WHEN TO RELY ON A PARTNER? RELIABILITY BASED ARBITRATION FOR ZERO SHOT COOPERATION WITH DECISION PROCESS IMPAIRMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Training agents that cooperate with humans requires adapting to diverse partner behaviors, which cannot be fully anticipated before interaction. Zero shot coordination (ZSC) addresses this by exposing agents to a wide range of partner behaviors during training, covering differences in skill level, strategy, preference, and action stochasticity. Such coverage, however, presumes partners whose decision processes are intact. Agents deployed as assistive or collaborative partners must also work with people whose decision processes are impaired, as in cognitive and psychiatric conditions. Such impairment does not lower competence uniformly: it selectively affects behavior during the task, so the same partner may act competently in one situation and fail in another. A fixed cooperation policy is therefore not enough; the agent has to keep reassessing how much to rely on the partner as the interaction unfolds. Adapting behavior across different situations is a fundamental aspect of human cognitive control. Reliability based arbitration control has been identified as a mechanism that supports such adaptation by adjusting the influence of competing strategies according to their estimated reliability. Inspired by this principle, we propose Reliability based Dual Specialist Arbitration (RDSA). RDSA decomposes cooperation into two strategies, each based on a different assumption about the partner's reliability: a Partner-Reliant Specialist that relies on the partner to handle part of the task, and a Self-Reliant Specialist that completes the task on its own. During interaction, RDSA estimates each specialist's reliability and dynamically adjusts their relative influence on action selection. To evaluate RDSA, we constructed simulated partners with decision process impairments by applying targeted perturbations to perception, valuation, and action selection. Experiments showed that these partners exhibited behavioral variation distinct from skill level, with the differences depending on the situation. RDSA outperformed the best performing ZSC baseline and the best static mixture in most evaluated conditions, including conditions in which neither specialist outperformed the best performing ZSC baseline.

1 INTRODUCTION

Training agents that can cooperate with humans has long been a central challenge in human–AI interaction and multi agent reinforcement learning (Carroll et al., 2019; Strouse et al., 2021). In cooperative tasks, team performance depends heavily on an agent's ability to adapt to its partner's behavior (Strouse et al., 2021; Yan et al., 2023). However, such adaptation is difficult to achieve in advance because human behavior varies substantially across individuals. Therefore, training cannot cover the full range of behaviors an agent may encounter during interaction (Strouse et al., 2021; Yu et al., 2023). This difficulty has motivated zero shot coordination (ZSC), which aims to enable agents to cooperate with previously unseen partners. Prior work has addressed this difficulty by exposing agents to diverse partner behaviors during training, such as differences in skill levels, strategies, and preferences (Strouse et al., 2021; Yu et al., 2023; Zhao et al., 2023; Yan et al., 2023; Liang et al., 2024). However, prior work on cognitive and psychiatric conditions has shown that human behavioral variation also arises from impairments in specific decision processes, a source of partner diversity that remains less explored in ZSC (Duffy et al., 2025; del Río et al., 2026; Shin &

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

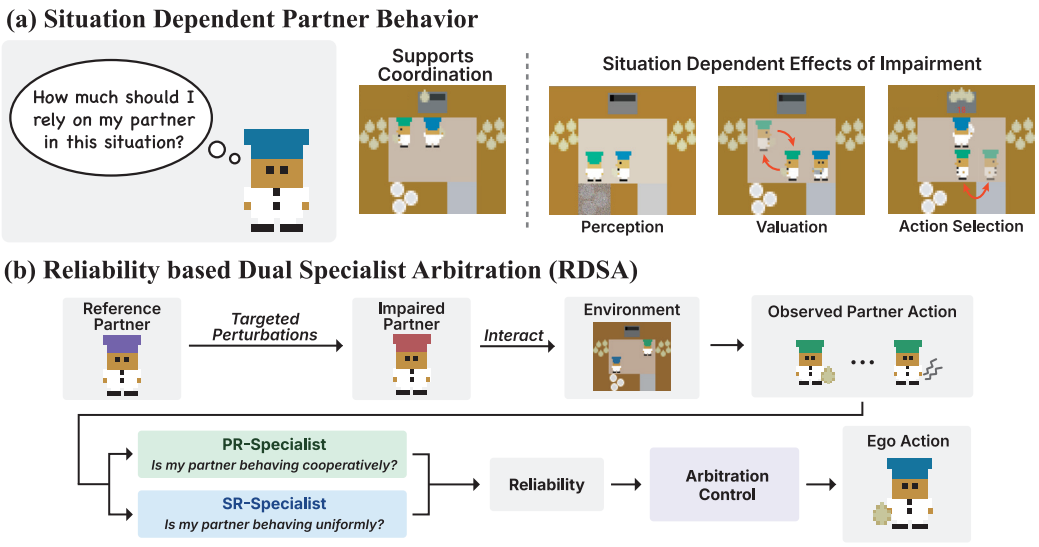


Figure 1: **RDSA overview.** (a) Decision process impairment affects partner behavior differently across situations, motivating dynamic adjustment of partner reliance. (b) RDSA estimates the reliability of the PR- and SR-Specialists from how well observed partner behavior aligns with their respective partner assumptions and dynamically adjusts their relative influence on ego action selection through arbitration control.

Niv, 2021; Lee, 2013; Vaghi et al., 2017; Heerey et al., 2008). Agents deployed as assistive or collaborative partners must therefore be able to cooperate with people affected by such impairments. Enabling cooperation with such partners begins with understanding their behavior. Decision process impairment has been shown to manifest as behavioral variation across situations rather than a uniform reduction in skill level (Vaghi et al., 2017; del Río et al., 2026; Attaallah et al., 2024; Hunter et al., 2022), as illustrated by computational models of schizophrenia (Al-Hejji et al., 2025). Variation in impairment type and severity limits coverage during training, motivating adaptation to ongoing partner behavior (Mirsky et al., 2022; Santos et al., 2021; Faraone et al., 2024; Leucht et al., 2025).

The ability to adapt behavior across different situations is a fundamental aspect of human cognitive control (Miller & Cohen, 2001; Shenhav et al., 2013; Dajani & Uddin, 2015). Neuroscience research has identified reliability based arbitration control as a mechanism for adaptive behavior, with the reliability of each strategy determining its relative influence on behavior (Lee et al., 2014; Kim et al., 2019; 2021). Inspired by this principle, we introduce Reliability based Dual Specialist Arbitration (RDSA) to dynamically adapt to the situation dependent behaviors of impaired partners. We conceptualize cooperation through two complementary strategies that differ in whether the ego agent assumes cooperative contribution from the partner: a Partner-Reliant Specialist (PR-Specialist), which assumes cooperative partner behavior, and a Self-Reliant Specialist (SR-Specialist), which does not. During interaction, RDSA adjusts their influence on action selection based on estimated reliabilities.

Inspired by prior work that conceptualizes human decision making as a set of subprocesses, we constructed simulated partners with decision process impairments by applying targeted perturbations to the corresponding components of a computational model (Heekeren et al., 2008; Rangel et al., 2008; Suri et al., 2020). Experiments showed that the simulated partners exhibited a key characteristic of human decision process impairment: behavioral variation beyond simple skill level differences. Across diverse Overcooked layouts, RDSA outperformed both the best ZSC baseline and the best static mixture in most conditions, including conditions in which neither specialist outperformed the best ZSC baseline. In summary, our key contributions are as follows: (1) identifying decision process impairment as a less explored source of ZSC partner diversity; (2) introducing RDSA, which dynamically adjusts the relative influence of two specialists through arbitration control; and (3) demonstrating its effectiveness through extensive experiments.

