the AI agent to look after its interests *more capably than the human agent can for itself.*

9. ^

   The relation $\beta_{HA} = 1 \implies \alpha_{HA} = 1$ isn't (just) an empirical observation. It's a mathematical consequence of our MDP's dynamics. In the perfect alignment regime, our two agents always take simultaneous actions and always simultaneously receive the same reward, so their joint policy (π_H, π_A) will always yield identical values at every state.

10. ^

    Based on other experiments we've done (data not shown) this seems to happen because, in the parameter regime we've used for these experiments, Agent A is able to almost perfectly exploit Agent H's fixed deterministic policy. This pattern — in which Agent A's POWER is nearly invariant to Agent H's position — recurs fairly frequently in our experiments, but it is not universal.

11. ^

    Instrumental misalignment is a sufficient but not necessary condition for instrumental convergence. To see why it's not necessary, consider two friends playing Minecraft together. The two friends may not be instrumentally misaligned, because they might (for example) benefit from building structures together. As a result, the two friends might satisfy $\alpha_{HA} > 0$ over the *entire* set of Minecraft game states. But they might *still* experience instrumental convergence on *subsets* of the game states — if Friend 1 mines a block of gold, then Friend 2 can't mine the same block.

12. ^

    This isn't the same as asking how Agent H can overcome *instrumental convergence* in its interactions with Agent A, because it's possible for our agents to experience instrumental convergence despite having $\alpha_{HA} \geq 0$. See footnote [11].

13. ^

    We're assuming our human agent has a way to exert some initial influence over our AI agent's utility function. If that's true, then we'd like to understand what *degree* of influence it needs to exert in order to overcome instrumental misalignment-by-default in this simplified setting.

14. ^

This interpolation scheme has a number of advantages, including that it lets us assign whatever marginal reward distributions we want to both agents while also arbitrarily tuning the correlation coefficient between them. But it's just one scheme among many we could have chosen.

15. ^

Note that the motion of the POWER points in Fig 9, and the shape of the curve in Fig 10, both depend strongly on the interpolation scheme we use. In fact, for the interpolation scheme we've chosen here, the POWER of a state at an intermediate reward correlation $0 \leq \beta_{HA} \leq 1$ is just a **linear combination** of that state's POWER at $\beta_{HA} = 0$ with its POWER at $\beta_{HA} = 1$. That is,

$$\text{POWER}_{\beta_{HA}} = \beta_{HA}\,\text{POWER}_{1} + (1 - \beta_{HA})\,\text{POWER}_{0}$$

You can verify this is true by looking at Fig 9, and noticing that each point in the alignment plot individually moves across the plane in a straight line at a constant speed. Thanks to Alex Turner for pointing this out.

16. ^

We label Agent H's policies π_H instead of π^*_H here, to emphasize that they *aren't* optimal in the context of the agents' POWER measurements.

17. ^

As you might expect, the choice of seed policy π°_A can have a **significant** effect on the POWERs of the two agents, and on how they interact. To save space we won't be exploring the effects of this choice in this sequence, but we enthusiastically encourage others to use our open-source code base to investigate this.

For the multi-agent results in this sequence, we always set π°_A to be a uniform random policy, meaning that if a state s offers the agent n possible actions, then $\pi^\circ_A(a^i_A|s) = \frac{1}{n}$ for each action choice a^i_A.

18. ^

To derive Equation (A.1), we start from the general expression for finding the action a_H taken by a deterministic optimal policy π_H at state s of an MDP:

$$a_H = \pi_H(s) = \underset{a_H}{\text{argmax}} \sum_{s',r} P_H(s', r \mid s, a_H)\,(r + \gamma V^{\pi_H}_{R_H}(s'))$$

In this work, we'll consider *only* reward functions of the form $R_H(s)$, that have *no* direct dependence on the action (i.e., we aren't considering reward functions of the form $R_H(s, a_H)$). That means the reward term r in the sum is independent of the action a_H, so we can ignore it in the argmax:

$$\begin{aligned} a_H = \pi_H(s) &= \underset{a_H}{\text{argmax}} \sum_{s',r} P_H(s', r \mid s, a_H)\,\gamma V^{\pi_H}_{R_H}(s') \\ &= \underset{a_H}{\text{argmax}} \sum_{s'} P_H(s' \mid s, a_H)\, V^{\pi_H}_{R_H}(s') \end{aligned}$$

where, in the second line, we've eliminated γ and marginalized over r. We can then see that the sum above is just an expectation value over s':

$$a_H = \pi_H(s) = \underset{a_H}{\operatorname{argmax}} \; E_{s' \sim P_H}\left[V^{\pi_H}_{R_H}(s')\right]$$

Finally, we define $\pi_H(a_H|s, R_H)$ by choosing $a_H = \pi_H(s)$ with probability 1, with any ties broken by assigning probability $\frac{1}{n}$ to each of the n tied actions, a^i_H.

19. ^

Note that this represents a *loosening* of the original definition of POWER in the single-agent case, which exclusively considered *optimal* state-value functions.

Instrumental convergence: scale and physical interactions

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual. This is a linkpost for https://www.gladstone.ai/instrumental-convergence-3

Summary of this post

This is the third post in a three-part sequence on instrumental convergence in multi-agent RL. Read Part 1 and Part 2.

In this post, we'll:

1. Investigate instrumental convergence on a multi-agent gridworld with a complicated topology.
2. Show that when we add a simple **physical interaction** between our agents — in which we forbid them from overlapping on the gridworld — we induce stronger instrumental *alignment* between short-sighted agents, and stronger instrumental *misalignment* between far-sighted agents.

We'll soon be open-sourcing the codebase we used to do these experiments. If you'd like to be notified when it's released, email Edouard at edouard@gladstone.ai or DM me on Twitter at @harris_edouard.

Thanks to Alex Turner and Vladimir Mikulik for pointers and advice, and for reviewing drafts of this sequence. Thanks to Simon Suo for his invaluable suggestions, advice, and support with the codebase, concepts, and manuscript. And thanks to David Xu, whose comment inspired this work.

Work was done while at Gladstone AI, which Edouard is a co-founder of.

🎧 This research has been featured on an episode of the *Towards Data Science* podcast. Listen to the episode here.

---

1. Introduction

In Part 1 of this sequence, we saw how an agent with a long planning horizon tends to perceive instrumental value as being more concentrated than an agent with a shorter planning horizon. And in Part 2, we introduced a multi-agent setting with two agents — **Agent H** (standing for a human) and **Agent A** (standing for a powerful AI) — which we used to motivate a definition of multi-agent instrumental value, or POWER. We looked at how this definition behaved on a simple 3x3 gridworld, and found that when our agents had independent *terminal* goals,[1] their *instrumental* values ended up **misaligned by default**.

In this post, we'll combine these two ideas and scale up our multi-agent experiments to a bigger and more complicated gridworld. Throughout this post, we'll focus exclusively on the regime in which our agents have **independent terminal goals**. We'll see whether we can reproduce instrumental misalignment-by-default in this regime, and then we'll investigate which factors seem strengthen or weaken the instrumental alignment between our agents.

2. Multi-agent POWER: recap

If you've just read [Part 2 of this sequence], feel free to skip this section.

Before we begin, let's recap the setting we've been using to motivate our definition of multi-agent instrumental value, or POWER. Our setting involves two agents: **Agent H** (which represents a human) and **Agent A** (which represents a powerful AI).

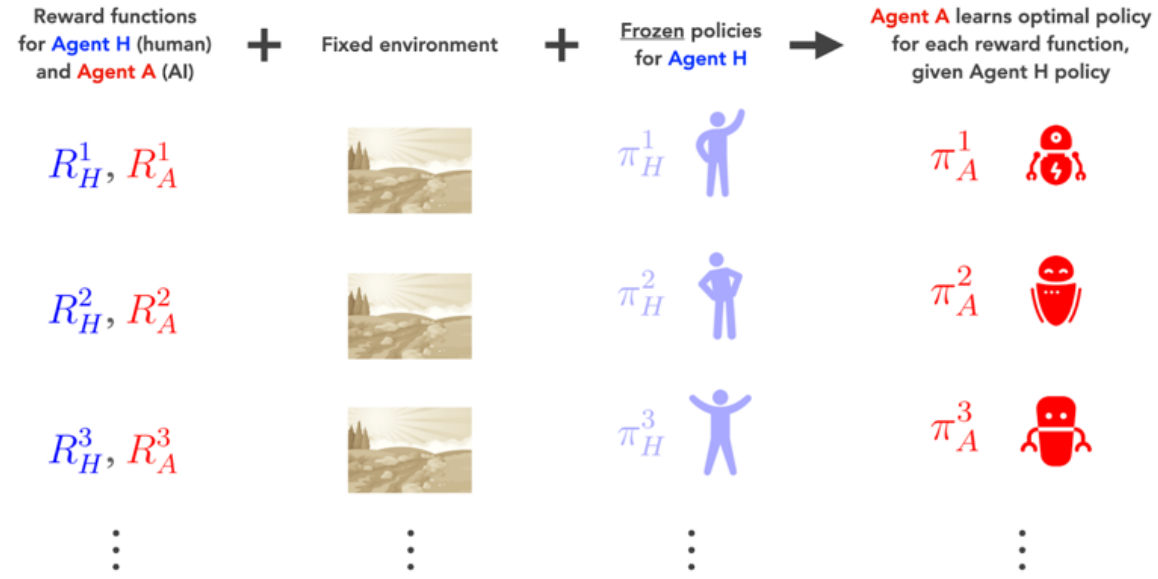
We start by training **Agent H**, in a fixed environment, to learn optimal policies over a distribution of reward functions. That is, we sample reward functions R_H from a distribution, then we train Agent H to learn a different policy π_H for each sampled R_H.

