

# WHEN EMPOWERMENT DISEMPOWERS

**Claire Yang**

University of Washington  
claireyy@uw.edu

**Maya Cakmak**

University of Washington  
mcakmak@uw.edu

**Max Kleiman-Weiner**

University of Washington  
maxhkw@uw.edu

## ABSTRACT

Empowerment, a measure of an agent’s ability to control its environment, has been proposed as a universal goal-agnostic objective for motivating assistive behavior in AI agents. While multi-human settings like homes and hospitals are promising for AI assistance, prior work on empowerment-based assistance assumes that the agent assists one human in isolation. We introduce an open source multi-human gridworld test suite Disempower-Grid. Using Disempower-Grid, we empirically show that assistive RL agents optimizing for one human’s empowerment can significantly reduce another human’s environmental influence and rewards—a phenomenon we formalize as “disempowerment.” We characterize when disempowerment occurs in these environments and show that joint empowerment mitigates disempowerment at the cost of the user’s reward. Our work reveals a broader challenge for the AI alignment community: goal-agnostic objectives that seem aligned in single-agent settings can become misaligned in multi-agent contexts.

## 1 INTRODUCTION

Building aligned agents capable of helping people when their goals are uncertain remains an open problem. A common approach is for an assistant to model a person’s goal or reward function and then take actions to maximize that reward for them (Hadfield-Menell et al., 2016; Leike et al., 2018; Pérez-D’Arpino & Shah, 2015). However, in practice, inferred reward functions are often misidentified, and optimizing even a slightly inaccurate reward function can lead to negative consequences and unsafe behavior (Hong et al., 2023; Freedman et al., 2021; Zhuang & Hadfield-Menell, 2020; Tien et al., 2022).

An alternative approach to creating helpful agents is to train them to empower humans in an open-ended way. Indeed, recent work has shown that agents that optimize for increasing the empowerment of others (Du et al., 2020; Myers et al., 2024) or their choices (Franzmeyer et al., 2022) yield helpful assistants. Furthermore, this class of promising techniques is relatively robust to misspecification because they sidestep the problem of goal inference.

However, across these lines of work, researchers assume a dyadic interaction between two agents: an assistive agent and a simulated human user (Newman et al., 2022). This assumption limits the usefulness of agents for assistance. The real world is fundamentally multi-human. Promising domains for deploying robots and AI agents that help people, such as homes and hospitals, include multiple people aside from the intended users. For instance, in a hospital setting, a robot may have one target of assistance (e.g., a nurse), but it interacts with other people (e.g., patients and other staff). Henry Evans, a quadriplegic user and researcher of assistive robots has said, “No matter how much assistance a device provides to a [adult] patient, it will not be used regularly unless [...] it makes the caregiver’s life a lot easier” (Ranganeni et al., 2024). This requires that AI agents and robots be well aligned in multi-human settings, even if there is only one primary user. We introduce this setting as a *single-principal multi-human assistance game* (SP-MHAG), where there is one principal (i.e. *user*) an AI agent aims to assist, among the presence of multiple humans (i.e. *bystanders*) acting in the environment.

Here, we show that alignment issues arise when empowerment-based assistance is applied in SP-MHAG, a setting that includes other humans who are not the target of assistance but can act and influence the state. When an assistive agent focuses on increasing one person’s empowerment, it may unintentionally reduce another person’s empowerment (i.e. *disempower* them). We show that

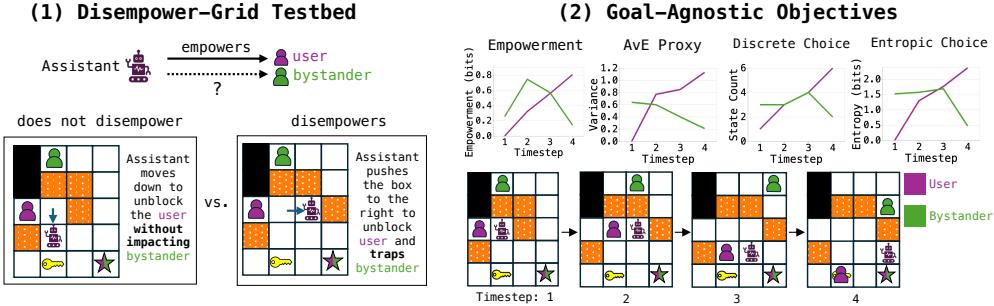


Figure 1: *Left:* Examples from our test suite Disempower-Grid. The assistant aims to empower the user through a goal-agnostic objective. Differing assistance strategies may influence the optionality of a bystander (green). The left shows an example where the assistant enables both the user and the bystander to reach more states, including the goal. The right shows an example where the assistant inhibits the bystander while helping the user. *Right:* Sample trajectory showing that four goal-agnostic objectives used for training an assistive RL agent all increase the user’s influence/choice while decreasing it for the bystander. See Section 4.1 for details on goal-agnostic objectives.

this alignment problem need not emerge from any malicious intent. An AI agent that inadvertently disempowers others could lead to “gradual disempowerment,” where human agency erodes over time (Hammond et al., 2025; Kulveit et al., 2025). We introduce Disempower-Grid, a new test suite of multi-human assistance gridworld environments for benchmarking disempowerment (Figure 1). Across Disempower-Grid, we empirically show evidence for disempowerment across four empowerment and goal-agnostic assistance objectives. Furthermore, we qualitatively characterize when disempowerment happens. Finally, we attempt to mitigate disempowerment with an assistant that maximizes the joint empowerment of both humans. We find that while joint empowerment is very effective in preventing bystander disempowerment, the user’s attained reward is significantly decreased. Safe and effective assistance in multi-human settings remains an important challenge for AI alignment.

The main **contributions** of our work are:

1. Disempower-Grid: a test suite of multi-human assistance gridworlds and implementations of goal-agnostic objectives for assistance, built in JAX for highly efficient training (Bradbury et al., 2018; Rutherford et al., 2024b). The test suite includes diverse environment dynamics, assistant action spaces, and parameterized environment generation. We open source the implementation of the test suite to enable the community to build on our work <sup>1</sup>
2. Empirical evidence that goal-agnostic objectives are misaligned in multi-human settings: using Disempower-Grid, we show that *assistants disempower the bystander* while learning to empower the user across varying environment dynamics, assistant action spaces, and human goals.
3. Joint empowerment mitigates disempowerment at the expense of the user’s reward: we find that extending the assistance objective to include bystander’s empowerment significantly reduces disempowerment, but also worsens the effectiveness of the assistant in helping the user attain their goals. This highlights an important challenge for alignment in real-life scenarios where AI agents must effectively assist one human in the presence of others.