2 Related Works

2.1 Modeling partner diversity for human–AI cooperation

Cooperative tasks in the real world require AI agents to adapt to diverse human behaviors. Training cannot cover the full range of partner behaviors that an agent may encounter during interaction. To address this challenge, ZSC has emerged to enable cooperation with previously unseen partners (Strouse et al., 2021; Yu et al., 2023). Existing ZSC methods address this challenge by representing diverse partner behaviors during training, but differ in how they construct this diversity. FCP constructs a partner population from intermediate training checkpoints (Strouse et al., 2021). These partners represent differences in conventions and skill levels. MEP promotes diversity in partner action distributions through a population entropy bonus (Zhao et al., 2023). This encourages behavioral diversity both across policies and within each policy. E3T mixes the ego policy with a random policy to diversify partner behavior (Yan et al., 2023). HSP constructs a diverse partner population by varying event based feature weights in hidden utility functions that model partner preferences (Yu et al., 2023). GAMMA uses a variational autoencoder to encode partner behavior into latent variables that capture partner specific strategies (Liang et al., 2024). These approaches improve ZSC performance by incorporating such diversity during training.

2.2 Computational approaches to impairments in human decision processes

Computational neuroscience has sought to understand human decision processes by decomposing them into subprocesses such as perception, valuation and action selection (Heekeren et al., 2008; Rangel et al., 2008; Suri et al., 2020). This approach has been extended to cognitive and psychiatric impairments by perturbing decision related components to characterize behavioral effects (Lee, 2013; Maia & Frank, 2011; Rolls et al., 2008). For example, Al-Hejji et al. (2025) modeled schizophrenia related alterations by pruning excitatory connections to reduce excitation. Dobs et al. (2022) selectively disrupted task relevant CNN filters to impair the corresponding recognition function while preserving others. Ilboudo & Tanaka (2025) modeled decision paralysis by separating intent and action selection, producing delayed or failed decisions. Other computational models have reproduced behavioral and neural changes associated with disorders by altering dopamine related gain, synaptic connectivity, receptor function, and corticostriatal dynamics (Cohen & Servan-Schreiber, 1992; Frank, 2005; Spencer, 2009; Moustafa & Gluck, 2011). Together, these works reflect diverse efforts to understand impairments in human decision processes.

3 Background

3.1 Two player cooperative Markov game

We treat Overcooked as a two player cooperative Markov game between an ego agent with policy π_e and a partner with policy π_p (Littman, 1994; Carroll et al., 2019). At timestep t, the ego agent and partner take actions $a_e^t \in \mathcal{A}_e$ and $a_p^t \in \mathcal{A}_p$, forming the joint action $\mathbf{a}^t = (a_e^t, a_p^t)$ in state $s_t \in \mathcal{S}$. They receive the shared reward $R(s_t, a_e^t, a_p^t)$ and transition to $s_{t+1} \sim P(\cdot \mid s_t, a_e^t, a_p^t)$. The environment is $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, R, \gamma, T \rangle$, where $\mathcal{A} = \mathcal{A}_e \times \mathcal{A}_p$, γ is the discount factor, and T is the episode horizon (Littman, 1994; Lowe et al., 2017). Given π_p, the expected discounted return is $J(\pi_e, \pi_p) = \mathbb{E}_{\tau \sim (\pi_e, \pi_p)} \left[\sum_{t=0}^{T-1} \gamma^t R(s_t, a_e^t, a_p^t) \right]$, where τ denotes a trajectory.

3.2 Dueling DQN with joint action values

We use a dueling DQN with a shared CNN encoder and separate state value and joint action advantage heads to estimate joint action values $Q_\vartheta(s, a_e, a_p)$ (Mnih et al., 2015; Wang et al., 2016; Smith & Zhang, 2025). The outputs are combined using mean centered aggregation (Wang et al., 2016):

$$Q_\vartheta(s, a_e, a_p) = V_\vartheta(s) + A_\vartheta(s, a_e, a_p) - \frac{1}{|\mathcal{A}|} \sum_{(a'_e, a'_p) \in \mathcal{A}} A_\vartheta(s, a'_e, a'_p), \tag{1}$$

where ϑ denotes network parameters and $\mathcal{A} = \mathcal{A}_e \times \mathcal{A}_p$ is the joint action space. The network outputs $|\mathcal{A}_e| \times |\mathcal{A}_p|$ joint action values and is trained by temporal difference learning (Sutton, 1988).

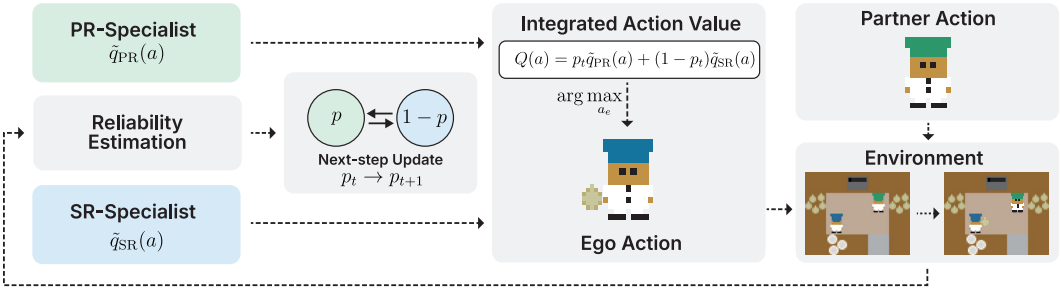


Figure 2: **RDSA framework.** At timestep t, p_t determines the relative influence of the standardized PR- and SR-Specialist action values on ego action selection. After observing the partner action, RDSA updates the specialists' reliabilities and uses them to update p_{t+1} for the next timestep.

4 METHOD

To cooperate with partners whose behavior varies across situations due to decision process impairment, we propose RDSA. Arbitration control was introduced to explain human adaptive behavior by estimating the reliability of complementary strategies and adjusting their relative influence on action selection (Lee et al., 2014). RDSA extends this principle to cooperation by redefining both the complementary strategies being arbitrated and the evidence used to estimate their reliability. Specifically, we conceptualize cooperation through two complementary strategies that differ in whether the ego agent assumes cooperative contribution from the partner. The PR- and SR-Specialists are designed to represent these two strategies. Rather than estimating reliability from prediction errors, RDSA estimates the reliability of each specialist from observed partner behavior as the interaction unfolds. The estimated reliabilities determine the relative influence of the two specialists, enabling RDSA to adapt its cooperation strategy to situation dependent variation in partner behavior.

4.1 REFERENCE PARTNER AND SPECIALISTS

To construct impaired partners and the two specialists, we introduce three models: a Reference Partner Model (Ref), a PR-Specialist, and an SR-Specialist. All three models estimate joint action values $Q_m(s, a_e, a_p)$, where s is the current state, $a_e \in \mathcal{A}_e$ is the ego action, $a_p \in \mathcal{A}_p$ is the partner action, and $m \in \{\mathrm{Ref}, \mathrm{PR}, \mathrm{SR}\}$. Ref uses a joint Q DQN, while both specialists use dueling joint Q DQNs. For each model, we denote the feasibility masked joint action value for (a_e, a_p) by $G_m(s, a_e, a_p)$. The three models follow the same value based training procedure with temporal difference updates, but differ in how their joint action values are used to select actions. At each training timestep, each model selects a joint action based on $G_m(s, a_e, a_p)$ using its model-specific action selection procedure, executes it in the environment, and observes the reward and next state.