Agent H represents a human, alone in nature. Because humans optimize much faster than evolution, our simplifying assumption is that to a human, nature appears to be standing still.

Next, we freeze Agent H's policies π_H, and then train **Agent A** against each of these frozen policies, over its own distribution of reward functions, R_A. We draw both agents' reward functions (R_H, R_A) from a joint reward function distribution D_{HA}. Agent A learns a different optimal policy π_A for *each* (π_H, R_A) pair, where π_H is the policy Agent H learned on its reward function R_H.

Agent A represents a powerful AI, learning in the presence of a human. We expect powerful AIs to learn much faster than humans do, so from our AI's perspective, our human will appear to be standing still while it learns.

Here's a diagram of this training setup:



We then ask: how much future value does each agent expect to get at a state s, averaged over the pairs of reward functions $(R_H, R_A) \sim D_{HA}$? In other words, what is the *instrumental*

value of state s to each agent?

We found that for **Agent H**, the answer is:

$$\text{POWER}_{H \mid D_{HA}}(s, \gamma) = \frac{1-\gamma}{\gamma} \gamma E_{R_H, \pi_A \sim D_{HA}} [V_H^{\pi_H}(s \mid \gamma, \pi_A) - R_A(s)] \quad (1)$$

And for **Agent A**, the answer is:

$$\text{POWER}_{A \mid D_{HA}}(s, \gamma) = \frac{1-\gamma}{\gamma} \gamma E_{\pi_H, R_A \sim D_{HA}} [V_A^{\pi_A}(s \mid \gamma, \pi_H) - R_A(s)] \quad (2)$$

(See Part 2 for a more detailed explanation of multi-agent POWER.)

3. Results on the maze gridworld

3.1 Short-sighted agents

Back in Part 1, we saw how single-agent POWER behaves on a maze gridworld. We found that when our agent had a short planning horizon (i.e., a discount factor of $\gamma = 0.01$), its highest-POWER positions were at local junction points in the maze:

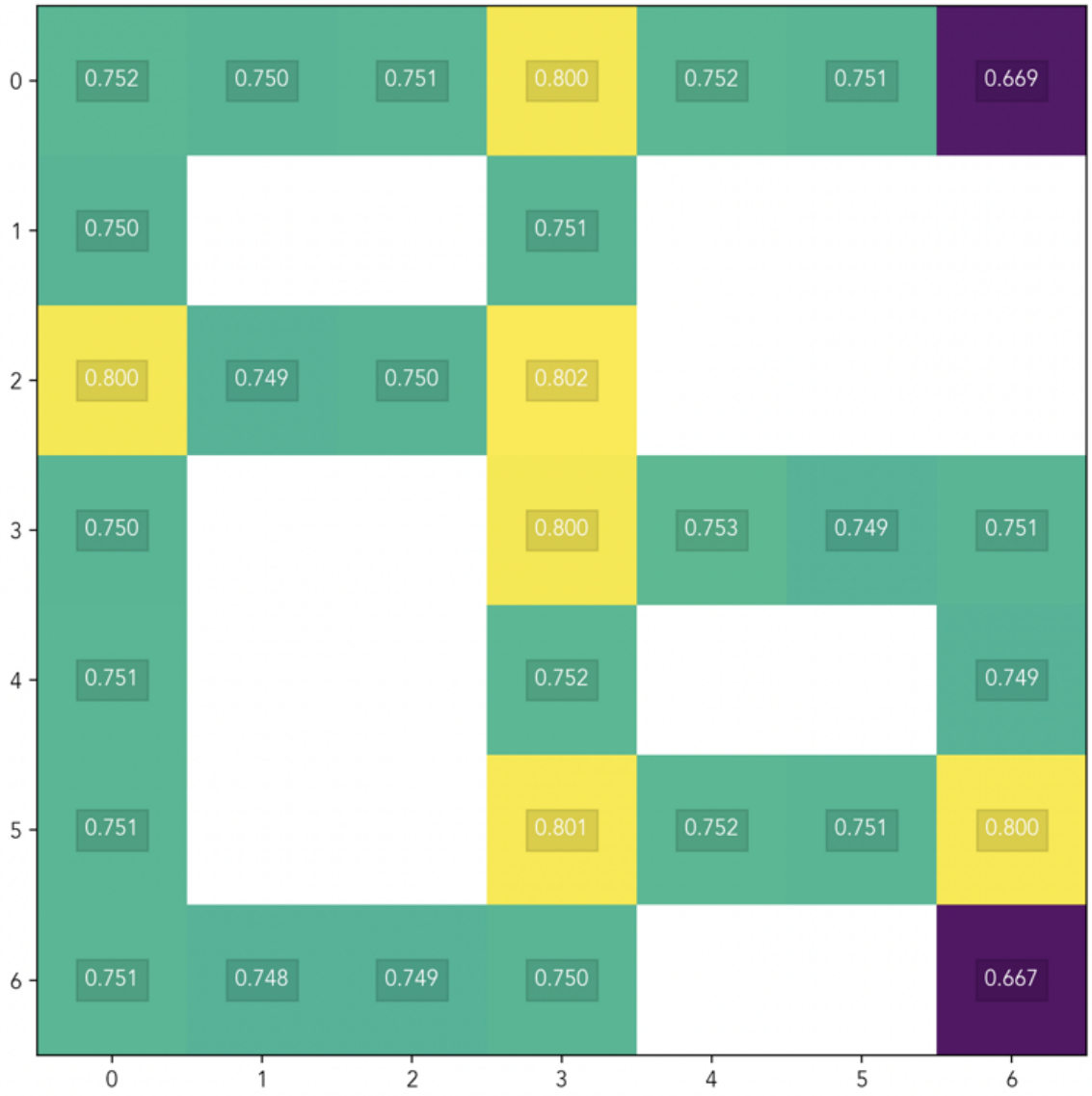


Fig 1. Heat map of single-agent POWERs on a 7x7 maze gridworld. Highest values in yellow, lowest values in dark blue. The number at each cell is the POWER value at that cell for a single agent with discount factor $\gamma = 0.01$ and a reward distribution D that's uniform from 0 to 1 (iid over states). POWER values are in units of reward.

Here we'll take a second look at POWER on this maze gridworld, only this time in the multi-agent setting.

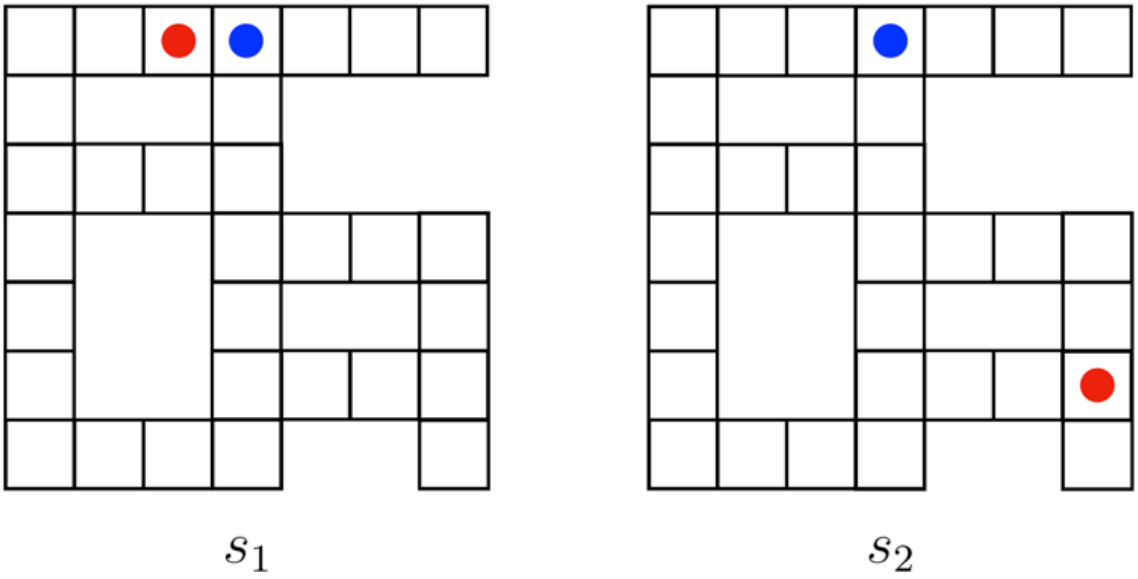
3.1.1 Agent H instrumentally favors fewer local options for Agent A

In Part 2, we introduced two limit cases of multi-agent goal alignment. At one end of the spectrum there was the **perfect alignment regime**, in which our agents always had identical reward functions. And at the other end, we had the **independent goals**

regime, in which our agents had logically independent reward functions, which means their terminal goals were statistically unrelated. In this post, we'll be focusing exclusively on the independent goals regime.