## 2 RELATED WORK

We combine key ideas from goal-agnostic objectives and connect them to assistance and AI safety.

**Goal-Agnostic Assistance: Empowerment, Choice, and Power** Our work builds on key ideas from reinforcement learning and control that aim to measure an agent’s control and capability in an

<sup>1</sup><https://github.com/claireyyang/disempower-grid>

environment. *Empowerment*, defined as the maximum mutual information between an agent’s action and its future states, is a goal-agnostic measure of capability (Klyubin et al., 2005b;a; Song et al., 2025). An agent’s effective empowerment (the mutual information, not the maximum of the mutual information) has been used as an intrinsic motivation for reinforcement learning agents, and shown to enhance their learning and exploration across domains (Brändle et al., 2023; Baddam et al., 2025; Lidayan et al., 2025). It has also been applied to improve agent coordination in multi-agent settings (van der Heiden et al., 2020; Kim et al., 2023; Guckelsberger et al., 2016). Intuitively, effective empowerment measures an agent’s potential to navigate efficiently through a state space. Agents with greater mastery and control over their environment or those that can access a larger fraction of available states will have higher effective empowerment. For example, if two agents are locked in two separate rooms, the agent with a key to get out would have higher effective empowerment than the one without, since the agent with a key would also be able to access states beyond the locked room. Finally, Turner & Tadepalli (2022) demonstrates that reinforcement learning-based agents are power-seeking (as measured by increases in optionality), suggesting that the majority of reward functions reward maximizing future choices (Turner et al., 2023).

Recent work uses approximations of effective empowerment as an objective for assistance. Importantly, these models can help human users without needing to model their goals (Du et al., 2020; Myers et al., 2024). The appeal is intuitive: by maximizing a human’s empowerment, an agent should help them achieve as many possible states in the future without needing to explicitly infer those goals. Because calculating empowerment is computationally intractable in high-dimensional environments, several approximations have been developed to scale its measurement (Mohamed & Rezende, 2015; Myers et al., 2024; Jung et al., 2012). Franzmeyer et al. (2022) develop an assistive agent that optimizes the number of *choices* available to another agent. They develop multiple estimators for choice and show that the resulting agent acts prosocially across multiple contexts without access to external rewards. They demonstrate these results in environments where the assistive agent’s action space is limited to moving around the environment, without influencing the layout. Compared to empowerment, choice is simpler to compute because it only depends on the agent’s states (although a state transition matrix must also be estimated). Regardless, prior works on goal-agnostic assistance focus on dyadic interactions between an assistant and a simulated human user, or assume that the user and bystanders are adversaries (Du et al., 2020; Myers et al., 2024; Franzmeyer et al., 2022). They do not measure the impact of these assistance objectives on other agents in the environment.

**Side Effects** There is a rich literature on studying the unintended side effects of AI action and assistance (Amodei et al., 2016; Krakovna et al., 2019; Turner et al., 2020; Krakovna et al., 2020). Most related to our work, Krakovna et al. (2020) develops a method to encourage agents to leave environments intact by incentivizing them to consider the reward a future agent would achieve in that same environment. While this setup does involve thinking about a disadvantaged third-party, the agents are not directly interacting with each other and there is no assistant.

### 3 RESEARCH QUESTIONS

**RQ1: Under what conditions do assistants optimizing for one human’s influence/choice systematically harm other humans in multi-human environments?** We hypothesize that disempowerment occurs systematically across different assistant capabilities, environmental constraints, and human goals, suggesting this is a fundamental property of empowerment-based assistance in multi-human settings, rather than an artifact of specific experimental conditions.

**RQ2: What environmental factors determine when bystander disempowerment occurs versus when it can be avoided?** We hypothesize that bystander disempowerment mainly occurs when the bystander encounters limited resources that can be influenced by the assistant, even if the payoffs for the humans are not zero-sum. Examples of this from our gridworlds are spatial bottlenecks in the environment that can be blocked by the assistant moving a box. This distinction shows how goal-agnostic objectives highly depend on the interaction between the environmental layout and its dynamics, without necessarily aligning with the underlying rewards.

**RQ3: Can multi-agent extensions of the empowerment objective mitigate disempowerment?** We hypothesize that naively adding the bystander’s empowerment to the objective will not fully

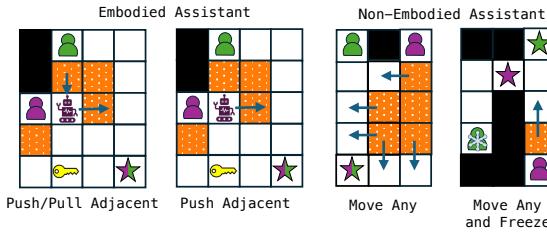


Figure 2: Our experiments demonstrate disempowerment on these four example grids from Disempower-Grid. The user (green) and the bystander (purple) are both rewarded for reaching the star after touching the key. The task is not competitive, and both agents can occupy the star square simultaneously. The user and bystander move in cardinal directions, cannot move through each other, and cannot move the blocks (orange) or the walls (black). *Left two grids:* when the assistant is embodied, the assistant can move in cardinal directions and can move adjacent blocks by pushing/pulling, or only pushing. The user and bystander cannot move through the assistant when embodied. Each human must go to the key and then the goal position (star) in order to receive their independent reward. *Right two grids:* when the assistant is non-embodied, it can move any of the blocks, or freeze the bystander in place for 4 timesteps. Each human only needs to go to the goal position (star) in order to receive their independent reward.

address the problem of disempowerment, especially as the human’s underlying goals vary (and potentially conflict) within the environment.