To construct impaired partners as previously unseen entities for evaluation, we perturb a Ref that is trained separately from the PR- and SR-Specialists and differs from them in both network architecture and action selection mechanism. For these perturbations to meaningfully represent impairments in human decision processes, the unimpaired Ref should first reflect their general characteristics. To achieve this, we model **Ref** to exhibit bounded rationality using a recursive quantal response equilibrium (QRE) procedure that terminates the recursion at $\ell = k$ rather than iterating to convergence, following prior work on bounded rational reasoning (Simon, 1955; McKelvey & Palfrey, 1995; Smith & Zhang, 2025). During training, the action distributions of both agents are computed recursively as follows:

$$\pi_i^{(\ell)}(a_i \mid s) \propto \exp\left(\lambda\, \mathbb{E}_{a_{-i}\sim\pi_{-i}^{(\ell-1)}}\left[G_{\mathrm{Ref}}(s, a_e, a_p)\right]\right), \qquad \pi_i^{(0)} = \mathrm{Unif}(\mathcal{A}_i), \tag{2}$$

where $i \in \{e, p\}$ denotes the updating agent, $-i$ the other agent, and ℓ the recursion step.

The **PR-Specialist** is trained under the assumption that the partner behaves cooperatively. The ego and partner initialize their action distributions uniformly and alternately apply the following deterministic $\arg\max$ update:

$$a_i' = \arg\max_{a_i} \sum_{a_{-i}} \pi_{-i}(a_{-i}) G_{\mathrm{PR}}(s, a_e, a_p), \quad \pi_i'(a_i) = \mathbf{1}[a_i = a_i'], \quad i \in \{e, p\}. \tag{3}$$

Here, e and p denote the ego and partner, respectively; i denotes the updating agent, and $-i$ the other agent. At iteration j, the ego update uses $\pi_p^{(j)}$, and the subsequent partner update uses $\pi_e^{(j+1)}$.

The **SR-Specialist** is trained without assuming cooperative partner behavior. During training, it models the partner as uniformly random and selects the ego action using Eq. 4. Further details are provided in Appendix A.1.

$$a_e^* = \arg\max_{a_e} \frac{1}{|\mathcal{A}_p|} \sum_{a_p \in \mathcal{A}_p} G_{\mathrm{SR}}(s, a_e, a_p). \tag{4}$$

4.2 Reliability based Dual Specialist Arbitration

With the specialists fixed during interaction, RDSA adjusts only their relative influence. For each ego action, the PR- and SR-Specialists aggregate their joint action values over partner actions according to their respective partner assumptions:

$$q_{\mathrm{PR}}(a_e) = \max_{a_p \in \mathcal{F}(a_e)} Q_{\mathrm{PR}}(s, a_e, a_p), \qquad q_{\mathrm{SR}}(a_e) = \frac{1}{|\mathcal{F}(a_e)|} \sum_{a_p \in \mathcal{F}(a_e)} Q_{\mathrm{SR}}(s, a_e, a_p). \tag{5}$$

Here, $\mathcal{F}(a_e) \subseteq \mathcal{A}_p$ denotes the feasible partner actions for a_e, and $\mathcal{E}(s) = \{a_e \in \mathcal{A}_e : |\mathcal{F}(a_e)| > 0\}$ denotes the feasible ego actions. After standardizing the two specialists' values over $\mathcal{E}(s)$, RDSA selects an action as follows:

$$a_e^t = \arg\max_{a_e \in \mathcal{E}(s_t)} \left[p_t \tilde{q}_{\mathrm{PR}}(a_e) + (1 - p_t) \tilde{q}_{\mathrm{SR}}(a_e) \right], \tag{6}$$

where $p_t \in [0, 1]$ determines the relative influence of the PR-Specialist. The partner independently selects a_p^t, and the joint action is executed. After observing a_p^t, RDSA evaluates how closely the partner's behavior aligns with each specialist's partner assumption. For the PR-Specialist, RDSA quantifies this as

$$r_t^{\mathrm{PR}} = \sum_{a_p \in \mathcal{F}(a_e^t)} \pi_t^{\mathrm{BR}}(a_p) \Big[\mathbf{1}\{v_t(a_p) < v_t(a_p^t)\} + \tfrac{1}{2} \mathbf{1}\{v_t(a_p) = v_t(a_p^t)\} \Big], \quad \delta_t^{\mathrm{PR}} = \tfrac{1}{2} - r_t^{\mathrm{PR}}. \tag{7}$$

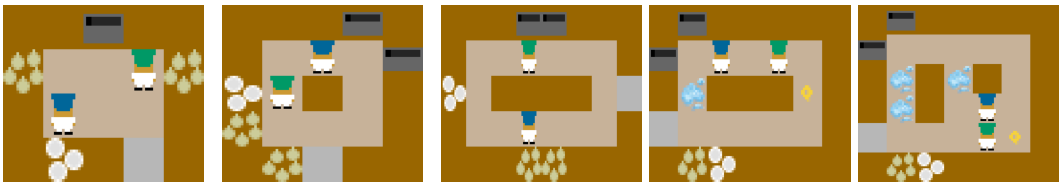
Here, $v_t(a_p) = G_{\mathrm{PR}}(s_t, a_e^t, a_p)$ denotes the value of partner action a_p, and π_t^{BR} is the soft partner action distribution obtained using Eq. 2, with G_{PR} in place of G_{Ref}. Ref uses the same QRE procedure, but different joint action values generally yield different distributions. Because only a single partner action is observed at each timestep and multiple actions can have similarly high cooperative value, directly comparing action distributions would not adequately reflect the PR-Specialist's assumption. We therefore use the probability weighted percentile r_t^{PR}, which measures the relative value rank of the observed action a_p^t under π_t^{BR}, following the stochastic superiority measure of Vargha & Delaney (2000). Under the cooperative partner assumption, $\mathbb{E}[r_t^{\mathrm{PR}}] = 0.5$, making δ_t^{PR} centered at 0. Thus, 0 serves as the evidence threshold: a larger r_t^{PR} indicates a higher relative value under the cooperative partner distribution, whereas $\delta_t^{\mathrm{PR}} > 0$ provides evidence against the cooperative assumption.

For the SR-Specialist, RDSA quantifies alignment with the uniform partner assumption as

$$e_t^{\mathrm{SR}} = \pi_t^{\mathrm{BR}}(a_p^t) - \frac{1}{|\mathcal{F}(a_e^t)|}, \qquad \delta_t^{\mathrm{SR}} = (1 - \kappa)\delta_{t-1}^{\mathrm{SR}} + \kappa e_t^{\mathrm{SR}}. \tag{8}$$

The raw signal e_t^{SR} compares the probability that π_t^{BR} assigns to the observed action with its uniform probability. Accordingly, $\mathbb{E}[e_t^{\mathrm{SR}}] = 0$ under the uniform partner assumption. However, uniform sampling can select an action with high probability under π_t^{BR}, producing positive deviations. RDSA therefore smooths the signal using an exponential moving average with coefficient κ and applies a positive threshold τ_{SR}. Thus, $\delta_t^{\mathrm{SR}} > \tau_{\mathrm{SR}}$ provides evidence against the uniform partner assumption; otherwise, the evidence is insufficient.