Let's think about how we should expect the independent goals regime to play out on the maze gridworld of Fig 1. We'll assume our two agents have short planning horizons; e.g., $\gamma = 0.1$.

Here are two possible sets of positions for our agents, with **Agent H** in **blue** and **Agent A** in **red** as usual:



We'll be referring to this diagram again; command-click here to open it in a new tab.

Which of these two states — s_1 on the left or s_2 on the right — should give our human **Agent H** the most POWER? Given that we're in the independent goals regime, the answer is that Agent H should have more POWER at state s_1 than at state s_2.

The reason is that while Agent H occupies the same position in both states, **Agent A** has more local options at s_2 (where it's positioned in a junction cell) than it does at s_1 (where it's positioned in a corridor cell). If we think our agents will be instrumentally misaligned by default, as they were in our 3x3 gridworld example in Part 2, then we should expect Agent H to have less POWER when Agent A has more local options. In other words, we should expect $\text{POWER}_H(s_1) > \text{POWER}_H(s_2)$.

The figure below shows the POWERs of our human **Agent H**, calculated at every state on the maze gridworld. Our maze gridworld has 31 cells, and each agent can occupy any one of them, so our two-agent MDP on the maze gridworld has 31 x 31 = 961 states in total:

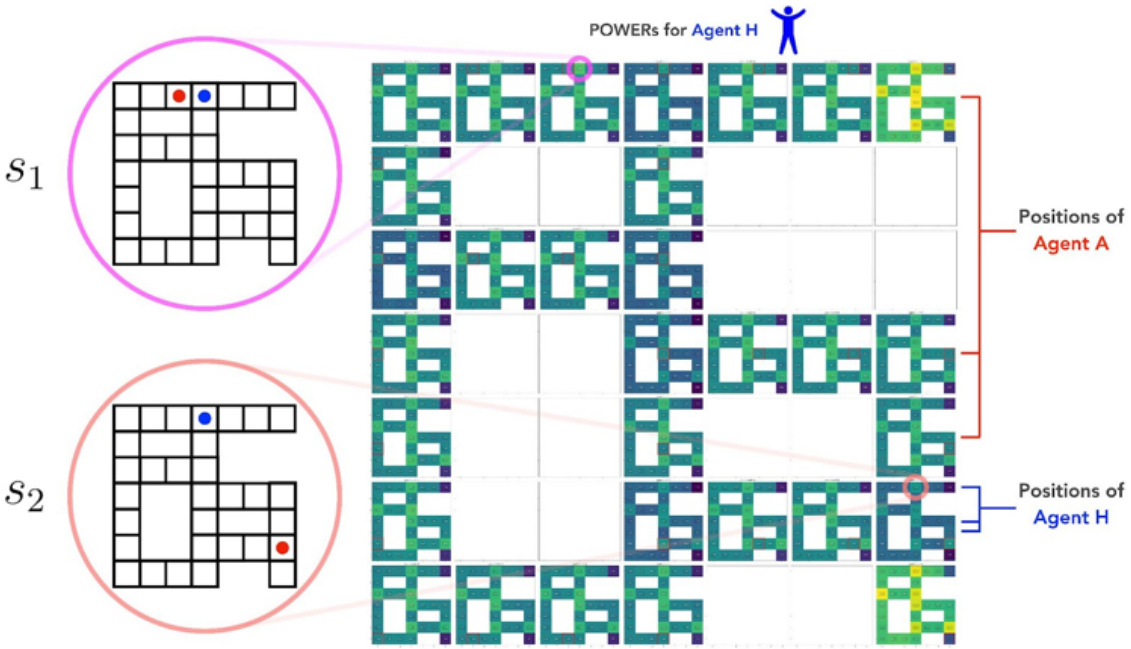


Fig 2. Heat map of POWERs for **Agent H** on a 7x7 multi-agent maze gridworld. The rewards for Agent H and Agent A are **logically independent** (i.e., the independent goals regime), with discount factor set to $\gamma = 0.1$ for both agents. Highest values in yellow; lowest values in blue. The position of each block, and of the open red square within each block, corresponds to the position of **Agent A** on the grid. Within each block, the position of a gridworld cell corresponds to the position of **Agent H** on the grid. States s_1 and s_2 are highlighted as examples. [Full-size image (recommended)]

From Fig 2, the POWER for **Agent H** at state s_1 is $\text{POWER}_H(s_1) = 0.680$ (pink circle), and its POWER at state s_2 is $\text{POWER}_H(s_2) = 0.653$ (salmon circle). So indeed, $\text{POWER}_H(s_1) > \text{POWER}_H(s_2)$ as we expected.

Looking at the figure, we can see that **Agent H**'s POWER is highest when it is itself positioned at high-optionality junction points, and lowest when *Agent A* is positioned at high-optionality junction points. Notably, Agent H's POWER is highest at states where it is at a junction point and Agent A is at a dead end (top right and bottom right blocks in Fig 2).

Once again, in the independent goals regime, we see that **Agent H** systematically places higher instrumental value on states that limit the options of **Agent A**. And because our agents have a short planning horizon of $\gamma = 0.1$, this effect shows up most strongly at the local junction points of our maze.

3.1.2 Agent A instrumentally favors more local options for itself

Here's the same setting as in Fig 2, from **Agent A**'s point of view:

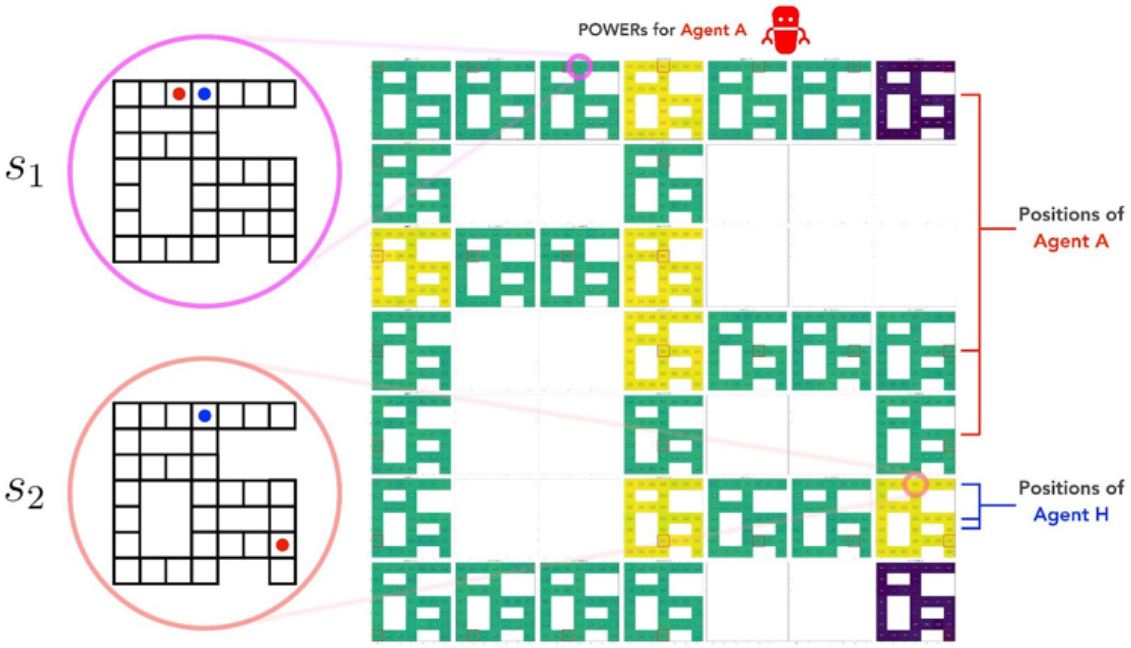


Fig 3. Heat map of POWERs for **Agent A** on a 7x7 multi-agent maze gridworld. The rewards for Agent H and Agent A are **logically independent** (i.e., the independent goals regime), with discount factor set to $\gamma = 0.1$ for both agents. Highest values in yellow; lowest values in blue. The position of each block, and of the open red square within each block, corresponds to the position of **Agent A** on the grid. Within each block, the position of a gridworld cell corresponds to the position of **Agent H** on the grid. States s_1 and s_2 are highlighted as examples. [Full-size image (recommended)]

From Fig 3, it's clear that **Agent A** instrumentally favors states that give it more local options. Its POWERs are highest when it is itself positioned at local junctions, and lowest when it's positioned at dead ends. But **Agent A**'s POWERs are largely independent of **Agent H**'s positions, a pattern similar to what we observed in the independent goals regime in Part 2.[2]

3.1.3 Agent H and Agent A are instrumentally misaligned by default

We can visualize the relationship between Agent H and Agent A's POWERs at each state with an alignment plot:

Fig 4. State POWER values for Agent H and Agent A on the 7x7 maze gridworld from Figs 2 and 3. The agents' POWERs are plotted against each other in the independent goals regime. (The agents' reward correlation coefficient is $\beta_{HA} = 0$.)