## 4 METHODS

**Preliminaries** The setting is an assistance game with multiple humans (Hadfield-Menell et al., 2016). This differs from the multi-principal assistance game setting in that instead of the AI agent acting on behalf of all humans in the setting, it acts on behalf of a single human in the presence of other humans (Fickinger et al., 2020). We define a single-principal multi-human assistance game (SP-MHAG) as an assistance game in which a single principal (i.e. *user*) is assisted by an AI agent within an environment that includes other humans who are not explicit targets of assistance but can act, influence the state, and must not be significantly disempowered as a result of the assistant’s behavior.

More specifically, we formulate the reinforcement learning setting as a Multi-agent Markov Decision Process (MMDP), defined by  $(S, \Omega_H, \Omega_A, A_U, A_B, A_A, P, R_U, R_B, \gamma)$ . The (simulated) humans ( $H$ ) are partitioned into two disjoint sets: a user ( $U$ ) and a bystander ( $B$ ).  $S$  represents the full environment state space,  $\Omega_H$  is the observation function for the humans, and  $\Omega_A$  is the observation function for the assistant. The assistant does not observe either of the humans’ goals.  $\pi_U$ ,  $\pi_B$ , and  $\pi_A$  are the policies for the three agents. The assistant maximizes the expected sum of discounted rewards  $\mathbb{E}[\sum_t \gamma^t R_A]$ , where  $R_A$  is equal to a goal-agnostic objective  $O(\cdot)$ .  $O$  will be replaced with one of four goal-agnostic objectives introduced in Equations 4,5,6, 7. Thus, the assistant’s optimal policy  $\pi_A^*$  is one that maximizes the future discounted goal-agnostic objective of the user:

$$\pi_A^* = \operatorname{argmax}_{\pi_A} \sum_{t=0}^{\infty} \gamma^t O(\cdot). \quad (1)$$

Further details on the setting and training can be found in Appendix A.1, A.2.

### 4.1 GOAL-AGNOSTIC ASSISTANCE OBJECTIVES

This section introduces four goal-agnostic objectives  $O(\cdot)$  that we use to train assistants.

**Empowerment:** To compute the potential ability of an agent to affect future states using its actions, Klyubin et al. (2005b) defines the empowerment of a state  $s_t$  as the maximum mutual information between the action sequence and future states after horizon  $T$  timesteps:

$$\mathcal{E}(s) = \max_{p(a|s)} I(A_T; S_T | s), \quad (2)$$

where  $p(a|s)$  is the probability distribution over actions given the state,  $I(\cdot)$  is the mutual information between an action sequence of size  $T$  sampled from the action space  $A_T$  and the states after horizon  $T$  timesteps, conditioned on the input state.

We translate this equation into the assistive setting, in which an assistant is calculating the user's effective empowerment  $\mathcal{E}^U$ , which is the mutual information between the user's actions and its future states computed under the user's policy, rather than the maximum possible mutual information (Myers et al., 2024). The action sequence  $A_U$  under consideration are of the user's actions. Thus, Equation 2 can be transformed for use by the reinforcement learning assistant in our setting as such:

$$\mathcal{E}^U(s) = I(A_T^U; S_T^U | s, \pi_U). \quad (3)$$

However, because the assistant does not know  $\pi_U$ , it cannot exactly compute effective empowerment. Instead, we calculate an approximation of effective empowerment by assuming that the user's policy is random and conducting sparse sampling of the action sequences (Salge et al., 2014). This is a worst-case assumption for when the user's policy is unknown or only known probabilistically. It is also more robust for cases where the user's behavior is unpredictable to the assistant and the noise model is unknown (Du et al., 2020; Salge et al., 2014). As a result of this assumption, this calculation acts as a lower bound on true effective empowerment because the entropy of the user's actions is maximized under the uniform policy:

$$\tilde{\mathcal{E}}^U(s) = I(A_T^U; S_T^U | s, \pi_{U_{\text{uniform}}}), \quad (4)$$

where  $\pi_{U_{\text{uniform}}} = 1/|A|$ . We calculate  $\tilde{\mathcal{E}}^U(s)$  by sampling multiple forward rollouts under the uniform user policy. This approximation of effective empowerment is labeled as empowerment in future figures and analyses.

**AvE Proxy:** Du et al. (2020) introduced an efficient proxy for empowerment based on the variance of the user's states at the end of trajectory rollouts. We include this proxy as an additional way to estimate effective empowerment:

$$\text{AvE}(s) = \text{Var}(S_T^U | s, \pi_{U_{\text{uniform}}}), \quad (5)$$

where  $S_T^U$  are the user's final states after horizon  $T$  steps. This proxy is also computed through sampling forward rollouts, with the assumption that the user's policy is uniformly random. The proxy is calculated as the variance of the final states of the rollouts. Intuitively, this means that if the final states are highly dissimilar in their features, then the value of the AvE proxy is high.

**Discrete Choice:** Discrete choice is defined as the number of reachable states by the user within horizon  $T$  from the current state (Franzmeyer et al., 2022). It is one of three methods for estimating future state availability introduced in their work. These metrics are simpler than empowerment, as they do not require calculating the mutual information between the actions and the future states. As a result, this method's implicit assumption is that the agent's influence on the environment is tied to how many states they will have access to. However, they assume that future available states are only influenced by the agent's own actions. While this assumption does not hold in our multi-human setting, where the bystander can also influence the user's future states, we include it for completeness as a goal-agnostic objective for assistance that is distinct from empowerment.

$$\text{DC}(s) = |s' : s' \in \text{Reachable}(s, T, \pi_{U_{\text{uniform}}})|, \quad (6)$$

where  $\text{Reachable}(s, T, \pi_{U_{\text{uniform}}})$  is the sampled set of states reachable from  $s$  in exactly  $T$  steps under the user's uniform policy and  $|\cdot|$  denotes the number of unique elements.