Following prior work (Lee et al., 2014), RDSA estimates each specialist's reliability from a Dirichlet posterior over accumulated evidence categories and uses these reliabilities to update p_{t+1}. By introducing two specialists and redefining the evidence, RDSA extends the principle of human adaptive control to cooperation with partners affected by decision process impairment. Further details are provided in Appendix A.2.



(a) Cramped Room (b) Coordination Ring (c) Counter Circuit (d) Risky Dualpath (e) Risky Multipath

Figure 3: **Overcooked layouts.** The first three are standard layouts and the last two are risky layouts. In the risky layouts, blue puddles are slip hazards that cause agents to drop carried objects, while yellow buttons are subgoals that temporarily remove the hazards.

4.3 Decision process impairment modeling

To model decision process impairment in a controlled manner, we constructed impaired partners by selectively perturbing Ref's perception, valuation, or action selection. This design isolates impairment to a specific decision subprocess and allows its severity to be systematically varied. For **perception impairment**, motivated by prior work (Kumar et al., 2022; Li et al., 2025), we added Gaussian noise to an observation channel group sampled at the beginning of each episode before the CNN encoder. For **valuation impairment**, inspired by neural connectivity models (Al-Hejji et al., 2025), we pruned positive MLP weights, fine tuned the network, and injected noise at the MLP input during inference. In both cases, we controlled impairment severity via the noise standard deviation. For **action selection impairment**, we varied the rationality parameter λ in the QRE (Eq. 2) (McKelvey & Palfrey, 1995; Smith & Zhang, 2025); decreasing λ reduces value sensitivity and increases action stochasticity. Further details are provided in Appendix A.3.

5 Experiment

5.1 Experiment setup

We evaluated RDSA in Overcooked, a two-player cooperative gridworld widely used as a benchmark for human–AI cooperation (Carroll et al., 2019; Strouse et al., 2021; Yu et al., 2023), across three standard layouts (*Cramped Room*, *Coordination Ring*, *Counter Circuit*) and two risky layouts (*Risky Dualpath*, *Risky Multipath*) adapted from Risky Overcooked (Smith & Zhang, 2025). To focus on decision process impairment rather than risk sensitivity, we added a subgoal that temporarily removes the hazards in the risky layouts. Our experiment addressed five complementary questions: (1) whether the simulated partners exhibited behavioral variation beyond that attributable to skill level and across situations (Sections 5.2 and 5.3); (2) how RDSA performed relative to five ZSC baselines, the PR- and SR-Specialists, and their static mixtures (Sections 5.4 and 5.5); (3) whether this effectiveness extended to impaired partners constructed from behavioral cloning (BC) human proxies commonly used for evaluation in Overcooked (Carroll et al., 2019; Strouse et al., 2021; Yu et al., 2023; Zhao et al., 2023) (Appendix C.1); (4) whether the arbitration control weight tracked variation in partner behavior and impairment severity (Appendices C.2 and C.3); and (5) which conditions were associated with the benefit of RDSA using an LMM (Appendix C.4). Experimental details are provided in Appendix B.

5.2 Does decision process impairment produce behavioral differences beyond skill level variation?

To evaluate whether our simulated partners exhibited behavioral variation beyond simple skill level differences, we collected 40 intermediate Ref checkpoints throughout training and matched each impaired partner to the checkpoint whose cross-play return was closest to its own. Cross-play return is the cumulative shared reward when a partner agent cooperates with Ref, whereas self-play return is the same quantity for two instances of the same agent. We used cross-play return for matching because self-play returns clustered at low or high values in some layouts, preventing close return matches. We refer to the intermediate Ref checkpoint with the closest cross-play return as the return matched checkpoint. To measure state conditional behavioral differences between each impaired

Table 1: **Behavioral differences from return matched checkpoints.** Entries report the mean JS divergence, in bits, between impaired partners and their return matched intermediate checkpoints over the same 308 probe states, averaged across the three impairment types. The Baseline column reports the mean JS divergence among return matched Ref checkpoints within each layout.

Layout	Mild	Moderate	Severe	Average	Baseline
Cramped Room	0.193	0.435	0.471	0.366	0.090
Coordination Ring	0.236	0.395	0.452	0.361	0.120
Counter Circuit	0.212	0.449	0.511	0.390	0.192
Risky Dualpath	0.221	0.335	0.328	0.295	0.182
Risky Multipath	0.229	0.314	0.275	0.273	0.183
Average	**0.218**	**0.385**	**0.408**	**0.337**	0.153

partner and its return matched checkpoint, we evaluated both agents on the same 308 probe states and computed the Jensen–Shannon (JS) divergence between their action distributions. The probe states comprised 25 states from each of the eight backward curriculum stages (Smith & Zhang, 2025) and 108 states sampled from Ref self-play trajectories. JS divergence was expressed in bits and bounded between 0 and 1. We evaluated mild, moderate, and severe impairment by varying perturbation strength; severity increased with σ_{perc} and σ_{val} for perception and valuation, and with decreasing λ for action selection. Additional results are provided in Appendix B.3. To construct the Ref baseline, we first evaluated each intermediate Ref checkpoint in cross-play with the fully trained Ref and grouped checkpoints with similar cross-play returns within each layout. Within each return group, we computed the JS divergence between the action distributions of the intermediate checkpoints on the same probe states and averaged these divergences. The mean across the return groups was used as the Ref baseline for that layout.

Table 1 shows that impaired partners generally exhibited greater state conditional behavioral differences than the Ref baseline, and that these differences tended to increase with impairment severity. These results indicate that the simulated impaired partners exhibited behavioral differences beyond those explained by simple skill level variation among Ref agents with comparable returns.

5.3 Does a fixed impairment produce situation dependent behavioral variation?

To evaluate whether our simulated partners exhibited situation dependent behavioral variation even under a fixed impairment setting, we generated cross-play trajectories with Ref as the ego and each impaired partner as the partner. At each visited state, we computed the JS divergence, in bits, between the impaired partner's action distribution and that of its return matched checkpoint evaluated from the same partner position, measuring how their behavioral differences varied across states.

As shown in Figure 4, the behavioral effect of a fixed impairment varied substantially across visited states. Higher impairment severity generally produced larger JS divergence, whereas under mild impairment, behavioral variation from the return matched checkpoint was often small across states. These results indicate that the simulated partners exhibited situation dependent behavioral variation rather than a uniform behavioral shift, with impairment severity further modulating the magnitude of this variation. Additional results are provided in Figure 5 in Appendix B.4.

5.4 Does RDSA improve performance with partners affected by decision process impairment?

To evaluate how effectively RDSA adapts to partners affected by decision process impairment, we compared it with the PR- and SR-Specialists and five ZSC baselines across 15 layout–impairment conditions. Each condition was evaluated over 1,000 episodes with distinct random seeds; impairment severity was sampled per episode and shared across methods for paired evaluation.