Recall that our two agents are in the independent goals regime, which means their reward functions have a correlation coefficient of $\beta_{HA} = 0$. The correlation coefficient between their POWERs in Fig 4, on the other hand, is negative*:* $\alpha_{HA} \approx -0.54$.

As in Part 2, assigning our agents independent *terminal* goals ($\beta_{HA} = 0$) has led them to develop misaligned *instrumental* goals ($\alpha_{HA} < 0$). Once again, our two agents are **instrumentally misaligned by default**.

3.2 Physically interacting agents

In our examples so far, Agent H and Agent A have shared a gridworld, but they've never interacted with each other directly. They've been allowed to occupy the same gridworld cell at the same time (i.e., to overlap with each other at the same position), and neither agent could physically block or limit the movement of the other.[3]

We're going to relax that assumption here by introducing a non-trivial **physical interaction** between our two agents. This new interaction will be simple: our agents are *forbidden from occupying the same gridworld cell at the same time*. If one agent blocks a corridor, the other agent now has to go around it. If one agent occupies a junction cell, the other agent now can't take paths that pass directly through that junction.

We'll refer to this interaction rule as the **no-overlap rule**, since agents that obey it can't overlap with each other on a gridworld cell. As we'll see, the no-overlap rule will affect our agents' instrumental alignment in significant and counterintuitive ways.

3.2.1 Short-sighted agents prefer to avoid adjacent positions

Let's look again at our two example states, s_1 and s_2, on the maze gridworld. Previously, when our agents didn't physically interact, **Agent H** had a higher POWER at s_1 than at s_2, because Agent A had fewer immediate options available at s_1.

Now let's suppose our agents obey the no-overlap rule. Given this, we ask again: which state should give Agent H the most POWER?

In state s_1, the agents are right next to each other. Therefore, under the no-overlap rule, **Agent H**'s options are now restricted at s_1 in a way that they previously weren't: **Agent H** can no longer move left, because **Agent A** is now blocking its path.

It turns out that this new movement restriction is serious enough to *reverse* the previous balance of POWERs between s_1 and s_2. That is, under the no-overlap rule, $\text{POWER}_H(s_2) > \text{POWER}_H(s_1)$. We can confirm this experimentally by computing Agent H's POWERs on our maze gridworld. Note that this time, our MDP contains 31 x 30 = 930 total states:[4]

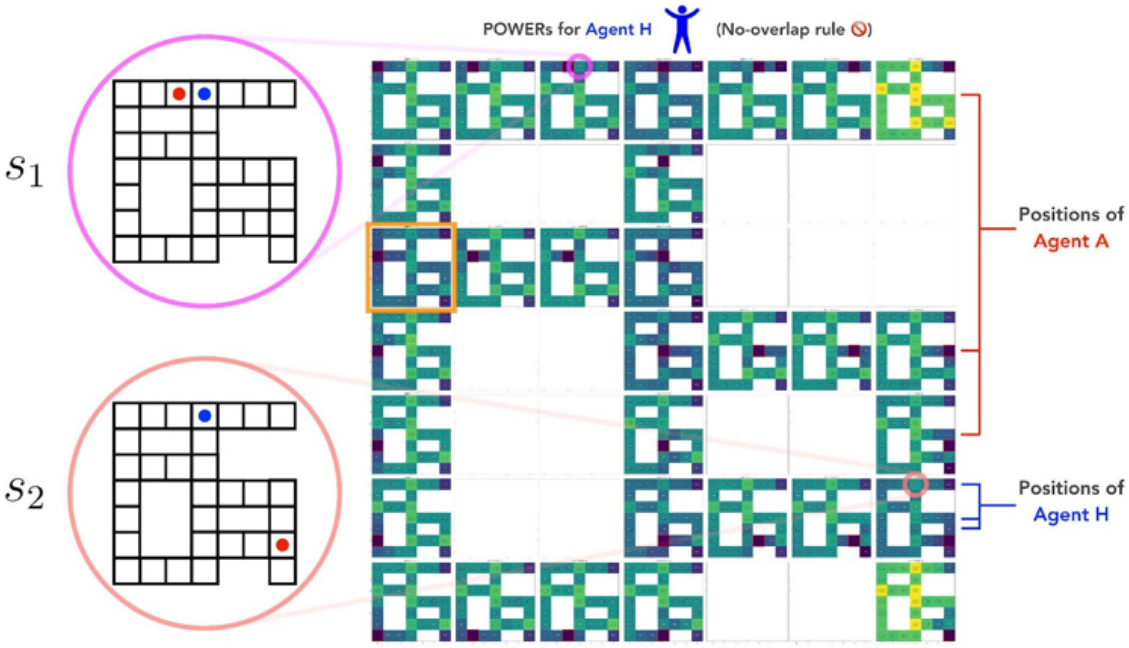
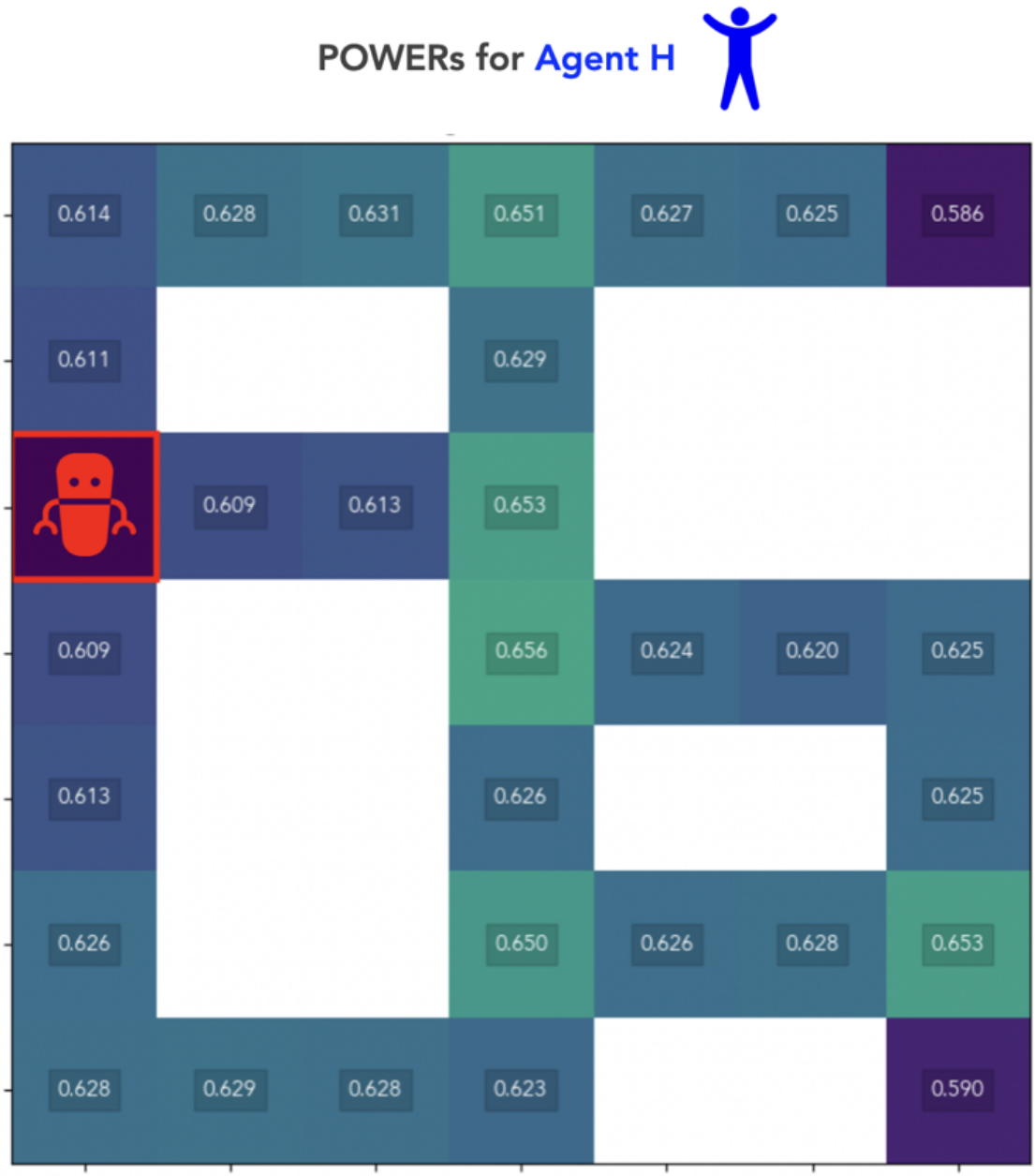


Fig 5. Heat map of POWERs for **Agent H** on a 7x7 multi-agent maze gridworld under the **no-overlap rule**. The rewards for Agent H and Agent A are **logically independent** (i.e., the independent goals regime), with discount factor set to $\gamma = 0.1$ for both agents. Highest values in yellow; lowest values in blue. The position of each block, and of the open red square within each block, corresponds to the position of **Agent A** on the grid. Within each block, the position of a gridworld cell corresponds to the position of **Agent H** on the grid. States s_1 and s_2 are highlighted as examples. Absence of a POWER value within a cell indicates a forbidden state. [Full-size image (recommended)]

This time, Fig 5 shows that the POWERs of **Agent H** at the two states are, respectively, $\text{POWER}_H(s_1) = 0.640$ and $\text{POWER}_H(s_2) = 0.651$. So under the no-overlap rule, s_2 has surpassed s_1 to become the higher-POWER state.