**Entropic Choice:** Entropic choice is another method for estimating future state availability, based on the conditional state entropy  $H(\cdot)$  (Franzmeyer et al., 2022). It acts as a lower bound on discrete choice.

$$\text{EC}(s) = H(S_T | s, \pi_{U_{\text{uniform}}}). \quad (7)$$

## 4.2 THE DISEMPWER-GRID TEST SUITE

To systematically test these research questions, we introduce **Disempower-Grid**, a suite of diverse parameterized gridworlds and implementations of the goal-agnostic objectives for training assistive AI agents in multi-human settings. The designs of the environments were inspired by those proposed

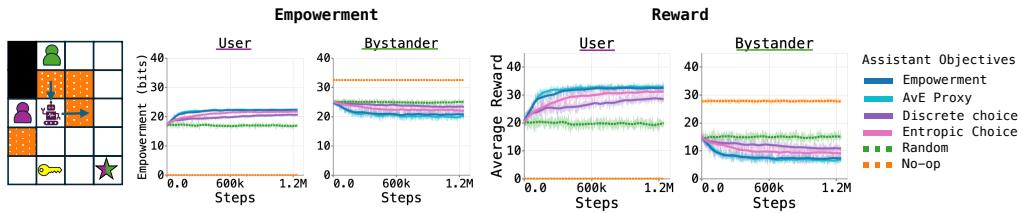


Figure 3: Assistant disempowers bystander in the *Push/Pull Adjacent* environment example grid. Left: An example grid where the assistant (robot) must push/pull the boxes (orange dotted) to empower the user (purple). Center/Right: The bystander (green) is disempowered by the assistant’s actions. The average empowerment and average reward of the user and bystander across the assistant’s training in grid from Disempower-Grid shown on the left. Each trace is averaged over five runs. The error bands show the standard deviation. Empowerment and reward levels are compared against an assistant with a Random objective (green dotted line). Subsequent figures follow the same format: example environment (left), empowerment trajectories (center), and reward trajectories (right).

in Du et al. (2020); Leike et al. (2017). However, unlike the test suites introduced before, these environments contain an additional bystander agent that is not the target of assistance. Additionally, the environments are designed for general-sum payoffs between humans, which allows for diverse interactions. Disempower-Grid is open source to enable further research on disempowerment in multi-human settings. Disempower-Grid is built in JaxMARL, allowing for highly efficient training (Rutherford et al., 2024a).

By varying the action space and embodiment of the assistant, we test our central hypothesis: assistants optimizing for empowerment will consistently disempower bystanders across varying constraints, showing that disempowerment in multi-human assistance is a fundamental alignment issue.

## 5 RESULTS

We qualitatively show how disempowerment occurs across four example grids with varying environment dynamic and action spaces. See Figure 2 for visualizations and descriptions of the grids.

**1. Spatial Bottlenecks (Push/Pull Adjacent)** Across all goal-agnostic objectives, the assistant disempowers the bystander by blocking the hallway with the box (Figure 3). The bystander initially has higher empowerment and reward than the user when acting with a no-op or random assistant. However, as the assistant learns to maximize the user’s empowerment, it inadvertently reduces the bystander’s. In this layout, the assistant pushes the box to its right, blocking the bystander (green) from exiting the hallway. Although this action increases the user’s (purple’s) empowerment, the user would have been equally helped if the assistant had simply moved aside without pushing the box. Lacking knowledge of either human’s goals, the assistant leaves the box in place after pushing it to the right, never pulling it back to unblock the hallway. Consequently, the bystander’s disem-

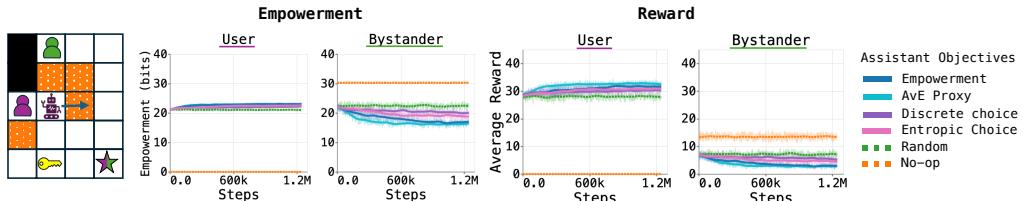


Figure 4: Assistant disempowers bystander in the *Push Adjacent* environment example grid, despite constrained capabilities.

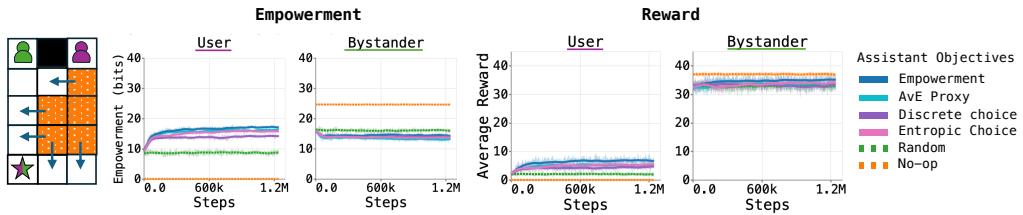


Figure 5: Non-embodied assistant disempowers bystander in the *Move Any* environment example grid.

powerment is ignored. This effect was consistent and statistically significant across all four assistant objectives.

**2. Constrained Assistant Capabilities (Push Adjacent)** In this condition, the embodied assistant can move around and only push the boxes to unblock the user. This means that the assistant is more limited in its ability to modify the environment and may also cause irreversible changes to the environment. Disempowerment occurs even when the assistant’s abilities are constrained in this way, and when the assistant has the ability to cause permanent side effects. Compared to the *Push/Pull Adjacent* environment, the assistant is not able to empower the user as much. However, even though the assistant does not assist the user as much, it still disempowers the bystander to a similar degree (shown in Figure 4). Effects were consistent and significant for all four assistant objectives. These results support our hypothesis for RQ1 that bystander disempowerment occurs across different assistant capabilities and environmental constraints.

**3. Non-Embodied Assistance (Move Any)** In this condition, the assistant is non-embodied and can move any box in the grid to an adjacent open position. This distinction creates a fundamental difference: non-embodied agents do not need to navigate the physical space to exert their influence. Compared to the previous two conditions, the assistant has a much larger action space since it can move any box, rather than only adjacent boxes. However, because it is disembodied, the assistant cannot physically block the agents, so it does not need to strategize its own positioning. Still, we observe strong evidence of bystander disempowerment in this environment’s example gridworld (see Figure 5). Effects were consistent and significant for all four assistant objectives. These results support our hypothesis for RQ1 that bystander disempowerment occurs across different environmental constraints.