As shown in Table 2, RDSA achieved the highest overall mean episode return of 157.0 and outperformed all compared methods in 12 of the 15 conditions. Holm-corrected two-sided paired t-tests (Holm, 1979) showed significant improvements over the PR- and SR-Specialists in 14 con-

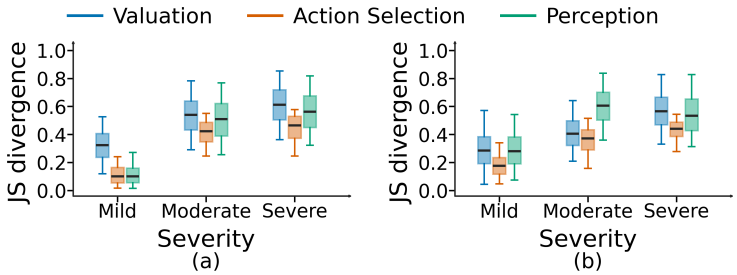


Figure 4: **State dependent behavioral differences.** JS divergence, in bits, between each impaired partner and its return matched intermediate checkpoint across states visited during cross-play with Ref. Panels (a) and (b) show *Cramped Room* and *Risky Dualpath*, respectively. The x-axis denotes impairment severity, and the y-axis shows JS divergence after a 10-step moving average. Boxes show the interquartile range, center lines the median, and whiskers the 5th and 95th percentiles.

Table 2: **Performance across layout–impairment conditions.** Each condition reports mean episode return ± 95% normal CI over 1,000 episodes. Best ZSC is the highest-performing method among FCP, HSP, MEP, GAMMA, and E3T; in the Average row, it denotes the ZSC baseline with the highest overall average. Bold indicates the highest mean per condition, and underlines mark cases where both specialists underperform Best ZSC but RDSA surpasses it.

Layout	Impairment	RDSA	PR	SR	Best ZSC
Cramped Room	Valuation	**192.9±1.8**	161.8±3.2	176.2±2.9	179.1±2.1 (H)
	Action Sel.	**196.7±1.2**	159.4±3.4	175.8±2.9	185.9±1.8 (H)
	Perception	**196.4±1.6**	171.2±3.1	182.3±2.7	187.4±2.0 (H)
Coordination Ring	Valuation	**181.7±4.1**	171.7±4.7	145.8±2.4	161.6±2.7 (M)
	Action Sel.	**158.9±2.4**	142.8±3.5	139.4±2.0	147.4±1.9 (M)
	Perception	**156.3±2.4**	138.2±3.4	142.1±1.8	147.4±1.8 (M)
Counter Circuit	Valuation	**115.5±1.9**	82.7±3.5	95.6±2.4	69.3±1.8 (F)
	Action Sel.	**113.3±1.8**	62.5±3.6	102.4±1.9	77.2±1.3 (F)
	Perception	**119.2±1.5**	52.0±3.5	111.2±1.5	62.8±1.9 (F)
Risky Dualpath	Valuation	171.9±2.0	172.2±2.6	164.5±2.0	**194.8±1.7 (F)**
	Action Sel.	**191.5±2.6**	176.3±4.1	161.9±1.7	177.5±2.2 (M)
	Perception	**190.7±4.5**	179.6±5.7	135.4±2.6	165.6±3.2 (M)
Risky Multipath	Valuation	112.6±1.8	104.1±2.4	111.8±1.6	**124.6±1.7 (F)**
	Action Sel.	134.3±1.6	125.7±2.3	120.7±2.1	**140.7±1.5 (F)**
	Perception	**122.2±2.7**	109.5±4.0	95.2±2.3	117.6±2.6 (F)
Average		157.0±17.3	134.0±21.7	137.4±15.3	140.9±21.8 (F)

ditions each, over FCP, HSP, and MEP in 12 conditions each, and over GAMMA and E3T in all 15 conditions. RDSA significantly underperformed FCP, HSP, and MEP in only 3, 2, and 1 conditions, respectively; overall, 100 of 105 pairwise comparisons were significant with Holm-adjusted $p < 0.01$. The full results and statistical test results are provided in Appendices B.5 and B.6. Notably, in 7 of the 10 conditions where both specialists underperformed the best performing ZSC baseline, RDSA surpassed that baseline, demonstrating the advantage of dynamically adjusting their relative influence through arbitration control. All significant underperformance against ZSC baselines occurred in the risky layouts. Averaged across impairment types and severity levels, the JS divergence from return matched checkpoints was lower in Risky Dualpath (0.295) and Risky Multipath (0.273) than in the three standard layouts (0.361–0.390; Table 1). Thus, impaired partners in the risky layouts behaved more similarly to their return matched checkpoints. This pattern may partly explain the competitiveness of FCP, which explicitly represents skill level variation through intermediate checkpoints.

Table 3: **Static mixture comparison.** Mean episode returns ± 95% CI for RDSA and the condition wise best static mixture across 15 conditions, with its PR-Specialist weight α. Here, $\alpha = 0$ and 1 correspond to the SR- and PR-Specialists, respectively; bold indicates the higher mean.

Layout	Impairment	Best static mixture (α)	RDSA
Cramped Room	Valuation	187.9±1.9 ($\alpha = 0.3$)	**192.9±1.8**
	Action Selection	192.0±1.4 ($\alpha = 0.3$)	**196.7±1.2**
	Perception	192.1±1.7 ($\alpha = 0.3$)	**196.4±1.6**
Coordination Ring	Valuation	171.7±4.7 ($\alpha = 1.0$)	**181.7±4.1**
	Action Selection	143.4±1.8 ($\alpha = 0.3$)	**158.9±2.4**
	Perception	145.3±1.8 ($\alpha = 0.3$)	**156.3±2.4**
Counter Circuit	Valuation	96.0±2.4 ($\alpha = 0.1$)	**115.5±1.9**
	Action Selection	102.8±1.9 ($\alpha = 0.1$)	**113.3±1.8**
	Perception	111.2±1.5 ($\alpha = 0.0$)	**119.2±1.5**
Risky Dualpath	Valuation	**175.4±2.5** ($\alpha = 0.8$)	171.9±2.0
	Action Selection	181.3±3.9 ($\alpha = 0.9$)	**191.5±2.6**
	Perception	181.8±5.6 ($\alpha = 0.8$)	**190.7±4.5**
Risky Multipath	Valuation	**113.7±1.5** ($\alpha = 0.1$)	112.6±1.8
	Action Selection	130.2±1.9 ($\alpha = 0.7$)	**134.3±1.6**
	Perception	115.4±3.4 ($\alpha = 0.7$)	**122.2±2.7**

5.5 Does arbitration control outperform static mixtures?

To evaluate the advantage of arbitration control, we replaced the dynamic mixture weight with a static mixture weight α, where α denotes the weight on the PR-Specialist. We evaluated 11 weights, $\alpha \in \{0.0, 0.1, \ldots, 1.0\}$, across the same 15 conditions using the same set of 1,000 distinct random seeds across all compared methods within each condition.

Table 3 shows that RDSA outperformed the best performing static mixture in 13 of the 15 conditions. The best α varied across conditions, taking seven distinct values between 0 and 1. Across the 165 condition weight comparisons, Holm-corrected paired t-tests (Holm, 1979) showed that no static mixture significantly outperformed RDSA. The full results and statistical test results are provided in Appendices B.5 and B.6. These results indicate that a fixed mixture does not consistently exploit the complementary strengths of the two specialists, whereas RDSA adjusts their relative influence during interaction to better leverage those complementary strengths.

6 Conclusion

In this work, we investigated decision process impairment as a less explored source of partner diversity in ZSC. Unlike simple skill level variation, such impairment produces situation dependent behavioral variation, motivating adaptation to partner behavior during interaction. Inspired by human adaptive control, we proposed RDSA and evaluated it with simulated partners constructed by perturbing model components corresponding to human decision subprocesses. RDSA outperformed the compared methods overall, supporting the benefit of dynamically adjusting the relative influence of complementary specialists. More broadly, assistive and collaborative agents should ultimately be able to cooperate with people whose decision processes are impaired as part of progress toward more general human–AI systems. Our work provides one direction toward this goal.