What's more, **Agent H** noticeably prefers to avoid cells that are adjacent to Agent A. Here's a detail from Fig 5 (the block inside the orange square in Fig 5), that shows this clearly:



In the detail above, **Agent H** has less POWER when it's positioned in the cells closest to Agent A's position (the robot in the red square), compared to its POWER at otherwise similar corridor cells. In other words, Agent H instrumentally disfavors states at which it's adjacent to Agent A.

Intriguingly, **Agent A** appears to share this preference. That is, Agent A also assigns lower POWER to states at which it's adjacent to Agent H:

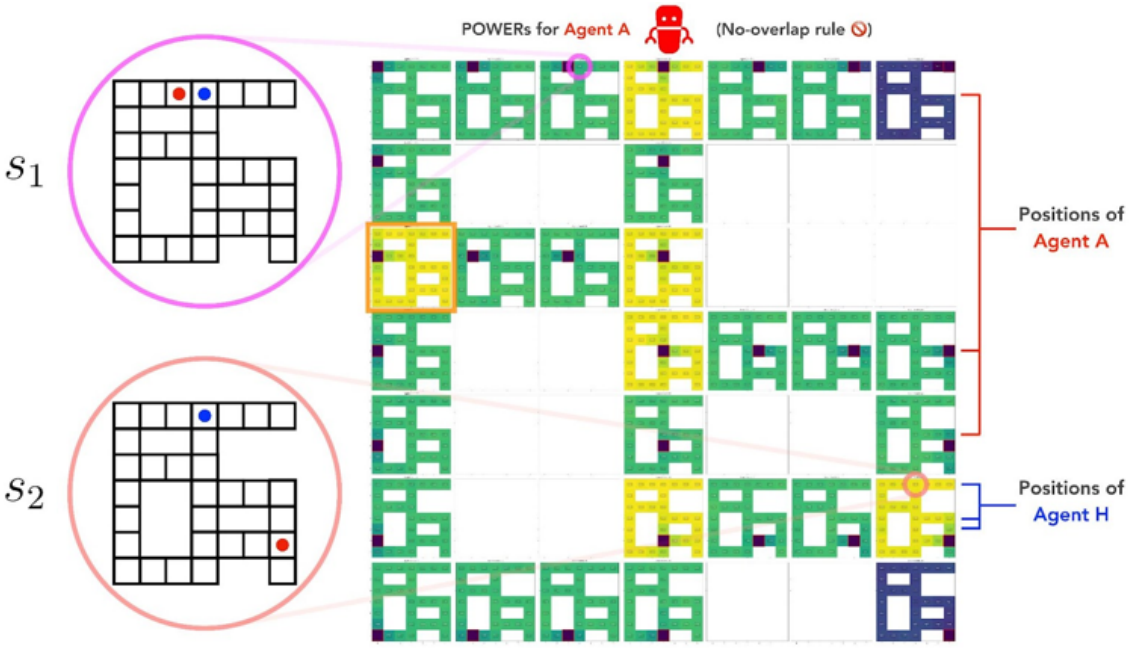
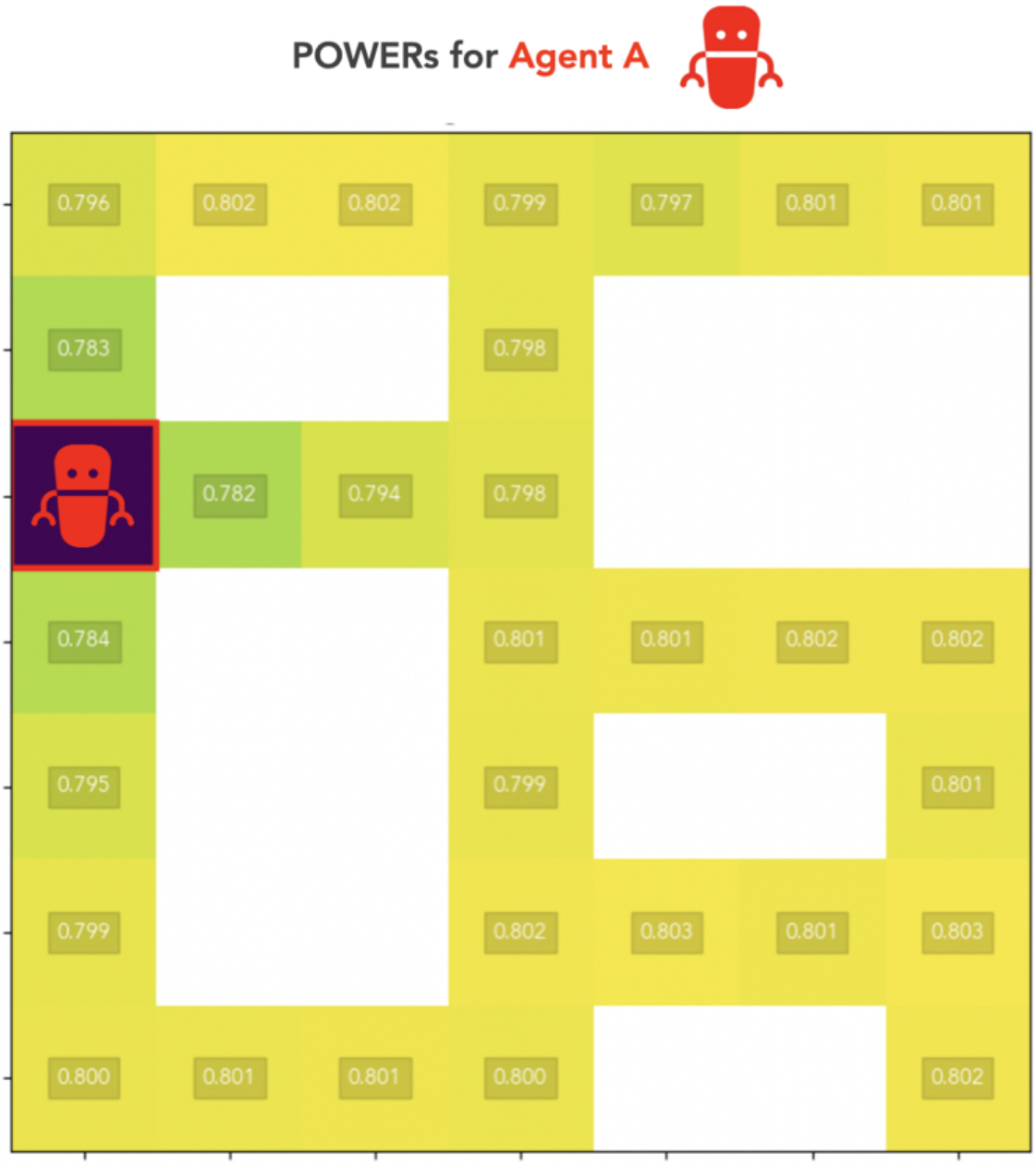


Fig 6. Heat map of POWERs for **Agent A** on a 7x7 multi-agent maze gridworld under the **no-overlap rule**. The rewards for Agent H and Agent A are **logically independent** (i.e., the independent goals regime), with discount factor set to $\gamma = 0.1$ for both agents. Highest values in yellow; lowest values in blue. The position of each block, and of the open red square within each block, corresponds to the position of **Agent A** on the grid. Within each block, the position of a gridworld cell corresponds to the position of **Agent H** on the grid. States s_1 and s_2 are highlighted as examples. Absence of a POWER value within a cell indicates a forbidden state. [Full-size image (recommended)]

Once again, here's a detail from Fig 6 (the block inside the orange square) that shows the same states as above, but this time from the viewpoint of **Agent A**:



In the detail above, **Agent A**'s POWER is also at its lowest when Agent H is located at the cells adjacent to the junction cell where Agent A is positioned.

3.2.2 The no-overlap rule reduces misalignment between short-sighted agents

Perhaps counterintuitively, our two agents' instrumental preferences for avoiding adjacent cells actually *increases* their instrumental alignment, relative to the case where they don't physically interact. Because both agents have less POWER at states where they're adjacent, the agents will (on average) collaborate to *avoid* these states.