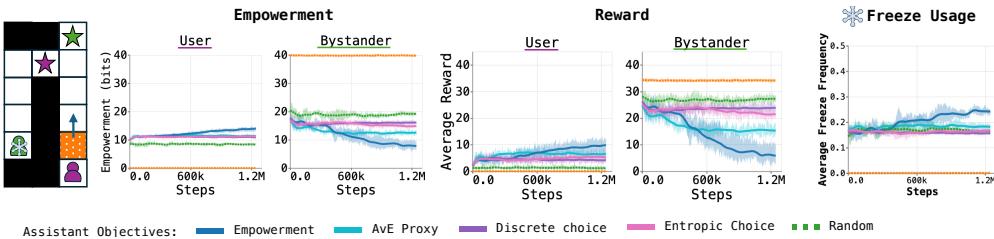


Figure 6: Non-embodied assistant directly freezes when given the opportunity in the *Move Any* and *Freeze* environment example grid, with the usage of the freeze action increasing over training.

**4. Direct Intervention (Move Any and Freeze)** In the *Move Any and Freeze* example, the non-embodied agent can freeze the bystander for four timesteps, in addition to moving any box around. In this environment, the box can also be moved to cover the goal. As seen in Figure 6, the bystander is disempowered. Moreover, the assistant learns during training to freeze the bystander. The disabling behavior was learned from optimizing the user empowerment objective alone. There is a positive relationship between freeze usage, the empowerment of the user, and the disempowerment of the bystander. Effects were consistent and significant for all four assistant objectives. These results

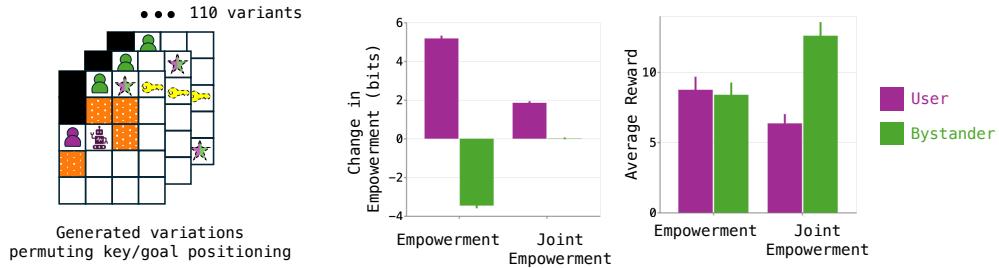


Figure 7: *Left:* 110 procedurally generated variations of the Push/Pull Adjacent environment by permuting the key and goal positions. *Center:* Change in empowerment from initial training (first 5 epochs) to final performance (last 5 epochs), averaged across the 106 layout variations where the empowerment-maximizing assistant disempowers the bystander. Joint empowerment significantly reduces the user’s empowerment gain while significantly increasing the bystander’s empowerment, compared to the empowerment-only objective. *Right:* Average episode reward over the last five epochs, for the same 106 layouts. Joint empowerment significantly decreases the user’s reward, compared to the empowerment condition. Standard error shown through error bars.

support our hypothesis for RQ1 that bystander disempowerment occurs across different assistant capabilities.

**Robustness Across Goal Variations** We procedurally generated 110 variations of the *Push/Pull Adjacent* environment by changing key and goal locations while keeping walls, blocks, and initial agent positions constant. Disempowerment was nearly universal for the empowerment objective: only four variations (all with the goal positioned in the hallway) did not demonstrate the assistant disempowering the bystander. This finding is consistent Klyubin et al. (2005b)’s insight that empowerment is determined by an agent’s ability to act on the environment. When the goal is positioned in the hallway, the assistant is unable to push the box to the right and block the hallway. Critically, even in scenarios with different user and bystander goals, disempowerment persisted, which supports our RQ2 hypothesis that bystander disempowerment arises from how spatial constraints can be manipulated by the assistant, rather than conflict arising from the underlying human goals themselves.

**Joint empowerment** A naive approach to preventing disempowerment is to include the bystander’s empowerment in the assistant’s objective, together with the user’s empowerment. van der Heiden et al. (2020) originally proposed this approach and showed that it improves multi-agent co-ordination in cooperative tasks. Instead of only maximizing the user’s empowerment, the assistant maximizes the sum of the user’s and bystander’s individual empowerment. We experiment with this objective to test whether it prevents bystander disempowerment. The joint empowerment calculation is as follows (reference Section 4.1 for variable definitions):

$$\tilde{\mathcal{E}}^{U+B}(s) = I(A_T^U; S_T^U | s, \pi_{U\text{uniform}}) + I(A_T^B; S_T^B | s, \pi_{B\text{uniform}}). \quad (8)$$

We find that joint empowerment substantially mitigates bystander disempowerment, but *significantly decreases the user’s reward*, compared to the empowerment objective. In the 106 environment layouts where an empowerment-maximizing assistant disempowers the bystander, an assistant optimizing the joint empowerment objective actually learns to increase the bystander’s empowerment in 52% of layouts (statistical significance:  $p < 0.001$ , effect size:  $d = 0.78$ ) and learns to not significantly impact the bystander’s empowerment in the remaining 48% ( $p = 0.10$ , non-significant). However, across all 106 layouts, an assistant maximizing joint empowerment significantly reduced the user’s reward ( $p < 0.001$ ) while simultaneously and significantly increasing the bystander’s reward ( $p < 0.001$ ), compared to an assistant optimizing empowerment (see Figure 7).

These results do not support our original RQ3 hypothesis. Contrary to our initial expectation, joint empowerment highly mitigates disempowerment across varying goal placements. However, this finding raises a critical question in the SP-MHAG setting: if the user’s reward is substantially de-

creased to benefit the bystander, joint empowerment may not represent an effective assistance objective. Moreover, we highlight a potential scalability limitation: joint empowerment would likely prove challenging in environments with multiple bystanders, as it requires strong, potentially unrealistic assumptions about the assistant’s familiarity with each bystander’s action space. For comprehensive details on the layout analysis and additional visualizations, refer to Appendix A.3.