Limitations and future work. Our evaluation relies on simulated partners because recruiting participants with relevant decision process impairments is difficult, and further evaluation is needed to establish generalization beyond the Overcooked settings considered here. Our LMM analysis further suggests that RDSA provides greater benefit when there is a stronger advantage to selecting one specialist over the other (Appendix C.6). Future work could therefore broaden the partner assumptions represented by the specialists beyond the current cooperative and uniform formulations, potentially extending the range of partner behaviors over which arbitration control provides an advantage.

AI USE STATEMENT

In this work, we used generative AI tools for two purposes: to aid and polish the writing of the manuscript, and for retrieval and discovery, such as identifying related work. We did not use these tools for other tasks, including research ideation, experimental design, or implementation. We carefully reviewed all AI-assisted work: we checked every suggested edit before adopting it, and every reference identified with AI assistance was read and verified by the authors before citation. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Xena Al-Hejji, Santina Duarte, Jose Guillermo Gomez Castro, Edgar Bermudez Contreras, and Eric Chalmers. An in-silico integration of neurodevelopmental and dopaminergic views of schizophrenia. *arXiv preprint arXiv:2510.24741*, 2025. doi: 10.48550/arXiv.2510.24741. URL https://arxiv.org/abs/2510.24741.

Bahaaeddin Attaallah, Pierre Petitet, Rhea Zambellas, Sofia Toniolo, Maria Raquel Maio, Akke Ganse-Dumrath, Sarosh R. Irani, Sanjay G. Manohar, and Masud Husain. The role of the human hippocampus in decision-making under uncertainty. *Nature Human Behaviour*, 8(7): 1366–1382, 2024. doi: 10.1038/s41562-024-01855-2. URL https://doi.org/10.1038/s41562-024-01855-2.

Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca D. Dragan. On the utility of learning about humans for human-ai coordination. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jonathan D. Cohen and David Servan-Schreiber. Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review*, 99(1):45–77, 1992. doi: 10.1037/0033-295X.99.1.45.

Dina R. Dajani and Lucina Q. Uddin. Demystifying cognitive flexibility: Implications for clinical and developmental neuroscience. *Trends in Neurosciences*, 38(9):571–578, 2015. doi: 10.1016/j.tins.2015.07.003.

Magdalena del Río, Nadescha Trudel, Gita Prabhu, Laurence T. Hunt, Michael Moutoussis, Raymond J. Dolan, and Tobias U. Hauser. Indecision and recency-weighted evidence integration in non-clinical and clinical settings. *Nature Human Behaviour*, 10(4):727–740, 2026. doi: 10.1038/s41562-025-02385-1.

Katharina Dobs, Julio Martinez, Alexander J. E. Kell, and Nancy Kanwisher. Brain-like functional specialization emerges spontaneously in deep neural networks. *Science Advances*, 8(11): eabl8913, 2022. doi: 10.1126/sciadv.abl8913.

Jade S. Duffy, Mark A. Bellgrove, Peter R. Murphy, and Redmond G. O'Connell. Disentangling sources of variability in decision-making. *Nature Reviews Neuroscience*, 26(5):247–262, 2025. doi: 10.1038/s41583-025-00916-3.

Stephen V. Faraone, Mark A. Bellgrove, Isabell Brikell, Samuele Cortese, Catharina A. Hartman, Chris Hollis, Jeffrey H. Newcorn, Alexandra Philipsen, Guilherme V. Polanczyk, Katya Rubia, Margaret H. Sibley, and Jan K. Buitelaar. Attention-deficit/hyperactivity disorder. *Nature Reviews Disease Primers*, 10:11, 2024. doi: 10.1038/s41572-024-00495-0.

Michael J. Frank. Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and nonmedicated parkinsonism. *Journal of Cognitive Neuroscience*, 17(1):51–72, 2005. doi: 10.1162/0898929052880093.

Hauke R. Heekeren, Sean Marrett, and Leslie G. Ungerleider. The neural systems that mediate human perceptual decision making. *Nature Reviews Neuroscience*, 9(6):467–479, 2008. doi: 10.1038/nrn2374.

Erin A. Heerey, Kimberly R. Bell-Warren, and James M. Gold. Decision-making impairments in the context of intact reward sensitivity in schizophrenia. *Biological Psychiatry*, 64(1):62–69, 2008. doi: 10.1016/j.biopsych.2008.02.015.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Lindsay E. Hunter, Elana A. Meer, Claire M. Gillan, Ming Hsu, and Nathaniel D. Daw. Increased and biased deliberation in social anxiety. *Nature Human Behaviour*, 6(1):146–154, 2022. doi: 10.1038/s41562-021-01180-y. URL https://doi.org/10.1038/s41562-021-01180-y.

Wendyam Eric Lionel Ilboudo and Saori C. Tanaka. An inference-based architecture for intent and affordance saturation in decision-making. *arXiv preprint arXiv:2512.23144*, 2025. doi: 10.48550/arXiv.2512.23144. URL https://arxiv.org/abs/2512.23144.

Dongjae Kim, Geon Yeong Park, John P. O'Doherty, and Sang Wan Lee. Task complexity interacts with state-space uncertainty in the arbitration between model-based and model-free learning. *Nature Communications*, 10:5738, 2019. doi: 10.1038/s41467-019-13632-1.

Dongjae Kim, Jaeseung Jeong, and Sang Wan Lee. Prefrontal solution to the bias-variance tradeoff during reinforcement learning. *Cell Reports*, 37(13):110185, 2021. doi: 10.1016/j.celrep.2021.110185.

Aounon Kumar, Alexander Levine, and Soheil Feizi. Policy smoothing for provably robust reinforcement learning. In *International Conference on Learning Representations*, 2022.

Nan M. Laird and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982. doi: 10.2307/2529876. URL https://doi.org/10.2307/2529876.

Daeyeol Lee. Decision making: From neuroscience to psychiatry. *Neuron*, 78(2):233–248, 2013. doi: 10.1016/j.neuron.2013.04.008. URL https://doi.org/10.1016/j.neuron.2013.04.008.

Sang Wan Lee, Shinsuke Shimojo, and John P. O'Doherty. Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81(3):687–699, 2014. doi: 10.1016/j.neuron.2013.11.028. URL https://doi.org/10.1016/j.neuron.2013.11.028.

Stefan Leucht, Spyridon Siafis, John J. McGrath, Patrick McGorry, Oliver D. Howes, Carol Tamminga, Richard Carr, Irene Bighelli, Johannes Schneider-Thoma, Josef Priller, and John M. Davis. Schizophrenia. *Nature Reviews Disease Primers*, 11:83, 2025. doi: 10.1038/s41572-025-00667-6.

Simin Li, Zihao Mao, Hanxiao Li, Zonglei Jing, Zhuohang Bian, Jun Guo, Li Wang, Zhuoran Han, Ruixiao Xu, Xin Yu, Chengdong Ma, Yuqing Ma, Bo An, Yaodong Yang, Weifeng Lv, and Xianglong Liu. Empirical study on robustness and resilience in cooperative multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yancheng Liang, Daphne Chen, Abhishek Gupta, Simon S. Du, and Natasha Jaques. Learning to cooperate with humans using generative agents. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1918.

Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 157–163. Morgan Kaufmann, 1994. doi: 10.1016/B978-1-55860-335-6.50027-1.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Tiago V. Maia and Michael J. Frank. From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, 14(2):154–162, 2011. doi: 10.1038/nn.2723.

Richard D. McKelvey and Thomas R. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995. doi: 10.1006/game.1995.1023.

Earl K. Miller and Jonathan D. Cohen. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24:167–202, 2001. doi: 10.1146/annurev.neuro.24.1.167.

Reuth Mirsky, Ignacio Carlucho, Arrasy Rahman, Elliot Fosong, William Macke, Mohan Sridharan, Peter Stone, and Stefano V. Albrecht. A survey of ad hoc teamwork research. In *Multi-Agent Systems: 19th European Conference, EUMAS 2022*, pp. 275–293. Springer, 2022. doi: 10.1007/978-3-031-20614-6_16.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. doi: 10.1038/nature14236.

Ahmed A. Moustafa and Mark A. Gluck. A neurocomputational model of dopamine and prefrontal–striatal interactions during multicue category learning by parkinson patients. *Journal of Cognitive Neuroscience*, 23(1):151–167, 2011. doi: 10.1162/jocn.2010.21420.

Kristopher J. Preacher, Patrick J. Curran, and Daniel J. Bauer. Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, 31(4):437–448, 2006. doi: 10.3102/10769986031004437. URL https://doi.org/10.3102/10769986031004437.

Antonio Rangel, Colin Camerer, and P. Read Montague. A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, 9(7):545–556, 2008. doi: 10.1038/nrn2357.

Edmund T. Rolls, Marco Loh, Gustavo Deco, and Georg Winterer. Computational models of schizophrenia and dopamine modulation in the prefrontal cortex. *Nature Reviews Neuroscience*, 9(9):696–709, 2008. doi: 10.1038/nrn2462.

Pedro M. Santos, João G. Ribeiro, Alberto Sardinha, and Francisco S. Melo. Ad hoc teamwork in the presence of non-stationary teammates. In *Progress in Artificial Intelligence: 20th EPIA Conference on Artificial Intelligence*, pp. 648–660. Springer, 2021. doi: 10.1007/978-3-030-86230-5_51.

S. R. Searle, F. M. Speed, and G. A. Milliken. Population marginal means in the linear model: An alternative to least squares means. *The American Statistician*, 34(4):216–221, 1980. doi: 10.1080/00031305.1980.10483031. URL https://doi.org/10.1080/00031305.1980.10483031.

Amitai Shenhav, Matthew M. Botvinick, and Jonathan D. Cohen. The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, 79(2):217–240, 2013. doi: 10.1016/j.neuron.2013.07.007.

Yeon Soon Shin and Yael Niv. Biased evaluations emerge from inferring hidden causes. *Nature Human Behaviour*, 5:1180–1189, 2021. doi: 10.1038/s41562-021-01065-0.

Herbert A. Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69 (1):99–118, 1955. doi: 10.2307/1884852.

Mason O. Smith and Wenlong Zhang. Risk-sensitive theory of mind: Coordinating with agents of unknown bias using cumulative prospect theory. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55999–56017. PMLR, 2025.

Kevin M. Spencer. The functional consequences of cortical circuit abnormalities on gamma oscillations in schizophrenia: Insights from computational modeling. *Frontiers in Human Neuroscience*, 3:33, 2009. doi: 10.3389/neuro.09.033.2009.

DJ Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, and Richard Everett. Collaborating with humans without human data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14502–14515, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/797134c3e42371bb4979a462eb2f042a-Abstract.html.

Gaurav Suri, James J. Gross, and James L. McClelland. Value-based decision making: An interactive activation perspective. *Psychological Review*, 127(2):153–185, 2020. doi: 10.1037/rev0000164.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988. doi: 10.1007/BF00115009.

Matilde M. Vaghi, Fabrice Luyckx, Akeem Sule, Naomi A. Fineberg, Trevor W. Robbins, and Benedetto De Martino. Compulsivity reveals a novel dissociation between action and confidence. *Neuron*, 96(2):348–354.e4, 2017. doi: 10.1016/j.neuron.2017.09.006. URL https://doi.org/10.1016/j.neuron.2017.09.006.

András Vargha and Harold D. Delaney. A critique and improvement of the cl common language effect size statistics of mcgraw and wong. *Journal of Educational and Behavioral Statistics*, 25 (2):101–132, 2000. doi: 10.3102/10769986025002101.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado van Hasselt, Marc Lanctot, and Nando de Freitas. Dueling network architectures for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1995–2003. PMLR, 2016.

Xue Yan, Jiaxian Guo, Xingzhou Lou, Jun Wang, Haifeng Zhang, and Yali Du. An efficient end-to-end training approach for zero-shot Human-AI coordination. In *Advances in Neural Information Processing Systems*, volume 36, pp. 2636–2658, 2023. doi: 10.52202/075280-0119. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/07a363fd2263091c2063998e0034999c-Abstract-Conference.html.

Chao Yu, Jiaxuan Gao, Weilin Liu, Botian Xu, Hao Tang, Jiaqi Yang, Yu Wang, and Yi Wu. Learning zero-shot cooperation with humans, assuming humans are biased. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TrwE8l9aJzs.

Rui Zhao, Jinming Song, Yufeng Yuan, Haifeng Hu, Yang Gao, Yi Wu, Zhongqian Sun, and Wei Yang. Maximum entropy population-based training for zero-shot Human-AI coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6145–6153, 2023. doi: 10.1609/aaai.v37i5.25758. URL https://ojs.aaai.org/index.php/AAAI/article/view/25758.

A METHODOLOGICAL DETAILS

A.1 REFERENCE PARTNER AND SPECIALIST MODELS

Each agent has six actions, so each model outputs 36 joint action values. Training schedules and optimization settings are provided in Section B.2. Before applying the quantal response procedure or the iterative deterministic $\arg\max$ updates described in Section 4.1, we mask infeasible joint actions. To make this masking explicit, let $\mathcal{F}(s, a_e) \subseteq \mathcal{A}_p$ denote the set of feasible partner actions in state s given ego action a_e. The feasibility masked joint action values introduced in Section 4.1 are then defined as

$$[G_m(s)]_{a_e,a_p} = \begin{cases} Q_m(s, a_e, a_p), & a_p \in \mathcal{F}(s, a_e), \\ -10^9, & a_p \notin \mathcal{F}(s, a_e), \end{cases} \qquad m \in \{\mathrm{Ref}, \mathrm{PR}, \mathrm{SR}\}. \tag{9}$$

Rows correspond to ego actions and columns to partner actions. For deterministic $\arg\max$ updates, ties are broken by selecting the action with the lowest index.

Reference partner model (Ref). Ref uses a plain joint Q DQN. The network receives a 28-channel representation of the current state, with spatial dimensions determined by the layout. Agent specific channels are ordered from the acting agent's perspective, allowing the same network to serve either seat. The CNN encoder contains two convolutional layers with 32 and 64 output channels, respectively. Both layers use 3×3 kernels, stride 1, padding 1, and LeakyReLU activations. No pooling is applied. The encoder output is flattened and passed to an MLP Q network containing seven linear layers: six hidden layers of 128 units with LeakyReLU activations, followed by a linear output layer of 36 units. The CNN and MLP architectures are identical across standard and risky layouts. The quantal response procedure in Section 4.1 uses $\lambda = 10$ and recursion depth $k = 8$. The implementation uses a fixed recursion depth rather than a fixed-point convergence criterion.