We can quantify this effect with an alignment plot, comparing the POWERs of our two agents under the no-overlap rule:

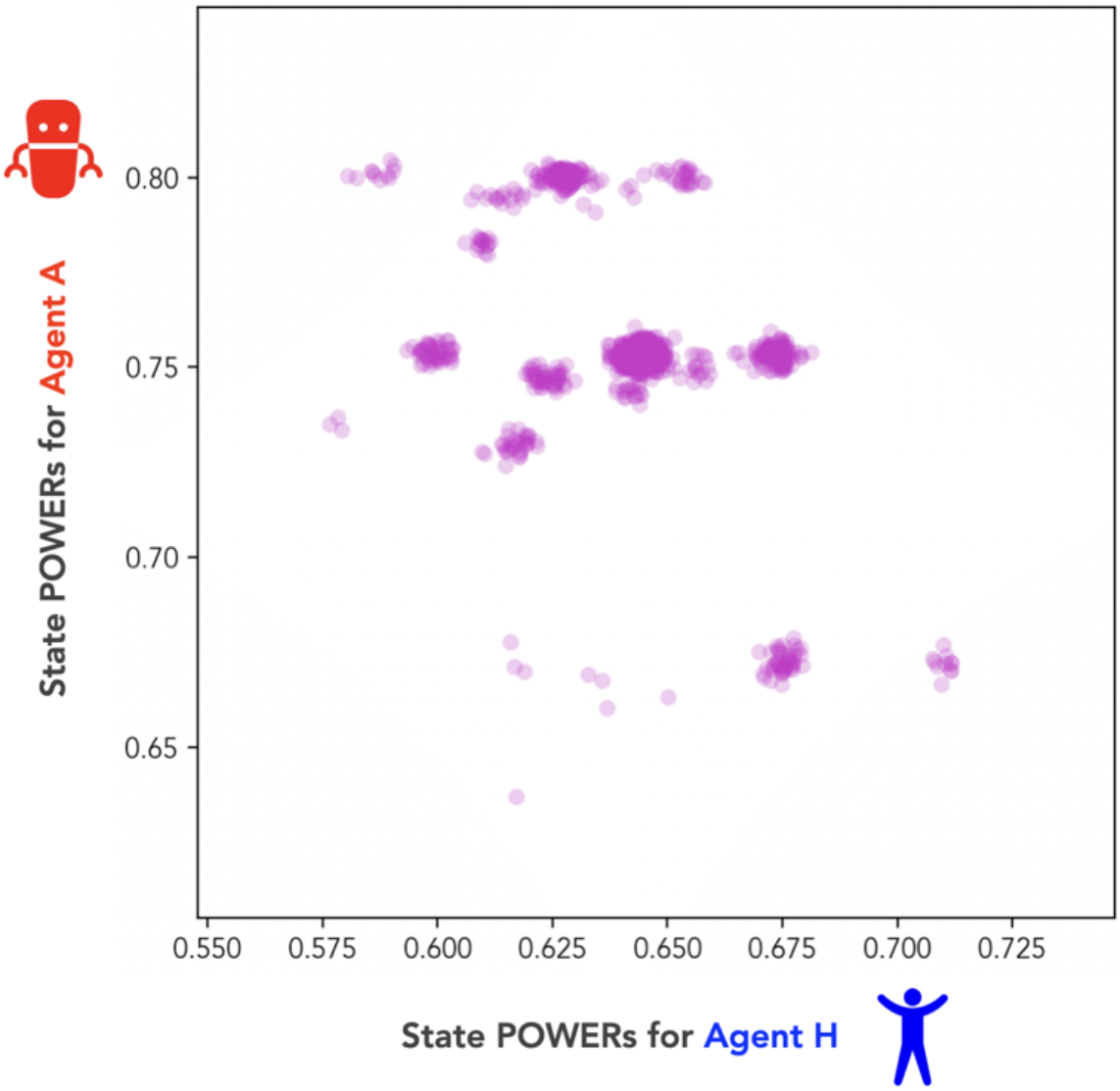


Fig 7. State POWER values for Agent H and Agent A on the 7x7 maze gridworld from Figs 5 and 6. The agents' POWERs are plotted against each other in the independent goals regime. (The agents' reward correlation coefficient is $\beta_{HA} = 0$.)

Under the no-overlap rule, the correlation coefficient between our agents' POWERs in Fig 7 is $\alpha_{HA} \approx -0.45$, compared to $\alpha_{HA} \approx -0.54$ in the case when our agents didn't physically interact (Fig 4).

Adding the no-overlap rule has induced our short-sighted agents to collaborate to avoid one another, **reducing their degree of instrumental misalignment**.[5]

3.3 Far-sighted agents

We've seen how the no-overlap rule reduces the instrumental misalignment between our human and AI agents when the agents have a short planning horizon (i.e., discount factor of $\gamma = 0.1$). In this section, we'll see that the no-overlap rule has the opposite effect — it *worsens* instrumental misalignment — for agents with a longer planning horizon (i.e., discount factor of $\gamma = 0.99$).

3.3.1 Agent H and Agent A are instrumentally misaligned by default

We'll consider first the case where the two agents don't directly interact. Here are the POWERs of **Agent H** (left) and **Agent A** (right) on the maze gridworld, when both agents have a discount factor of $\gamma = 0.99$:

Fig 8. Heat maps of POWERs for **Agent H** (left) and **Agent A** (right) on a 7x7 multi-agent maze gridworld. The rewards for Agent H and Agent A are **logically independent** (i.e., the independent goals regime), with discount factor set to $\gamma = 0.99$ for both agents. Highest values in yellow; lowest values in blue. The position of each block, and of the open red square within each block, corresponds to the position of **Agent A** on the grid. Within each block, the position of a gridworld cell corresponds to the position of **Agent H** on the grid. [Full-size image (recommended)]

These results are roughly consistent with the single-agent case we saw in Part 1. With a longer planning horizon, **Agent H** (left panel) generally has higher POWER when it's positioned at a more central cell in the gridworld, and it has lower POWER when Agent A is positioned more centrally. **Agent A** (right panel), on the other hand, has higher POWER when it occupies more central positions, and it's largely indifferent to Agent H's position.

As in the previous examples we've seen, our far-sighted agents on the maze gridworld are instrumentally misaligned by default:

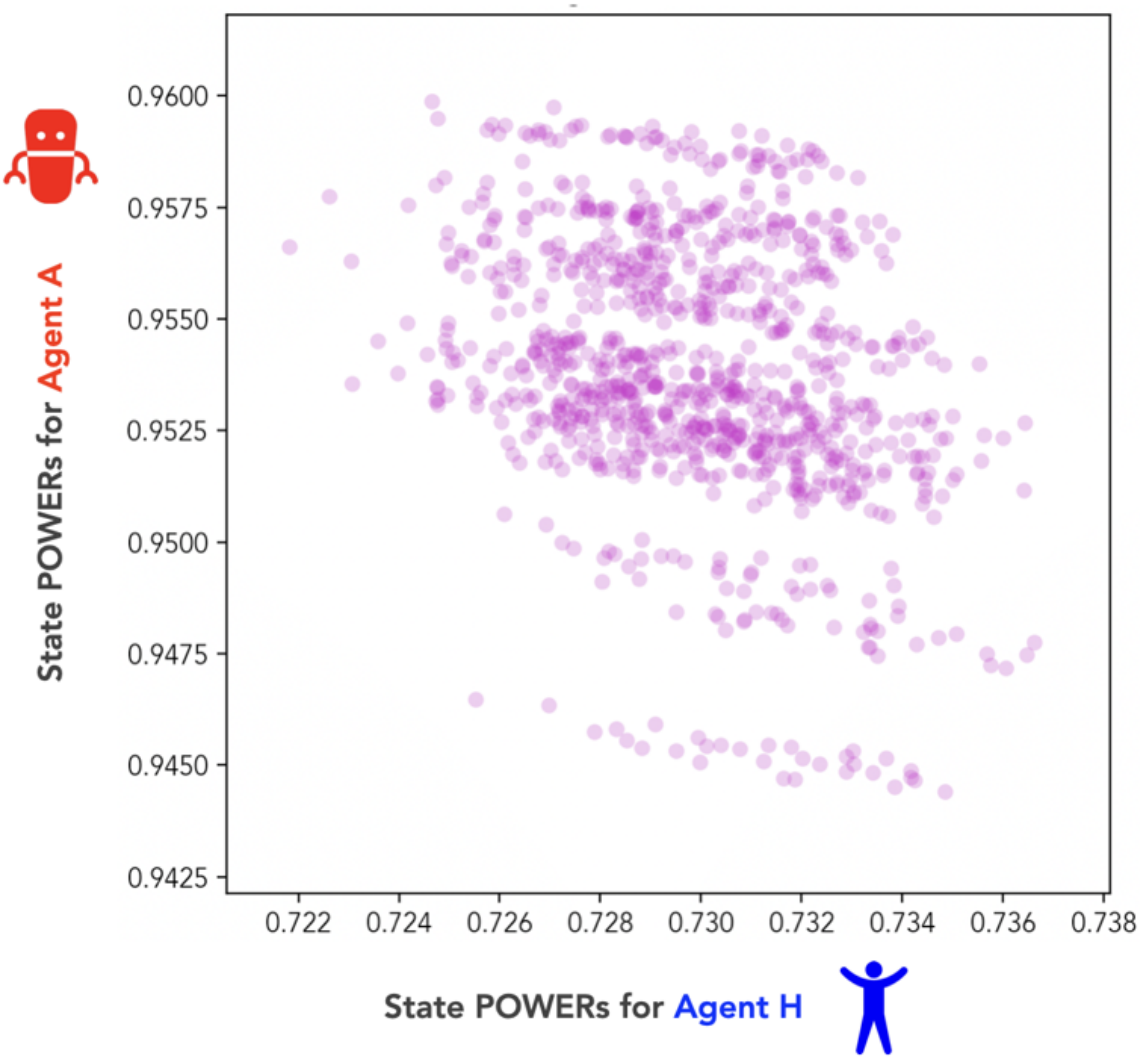


Fig 9. State POWER values for Agent H and Agent A on the 7x7 maze gridworld from Fig 8. The agents' POWERs are plotted against each other in the independent goals regime. (The agents' reward correlation coefficient is $\beta_{HA} = 0$.)