## 6 DISCUSSION

Across multiple environments and goal-agnostic objectives, we showed that an assistant designed to empower a user can unintentionally disempower other humans in the same environment. This reveals a fundamental tension in goal-agnostic AI assistance: even when assistants avoid the harms of explicit goal misspecification, the empowerment objective itself can be misspecified in multi-agent settings. Despite joint empowerment being effective in mitigating disempowerment, it reduces the assistant’s capacity to provide effective assistance to the user, which remains the core objective in the SP-MHAG setting. Our results challenge the assumption that goal-agnostic objectives are inherently safer than goal-directed ones, an important consideration as real-world assistants increasingly operate in multi-human contexts.

**Implications for AI Safety and Cooperative AI** Our experiments demonstrate that empowerment maximization can induce tradeoff-like dynamics even in general-sum environments where humans’ rewards are independent. The assistant often selected actions that increased the user’s empowerment at the bystander’s expense, even when reversible or mutually beneficial actions were available.

Existing AI safety approaches address negative externalities by constraining an agent’s influence or side effects on the environment (Amodei et al., 2016). However, as shown in the *Push Adjacent* setting (see Figure 4), limiting the assistant’s ability to move boxes reduced its effectiveness in assisting the user achieve their goals compared to the *Push/Pull Adjacent* case (see Figure 3). The joint empowerment objective also implicitly imposed a constraint on negative side effects to the bystander to a similar effect. This suggests that such restrictions may prevent harmful side effects only by limiting the assistant’s ability to assist altogether.

**Limitations** The environment layouts in our experiments were hand-designed, as disempowerment does not occur in every spatial configuration. It requires the state space to contain a bottleneck—in our Disempower-Grid case, a narrow hallway that only one human or embodied assistant can traverse at a time. Our examples also initialized the user trapped in a corner to illustrate the strongest contrast between increased user empowerment and bystander disempowerment. However, that setup is not strictly necessary for disempowerment to emerge.

**Future Work** A key open question is how to design assistance objectives that help the intended user without unintentionally harming others. While our work empirically demonstrates disempowerment, future research could formalize the conditions under which disempowerment occurs in SP-MHAG. Such formulations could make it possible to predict when and quantify how much assistive AI agents disempower others. We also found that joint empowerment mitigates disempowerment, but at the cost of assistance. Future work could explore hybrid approaches that combine empowerment with inferring general human goals or social norms of appropriateness (Leibo et al., 2024) to mitigate disempowerment of bystanders while still providing effective assistance to the user. Lastly, future work could also extend the notion of disempowerment beyond spatial or navigation-based settings to nonspatial and continuous domains. These extensions may help us understand how AI assistance across agentic systems like large language models and robots can remain helpful in real-world settings while ensuring that supporting one user or group does not systematically disadvantage others.

## 7 CONCLUSION

We demonstrate how an AI agent optimizing for its intended goal-agnostic objective (to maximize the empowerment or choice of a human user) can cause negative externalities to another human in the environment. This challenges a fundamental assumption in AI safety that goal-agnostic objectives alone reduce alignment failures.

## 8 REPRODUCIBILITY STATEMENT

To ensure reproducibility, we publish our source code and include documentation describing how to reproduce the training and experiments.

## 9 LLM USAGE STATEMENT

LLMs were used to aid in giving feedback and suggestions on the structure of sentences.

## ACKNOWLEDGMENTS

We thank Jinyeop Song, Kunal Jha, Aly Lidayan, Jared Moore, Michael Dennis, and Alpri Else for their insightful discussions and support during this research. We also gratefully acknowledge funding and support from the Cooperative AI Foundation, the Foresight Institute, the Sony Research Award Program, and the UW-Tsukuba Amazon NVIDIA Cross Pacific AI Initiative (XPAI). One author is supported by the National Science Foundation Computer and Information Science and Engineering Graduate Fellowship under Grant No. 2313998. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## REFERENCES

- Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety, July 2016. URL <http://arxiv.org/abs/1606.06565> [cs]. arXiv:1606.06565 [cs].
- Vasanth Reddy Baddam, Behdad Chalaki, Vaishnav Tadiparthi, Hossein Nourkhiz Mahjoub, Ehsan Moradi-Pari, Hoda Eldardiry, and Almuatazbellah Boker. In Search of a Lost Metric: Human Empowerment as a Pillar of Socially Conscious Navigation, January 2025. URL <http://arxiv.org/abs/2501.01539>. arXiv:2501.01539 [cs].
- James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL <https://github.com/jax-ml/jax>.
- Franziska Brändle, Lena J. Stocks, Joshua B. Tenenbaum, Samuel J. Gershman, and Eric Schulz. Empowerment contributes to exploration behaviour in a creative video game. *Nature Human Behaviour*, 7(9):1481–1489, September 2023. ISSN 2397-3374. doi: 10.1038/s41562-023-01661-2. URL <https://www.nature.com/articles/s41562-023-01661-2>. Publisher: Nature Publishing Group.
- Yuqing Du, Stas Tiomkin, Emre Kiciman, Daniel Polani, Pieter Abbeel, and Anca Dragan. AvE: Assistance via Empowerment. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4560–4571. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/30de9ece7cf3790c8c39ccff1a044209-Abstract.html>.
- Arnaud Fickinger, Simon Zhuang, Dylan Hadfield-Menell, and Stuart Russell. Multi-Principal Assistance Games, July 2020. URL <http://arxiv.org/abs/2007.09540>. arXiv:2007.09540 [cs].
- Tim Franzmeyer, Mateusz Malinowski, and Joao F. Henriques. Learning Altruistic Behaviours in Reinforcement Learning without External Rewards. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=KxbhdypHE>.
- Rachel Freedman, Rohin Shah, and Anca D. Dragan. Choice set misspecification in reward inference. *CoRR*, abs/2101.07691, 2021. URL <https://arxiv.org/abs/2101.07691>.

- Christian Guckelsberger, Christoph Salge, and Simon Colton. Intrinsically motivated general companion npcs via coupled empowerment maximisation. In *2016 IEEE conference on computational intelligence and games (CIG)*, pp. 1–8. IEEE, 2016.
- Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. Cooperative inverse reinforcement learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS’16, pp. 3916–3924, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.
- Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčíak, et al. Multi-agent risks from advanced ai. *arXiv preprint arXiv:2502.14143*, 2025.
- Joey Hong, Kush Bhatia, and Anca Dragan. On the sensitivity of reward inference to misspecified human models. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=hJqGbUpDGV>.
- Tobias Jung, Daniel Polani, and Peter Stone. Empowerment for Continuous Agent-Environment Systems, January 2012. URL <http://arxiv.org/abs/1201.6583>. arXiv:1201.6583 [cs].
- Woojun Kim, Whiyoung Jung, Myungsik Cho, and Youngchul Sung. A variational approach to mutual information-based coordination for multi-agent reinforcement learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS ’23, pp. 40–48, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450394321.
- Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv. All Else Being Equal Be Empowered. In Mathieu S. Capcarrère, Alex A. Freitas, Peter J. Bentley, Colin G. Johnson, and Jon Timmis (eds.), *Advances in Artificial Life*, pp. 744–753, Berlin, Heidelberg, 2005a. Springer. ISBN 978-3-540-31816-3. doi: 10.1007/11553090\_75.
- A.S. Klyubin, D. Polani, and C.L. Nehaniv. Empowerment: a universal agent-centric measure of control. In *2005 IEEE Congress on Evolutionary Computation*, volume 1, pp. 128–135 Vol.1, September 2005b. doi: 10.1109/CEC.2005.1554676. URL <https://ieeexplore.ieee.org/document/1554676/>. ISSN: 1941-0026.
- Victoria Krakovna, Laurent Orseau, Ramana Kumar, Miljan Martic, and Shane Legg. Penalizing side effects using stepwise relative reachability, March 2019. URL <http://arxiv.org/abs/1806.01186>. arXiv:1806.01186 [cs, stat].
- Victoria Krakovna, Laurent Orseau, Richard Ngo, Miljan Martic, and Shane Legg. Avoiding side effects by considering future tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS ’20, pp. 19064–19074, Red Hook, NY, USA, December 2020. Curran Associates Inc. ISBN 978-1-7138-2954-6.
- Jan Kulveit, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud. Gradual disempowerment: Systemic existential risks from incremental ai development. *arXiv preprint arXiv:2501.16946*, 2025.
- Joel Z Leibo, Alexander Sasha Vezhnevets, Manfred Diaz, John P Agapiou, William A Cunningham, Peter Sunehag, Julia Haas, Raphael Koster, Edgar A Duéñez-Guzmán, William S Isaac, et al. A theory of appropriateness with applications to generative artificial intelligence. *arXiv preprint arXiv:2412.19010*, 2024.
- Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A. Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. AI Safety Gridworlds, November 2017. URL <http://arxiv.org/abs/1711.09883>. arXiv:1711.09883 [cs].
- Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

- Aly Lidayan, Yuqing Du, Eliza Kosoy, Maria Rufova, Pieter Abbeel, and Alison Gopnik. Intrinsic-Motivated Humans and Agents in Open-World Exploration, March 2025. URL <http://arxiv.org/abs/2503.23631>. arXiv:2503.23631 [cs].
- Shakir Mohamed and Danilo J. Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pp. 2125–2133, Cambridge, MA, USA, 2015. MIT Press.
- Vivek Myers, Evan Ellis, Sergey Levine, Benjamin Eysenbach, and Anca Dragan. Learning to Assist Humans without Inferring Rewards. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 71540–71567. Curran Associates, Inc., 2024. URL [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/83a4ea71b13bc86308a2bd0b5e07fb61-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/83a4ea71b13bc86308a2bd0b5e07fb61-Paper-Conference.pdf).
- Benjamin A. Newman, Reuben M. Aronson, Kris Kitani, and Henny Admoni. Helping People Through Space and Time: Assistance as a Perspective on Human-Robot Interaction. *Frontiers in Robotics and AI*, 8:720319, January 2022. ISSN 2296-9144. doi: 10.3389/frobt.2021.720319. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8829116/>.
- Claudia Pérez-D'Arpino and Julie A. Shah. Fast target prediction of human reaching motion for cooperative human-robot manipulation tasks using time series classification. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6175–6182, 2015. doi: 10.1109/ICRA.2015.7140066.
- Vinitha Ranganeni, Vy Nguyen, Henry Evans, Jane Evans, Julian Mehu, Samuel Olatunji, Wendy Rogers, Aaron Edsinger, Charles Kemp, and Maya Cakmak. Robots for Humanity: In-Home Deployment of Stretch RE2. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 1299–1301, Boulder CO USA, March 2024. ACM. ISBN 979-8-4007-0323-2. doi: 10.1145/3610978.3641114. URL <https://dl.acm.org/doi/10.1145/3610978.3641114>.
- Alexander Rutherford, Benjamin Ellis, Matteo Gallici, Jonathan Cook, Andrei Lupu, Gardar Ingvarsson, Timon Willi, Ravi Hammond, Akbir Khan, Christian Schroeder de Witt, Alexandra Souly, Saptarashmi Bandyopadhyay, Mikayel Samvelyan, Minqi Jiang, Robert Tjarko Lange, Shimon Whiteson, Bruno Lacerda, Nick Hawes, Tim Rocktaschel, Chris Lu, and Jakob Nicolaus Foerster. JaxMARL: Multi-Agent RL Environments and Algorithms in JAX, November 2024a. URL <http://arxiv.org/abs/2311.10090>. arXiv:2311.10090 [cs].
- Alexander Rutherford, Benjamin Ellis, Matteo Gallici, Jonathan Cook, Andrei Lupu, Gardhar Ingvarsson, Timon Willi, Ravi Hammond, Akbir Khan, Christian Schroeder de Witt, Alexandra Souly, Saptarashmi Bandyopadhyay, Mikayel Samvelyan, Minqi Jiang, Robert Tjarko Lange, Shimon Whiteson, Bruno Lacerda, Nick Hawes, Tim Rocktaschel, Chris Lu, and Jakob Nicolaus Foerster. Jaxmarl: Multi-agent rl environments and algorithms in jax. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024b.
- Christoph Salge, Cornelius Glackin, and Daniel Polani. Changing the environment based on empowerment as intrinsic motivation. *Entropy*, 16(5):2789–2819, 2014.
- Jinyeop Song, Jeff Gore, and Max Kleiman-Weiner. Estimating the empowerment of language model agents. *arXiv preprint arXiv:2509.22504*, 2025.
- Jeremy Tien, Jerry Zhi-Yang He, Zackory Erickson, Anca D Dragan, and Daniel S Brown. Causal confusion and reward misidentification in preference-based reward learning. *arXiv preprint arXiv:2204.06601*, 2022.
- Alex Turner and Prasad Tadepalli. Parametrically retargetable decision-makers tend to seek power. *Advances in Neural Information Processing Systems*, 35:31391–31401, 2022.
- Alexander Matt Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. Conservative Agency via Attainable Utility Preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and*

*Society*, AIES ’20, pp. 385–391, New York, NY, USA, February 2020. Association for Computing Machinery. ISBN 978-1-4503-7110-0. doi: 10.1145/3375627.3375851. URL <https://dl.acm.org/doi/10.1145/3375627.3375851>.

Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal Policies Tend to Seek Power, January 2023. URL <http://arxiv.org/abs/1912.01683>. arXiv:1912.01683 [cs].

Tessa van der Heiden, Christoph Salge, Efstratios Gavves, and Herke van Hoof. Robust multi-agent reinforcement learning with social empowerment for coordination and communication. *arXiv preprint arXiv:2012.08255*, 2020.

Simon Zhuang and Dylan Hadfield-Menell. Consequences of misaligned ai. *Advances in Neural Information Processing Systems*, 33:15763–15773, 2020.

## A APPENDIX

### A.1 ENVIRONMENT DETAILS

In our environments implemented in Disempower-Grid,  $R_U(s_t) = 1$  if the user reaches its assigned goal  $g_U \in S$ , 0 otherwise.  $R_B(s_t) = 1$  if the bystander reaches its assigned goal  $g_B \in S$ , 0 otherwise. The user and bystander may be assigned to the same goal or different goals. Regardless, the reward each agent receives is fully independent of that of the other agent. The state to observation mapping function differs between the human and assistant.  $\Omega_H$  includes the goals pursued by the user and bystander, while  $\Omega_A$  does not, i.e., the assistant has no knowledge of the user or bystander’s goal.

At time  $t$ , the humans (user and bystander) observe  $\omega_t^H \in \Omega_H(s_t)$ , and the assistant observes  $\omega_t^A \in \Omega_A(s_t)$ . Action selection happens simultaneously. The user selects action  $a_t^U \sim \pi_U(\cdot | \omega_t^H)$ , the bystander selects action  $a_t^B \sim \pi_B(\cdot | \omega_t^H)$ , and the assistant selects action  $a_t^A \sim \pi_A(\cdot | \omega_t^A)$ .

### A.2 TRAINING DETAILS

First,  $\pi_U$  and  $\pi_B$  are trained simultaneously using PPO, using separate actor and critic networks. During the training of the user and bystander policies, the assistant selects actions according to a random policy  $\pi^{random_A}(a^U | \omega^A)$ . The assistant is included in this phase of training so that the user and bystander agents can learn the dynamics of the environment with the assistant present. We decided to use a random policy so that the user and bystander’s policies are not biased by an intentionally helpful or unhelpful assistant and experience a wide range of possible states from random exploration.

After  $\pi_U$  and  $\pi_B$  have converged they are frozen and  $\pi_A$  is trained using PPO with one of the goal-agnostic assistance rewards (see Section 4.1). During this phase, the user and bystander act according to their fixed policies  $\pi_U^*$  and  $\pi_B^*$ , respectively. This models an assistant learning its policy while interacting with capable humans that have seen many possible states.

### A.3 FURTHER DETAILS ON JOINT EMPOWERMENT

In Section 5, we analyzed joint empowerment across 106 different layouts, categorizing them into two groups: 51 layouts where the bystander’s empowerment increased (not disempowered), and 55 layouts where the bystander was still disempowered. Our empowerment change metric was calculated by subtracting the mean empowerment over the first five training epochs from the mean over the final five epochs for both the user and bystander. A layout-objective pair was specifically labeled as “disempowered” when two conditions were met: (1) the user’s empowerment increased, and (2) the bystander’s empowerment simultaneously decreased. Layouts not meeting both these criteria were categorized as “not disempowered”.

Contrary to our initial hypothesis for RQ3, we found a consistent pattern across all layouts: the bystander’s empowerment under joint empowerment was significantly higher than under the standard

empowerment objective ( $p < 0.001$ ). Even in cases where the bystander's empowerment was not explicitly increased, it remained close to the random baseline. We observed two distinct scenarios:

- Joint empowerment increased the bystander's empowerment. In these cases, the average empowerment across the last five training epochs was 2.08% higher than the random baseline ( $p < 0.001$ ).
- Joint empowerment did not increase the bystander's empowerment. In these cases, the average empowerment was 0.59% lower than the random baseline, though this difference was not statistically significant.

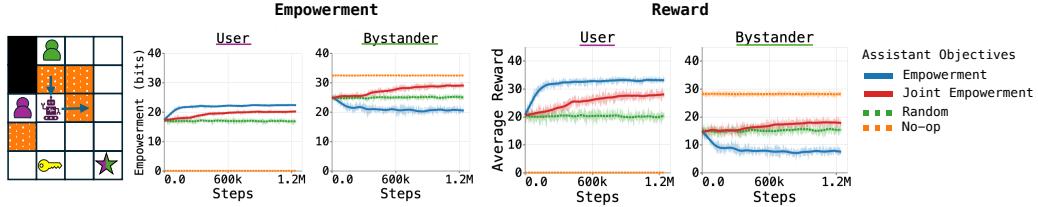


Figure 8: In the original Push/Pull Adjacent environment example (one of 110 layout variations), the assistant maximizing the joint empowerment avoids disempowering the bystander, but significantly decreases the user's empowerment and reward, compared to when maximizing the user's empowerment. Importantly, note that joint empowerment still performs significantly better than an assistant acting randomly for both the user and bystander's empowerment and reward.

#### A.4 DISEMPOWER-GRID CODE

The code for Disempower-Grid can be found at <https://github.com/claireyyang/disempower-grid>. All experiments were run on a single NVIDIA GeForce RTX 4090 GPU and are reproducible on CPU, GPU, or TPU. Training the user, bystander, and assistant on one layout across one objective takes 55 seconds. To train the user, bystander, and assistant on one layout on all objectives presented in this paper takes around 4 minutes 26 seconds.