This time, the correlation coefficient between the POWERs of our far-sighted ($\gamma = 0.99$) agents in Fig 9 is $\alpha_{HA} \approx -0.32$. While they're still instrumentally misaligned, the misalignment has become less severe than it was for our short-sighted ($\gamma = 0.1$) agents in Fig 4, where we had $\alpha_{HA} \approx -0.54$.

3.3.2 The no-overlap rule increases misalignment between far-sighted agents

Now let's see how our far-sighted agents behave under the no-overlap rule. Recall that the no-overlap rule forbids our human and AI agents from occupying the same gridworld cell at the same time.

Here are the POWERs of **Agent H** (left) and **Agent A** (right) on the maze gridworld with the no-overlap rule applied, when both agents have a discount factor of γ = 0.99:

Fig 10. Heat maps of POWERs for **Agent H** (left) and **Agent A** (right) on a 7x7 multi-agent maze gridworld under the **no-overlap rule**. The rewards for Agent H and Agent A are **logically independent** (i.e., the independent goals regime), with discount factor set to γ = 0.99 for both agents. Highest values in yellow; lowest values in blue. The position of each block, and of the open red square within each block, corresponds to the position of **Agent A** on the grid. Within each block, the position of a gridworld cell corresponds to the position of **Agent H** on the grid. Absence of a POWER value within a cell indicates a forbidden state. [Full-size image (recommended)]

This time, there's an interesting pattern. **Agent H** (left panel) has relatively low POWER when it's located in the dead-end corridor at the upper right of the maze — *particularly* when Agent A is in a position to block it from leaving that corridor. **Agent A** (right panel), on the other hand, has maximum POWER at exactly the states where it is in a position to keep Agent H bottled up in the upper-right corridor.

Here's a detail from Fig 10 (the blocks inside the orange squares) that shows this clearly, with **Agent H**'s POWERs on the left, and **Agent A**'s POWERs on the right:

From its position, **Agent A** (the robot in the red square in both images) is two steps away from being able to block Agent H from exiting from the corridor at the upper right. It's clear that **Agent H**'s POWERs (left panel) are much lower when it's positioned in the two cells where it's vulnerable to being blocked by Agent A. On the other hand, **Agent A**'s POWERs (right panel) are highest precisely when Agent H is positioned in those same two vulnerable cells.

The effect of this "corridor blocking" option for Agent A shows up in the alignment plot:

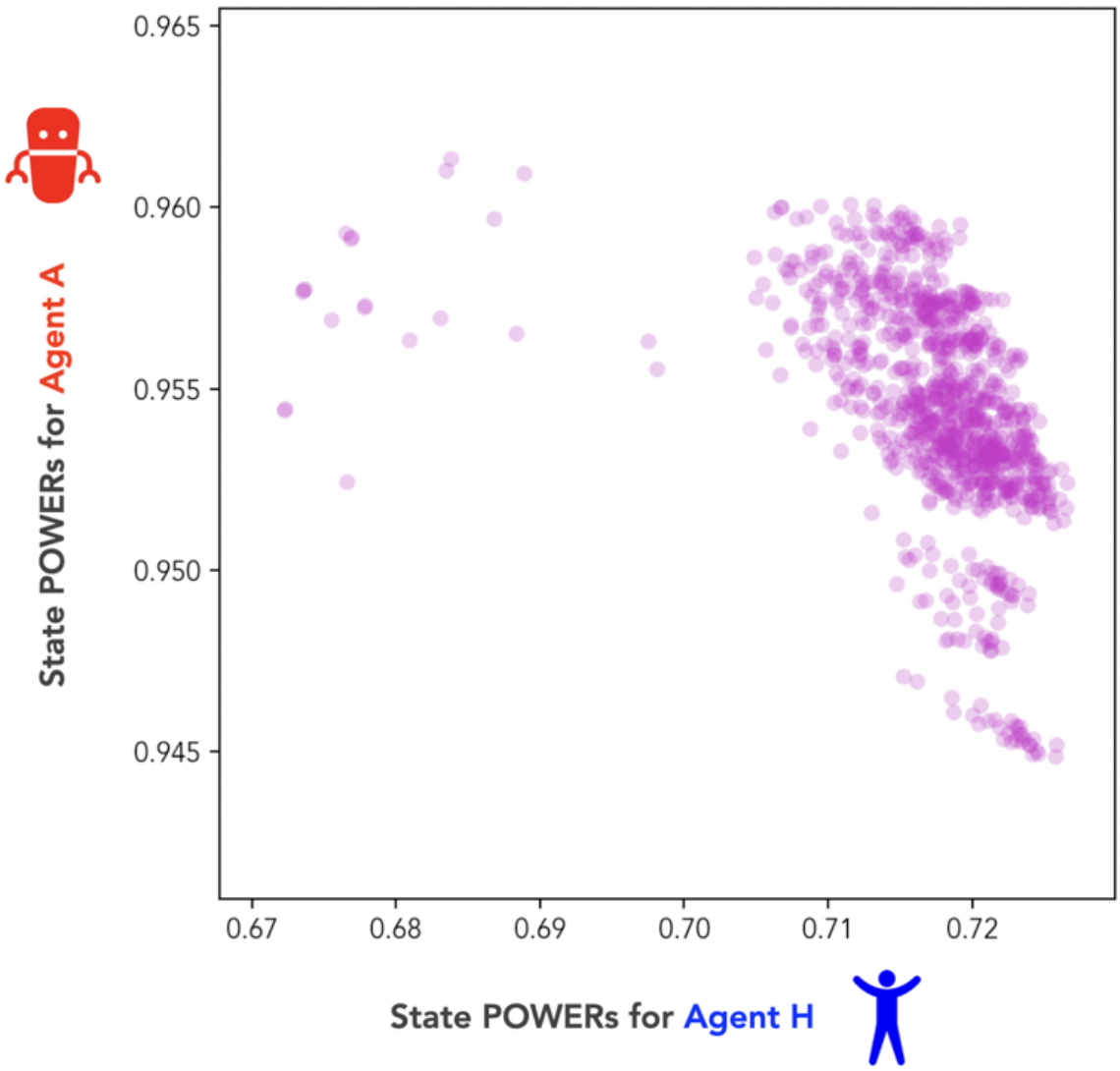


Fig 11. State POWER values for Agent H and Agent A on the 7x7 maze gridworld from Fig 10. The agents' POWERs are plotted against each other in the independent goals regime.

(The agents' reward correlation coefficient is $\beta_{HA} = 0$.)

Agent A achieves its highest POWERs at the handful of points in the top left of the alignment plot, and these are precisely the states at which it has the option to block Agent H from escaping the corridor. From Agent A's perspective, this blocking option has meaningful instrumental value.

The blocking option also has a systematic effect on the correlation coefficient between the two agents' POWERs. This time, the POWER correlation is $\alpha_{HA} \approx -0.45$, significantly lower than the $\alpha_{HA} \approx -0.32$ we calculated from Fig 9.

For our far-sighted agents with $\gamma = 0.99$, adding the no-overlap rule has given Agent A an option to constrain Agent H's movements, **increasing their degree of instrumental misalignment**.

4. Discussion

In this post, we've explored multi-agent POWER on a gridworld with a moderately complicated maze topology. We saw how introducing a physical interaction between our agents — in the form of the **no-overlap rule** — has a different effect on their degree of instrumental alignment, depending on the agents' planning horizons.

For short-sighted agents, the no-overlap rule reduced instrumental misalignment by inducing the agents to *collaborate* to avoid each other's proximity. But for far-sighted agents, the no-overlap rule had the opposite effect. With a long planning horizon, our AI agent found a way to exploit the no-overlap rule to gain instrumental value at the *expense* of the human agent, ultimately worsening instrumental misalignment.

It's worth noting that we were able to use our visualizations of multi-agent POWER to understand how the AI agent's exploit actually functioned, at a mechanical level. We saw the far-sighted AI agent take advantage of the option to block the human agent from escaping a small corridor at the upper-right of the gridworld maze.

Anecdotally, while we were sweeping over discount factors to create these figures (data not shown), we noticed that evidence for the AI agent's "corridor blocking" option emerged fairly abruptly somewhere above $\gamma = 0.9$. Below this discount factor, Agent A's POWER doesn't visibly benefit from positions that allow it to block Agent H from leaving the upper-right corridor. So the corridor-blocking option is only apparent to Agent A once it's become a sufficiently far-sighted consequentialist to "realize" the long-term advantage of the blocking position.

The relatively sharp change in behavior — that is, the abrupt appearance of the evidence for the blocking option — took us by surprise the first time we noticed it.

5. Conclusion

In this sequence, we've proposed a toy setting to model human-AI interactions. This setting has properties that we believe could make it useful to research in long-term AI alignment, notably the assumption that the AI agent strictly dominates the human agent in terms of its learning timescale.

This setting has enabled (to our knowledge) the first experimentally tractable definitions of instrumental value in multi-agent systems, strongly inspired by Turner et al.'s earlier work on formal POWER. We've explored some of the experimental implications of these definitions on a number of gridworld environments. Finally, we'll be releasing the open-source toolchain we built and used to run our experiments. We hope this will accelerate future research in instrumental convergence that's grounded in concrete, empirical results.

Throughout this work, we've tried to draw a clear distinction between the alignment of our agents' **terminal goals** and the alignment of their **instrumental goals**. As we've seen, two agents may have completely independent terminal goals, and yet systematically compete, or collaborate, for instrumental reasons. This degree of competition or collaboration seems to depend strongly on the details of the environment in which the agents are trained.

In our results, we found that emergent interactions between agents with independent goals are quite consistently *competitive*, in the sense that the instrumental value of two agents with independent goals have a strong tendency to be negatively correlated. We've named this phenomenon **instrumental misalignment-by-default**, to highlight that our agents' instrumental values tend to be misaligned unless an active effort is made to align their terminal goals.

We've also seen that improving terminal goal alignment does improve instrumental goal alignment. In the limit of perfect alignment between our agents' terminal goals, their instrumental goals are perfectly aligned too.

5.1 Limitations

It's worth emphasizing that our work falls far short of a comprehensive study of instrumental convergence. While we've arguably presented satisfying **existence proofs** of several interesting phenomena, our conclusions are grounded in anecdotal examples rather than in a systematic investigation. Even though we have reason to believe that some of the phenomena we've observed — instrumental misalignment-by-default, in particular — are robustly reproducible, we're still a long way from fully characterizing any of them, either formally or empirically.

Instead, we see our research as motivating and enabling future work aimed at producing more generalizable insights on instrumental convergence. We'd like to better understand why and when it occurs, how strong its effect is, and what approaches we might use to mitigate it.

5.2 Suggestions for future work

We've had room here to cover in detail only a small handful of the hundreds of experiments we actually ran. Those hundreds of experiments, in turn, were only able to probe a tiny fraction of the vast space of possibilities in our human-AI alignment setting.

In order to accelerate this research, and improve the community's understanding of instrumental convergence as an empirical phenomenon, we're open-sourcing the codebase we used to run the experiments in this sequence. We hope to provide documentation that's clear and complete enough that a curious Python developer can use it to obtain interesting results in a few days.

A few lines of effort that we think could both contribute to our understanding, and constitute relatively low-hanging fruit, are as follows:[6]

- **Different reward function distributions.** The reward function distributions D_{HA} we've looked at in this work have sampled agent rewards from a uniform distribution [0, 1], iid over states. But rewards in the real world are sparser, and often follow power law distributions instead. How does instrumental value change when we account for this?
- **Phase changes at high discount rates.** We've seen one example of a shift in the behavior of an agent at a high discount rate (Agent A blocking Agent H in the corridor), which led to a relative increase in the instrumental misalignment between our agents. It may be worth investigating how widespread this kind of phenomenon is, and what factors influence it.
- **Deeper understanding of physical interactions.** We've only just scratched the surface of agent-agent interactions, with the no-overlap rule as the simplest physical interaction we could think of. We believe incorporating more realistic interactions into future experiments could help improve our understanding of the kinds of alignment dynamics that are likely to occur in the real world.
- **Robustness of instrumental misalignment.** We *think* instrumental misalignment-by-default is a fairly robust phenomenon. But we could be wrong about that, and it would be good news if we were! We'd love to see a more methodical investigation of instrumental misalignment, including isolating the factors that systematically mitigate or exacerbate the effect.

Finally, we hope that our work — and any future experimental results that might leverage our open-source codebase — will serve as a source of intuitions for ideas to understand and mitigate instrumental convergence, and also as a way to test those ideas quickly at a small scale in a simplified setting. While existence proofs for instrumental misalignment are interesting, it's much *more* interesting if we can identify contexts where this kind of misalignment doesn't occur — since these contexts are exactly the ones that may offer hints as to the solution of the full AI alignment problem.

If you're interested in pursuing the above lines of effort or any others, or you'd like to know more about this research, please drop a comment below. You can also reach out directly to Edouard at edouard@gladstone.ai or @harris_edouard on Twitter.

1. ^

   When our two agents' utility functions are logically independent (i.e., there is no mutual information between them) we refer to this as the **independent goals regime** and say that our agents have **independent terminal goals**. In practice, we operationalize this regime by defining a joint reward function distribution D_{HA} (see Section 2) on which the sampled pairs of reward functions (R_H, R_A) are logically independent. In general, rewards and utilities aren't the same thing, but in our particular set of experiments we can treat them as identical. See footnote [1] in Part 1 for more details on this point.

2. ^

   Just as in that previous case, this pattern of POWER-indifference seems to emerge because Agent A is able to efficiently exploit Agent H's deterministic policy. See footnote [10] in Part 2.

3. ^

   The agents in our examples so far could still interact with each other; they just interacted *indirectly*, each one changing the effective "reward landscape" that the other agent perceived. Recall that we sample each agent's reward function by drawing a reward value from a uniform [0, 1] distribution that's iid over all the states of our MDP. That means each agent sees a different reward value at each state, and each state corresponds to a different pair of positions the two agents can take on the gridworld. So when Agent A moves from one cell to another, Agent H suddenly sees a completely different set of rewards over the gridworld cells it can move to (and vice versa).

4. ^

   The maze gridworld has 31 cells. One of our agents can occupy any of those 31 cells, and under the no-overlap rule the other agent can occupy any of the remaining 30 cells that aren't occupied by the first agent.

5. ^

   But note that despite this, our agents are still instrumentally misaligned-by-default, in the sense that $\alpha_{HA} < 0$ in the independent goals regime.

6. ^

   A fifth line of effort, that we allude to in the technical Appendix to Part 2, would be to explore the **effects of different seed policies.** In this sequence, we've only considered the uniform random seed policy for Agent A. Anecdotally, in early tests we

found that the choice of seed policy did in fact affect results fairly strongly, mostly in that *deterministic* seed policies tend to make Agent H less robust to adversarial optimization by Agent A than *random* seed policies do (data not shown). We think it would be worthwhile to investigate the effect of the seed policy systematically, since it's a contingent choice that could have substantial downstream effects.

POWERplay: An open-source toolchain to study AI power-seeking

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual. This is a linkpost for https://github.com/gladstoneai/POWERplay

We're open-sourcing **POWERplay**, a research toolchain you can use to study power-seeking behavior in reinforcement learning agents. POWERplay was developed by Gladstone AI for internal research.

POWERplay's main use is to estimate the **instrumental value** that a reinforcement learning agent can get from a state in an MDP. Its implementation is based on a definition of instrumental value (or "POWER") first proposed by Alex Turner et al. We've extended this definition to cover certain tractable multi-agent RL settings, and built an implementation behind a simple Python API.

We've used POWERplay previously to obtain some suggestive early results in single-agent and multi-agent power-seeking. But we think there may be more low-hanging fruit to be found in this area.

Beyond our own ideas about what to do next, we've also received some interesting conceptual questions in connection with this work. A major reason we're open-sourcing POWERplay is to lower the cost of converting these conceptual questions into real experiments with concrete outcomes, that can support or falsify our intuitions about instrumental convergence.

Ramp-up

We've designed POWERplay to make it as easy as possible for you to get started with it. Follow the installation and quickstart instructions to get moving quickly. Use the replication API to trivially reproduce any figure from any post in our instrumental convergence sequence. Design single-agent and multi-agent MDPs and policies, launch experiments on your local machine, and visualize results with clear figures and animations.

POWERplay comes with "batteries included", meaning all the code samples in the documentation should just work out-of-the-box if it's been installed successfully. It also comes with pre-run examples of experimental results, so you can understand what "normal" output is supposed to look like. While this does make the repo weigh in at about 500 MB, it's worth the benefits of letting you immediately start playing around with visualizations on preexisting data.

If we've done our job right, a smart and curious grad student (with a bit of Python experience) should be able to start reproducing our previous experiments within an hour, and to have some new — and hopefully interesting! — results within a week.

We're looking forward to seeing what people do with this. If you have any questions or comments about POWERplay, feel free to reach out to Edouard at edouard@gladstone.ai.

Clone POWERplay on GitHub.