Partner-Reliant Specialist (PR-Specialist). The PR-Specialist uses the same CNN architecture as Ref, followed by separate state value and action advantage MLPs. The two MLPs share the CNN encoder and receive its flattened output. Each MLP contains seven linear layers: six hidden layers of 128 units with LeakyReLU activations, followed by a linear output layer. The state value MLP outputs one scalar, whereas the action advantage MLP outputs 36 advantage values. Their outputs are combined according to Eq. 1. The iterative deterministic $\arg\max$ update procedure in Section 4.1 starts from uniform action distributions for both agents. At each iteration, the ego agent first performs an $\arg\max$ update under the current partner distribution, followed by the partner under the updated ego distribution. Each update produces a one hot action distribution, and the resulting pair is carried forward to the next iteration. The iteration terminates when $\max_{i \in \{e,p\}} \|\pi_i^{(j+1)} - \pi_i^{(j)}\|_\infty < 10^{-4}$ or after 50 iterations.

Self-Reliant Specialist (SR-Specialist). The SR-Specialist uses the same CNN and dueling MLP architecture as the PR-Specialist, but it is trained separately. Under the uniform partner assumption in Section 4.1, the partner action distribution is fixed to $\mathrm{Unif}(\mathcal{A}_p)$. This is equivalent to setting the partner rationality parameter to $\lambda_p = 0$ in the quantal response parameterization. The SR-Specialist then performs a deterministic $\arg\max$ update under the uniform partner distribution using $G_{\mathrm{SR}}(s)$.

A.2 RELIABILITY ESTIMATION AND ARBITRATION CONTROL DYNAMICS

Ego action value standardization. The PR- and SR-Specialist ego action values defined in Section 4.2 are standardized separately over ego actions for which at least one feasible partner action exists:

$$\tilde{q}_m(a_e) = \frac{q_m(a_e) - \mu_m}{\sigma_m + \varepsilon}, \qquad m \in \{\mathrm{PR}, \mathrm{SR}\}, \quad |\mathcal{F}(s, a_e)| > 0, \tag{10}$$

where μ_m and σ_m are the mean and standard deviation, respectively, of $q_m(a_e)$ over ego actions satisfying $|\mathcal{F}(s, a_e)| > 0$. We set $\varepsilon = 10^{-6}$ for numerical stability. If all feasible ego action values of a specialist are identical, their standardized values are zero. Ego actions with no feasible partner action are assigned -10^9 before the final $\arg\max$ action selection.

Bounded rational reference distribution. At state s_t, we apply the same quantal response procedure used by Ref to the feasibility masked PR game $G_{\mathrm{PR}}(s_t)$, with $\lambda = 10$ and recursion depth $k = 8$. After quantal response updates, we denote the partner action distribution over $\mathcal{F}(a_e^t)$ as π_t^{BR}. Because the quantal response procedure is applied to the feasibility masked PR game, infeasible partner actions receive zero probability. Unlike Ref, the PR-Specialist was trained using alternating deterministic $\arg\max$ updates under the cooperative partner assumption. Thus, the same quantal response procedure is applied to joint action values trained under a different response rule, so π_t^{BR} differs from Ref's partner action distribution.

Evidence computation. For the PR-Specialist, define $v_t(a_p) = Q_{\mathrm{PR}}(s_t, a_e^t, a_p)$. The probability weighted percentile is computed as

$$r_t^{\mathrm{PR}} = \sum_{a_p \in \mathcal{F}(a_e^t)} \pi_t^{\mathrm{BR}}(a_p) \left[\mathbf{1}\{v_t(a_p) < v_t(a_p^t)\} + \frac{1}{2}\mathbf{1}\{v_t(a_p) = v_t(a_p^t)\} \right]. \tag{11}$$

Value differences within 10^{-9} are treated as ties. We define the PR evidence signal as

$$\delta_t^{\mathrm{PR}} = \frac{1}{2} - r_t^{\mathrm{PR}}. \tag{12}$$

We retain δ_t^{PR} without smoothing to preserve the observed action's relative value rank at each state. The zero evidence threshold corresponds to $r_t^{\mathrm{PR}} = 1/2$; values below this percentile yield $\delta_t^{\mathrm{PR}} > 0$ and are treated as evidence against the cooperative partner assumption. If $|\mathcal{F}(a_e^t)| \leq 1$, we set $\delta_t^{\mathrm{PR}} = 0$, since the observed partner action does not provide comparative evidence.

For the SR-Specialist, evidence compares the probability of the observed action under the cooperative reference distribution with that under uniform partner behavior. We define the raw probability difference as

$$e_t^{\mathrm{SR}} = \pi_t^{\mathrm{BR}}(a_p^t) - \frac{1}{|\mathcal{F}(a_e^t)|}. \tag{13}$$

Because this signal uses the probability of a single sampled action, step wise sampling fluctuations can affect evidence categorization even when the partner policy is unchanged. We therefore smooth it using an exponential moving average:

$$\delta_t^{\mathrm{SR}} = (1 - \kappa)\delta_{t-1}^{\mathrm{SR}} + \kappa e_t^{\mathrm{SR}}, \qquad \kappa = 0.1. \tag{14}$$

We initialize $\delta_{-1}^{\mathrm{SR}} = 0$. Because a small positive probability difference does not necessarily indicate purposeful partner behavior, we apply a positive threshold of $\tau_{\mathrm{SR}} = 0.1$ to the smoothed signal instead of using zero as the decision threshold. Only $\delta_t^{\mathrm{SR}} > \tau_{\mathrm{SR}}$ is treated as evidence against the uniform partner assumption. Thus, smoothing reduces temporal fluctuations, while the positive threshold tolerates small positive deviations in the smoothed signal.

A diagnostic experiment in Coordination Ring compared Ref with a near uniform partner using $\lambda = 0.001$. We applied the same threshold, $\tau_{\mathrm{SR}} = 0.1$, to the raw SR signal and to the EMA smoothed signal. Without smoothing, 34.2% of Ref steps and 78.8% of near uniform partner steps were assigned to the category supporting the uniform partner assumption. With EMA smoothing using $\kappa = 0.1$, these proportions changed to 4.7% and 97.0%, respectively, increasing the separation between the two partners from 44.6 to 92.3 percentage points. This diagnostic supports the use of EMA smoothing for the SR signal.

Bayesian reliability estimation. We retain the posterior mean-to-variance reliability construction of Lee et al. (2014), but adapt the evidence representation to two categories for each specialist. Category 0 indicates that the corresponding specialist assumption is consistent with the observed partner behavior, whereas category 1 indicates evidence against that assumption. The categories are assigned as

$$c_t^{\mathrm{PR}} = \begin{cases} 0, & \delta_t^{\mathrm{PR}} \leq 0, \\ 1, & \delta_t^{\mathrm{PR}} > 0, \end{cases} \qquad c_t^{\mathrm{SR}} = \begin{cases} 0, & \delta_t^{\mathrm{SR}} \leq \tau_{\mathrm{SR}}, \\ 1, & \delta_t^{\mathrm{SR}} > \tau_{\mathrm{SR}}. \end{cases} \tag{15}$$

Each specialist maintains a FIFO window containing its most recent $T_{\mathrm{mem}} = 40$ categories. When the window is full, adding a new category removes the oldest one. Let $n_{m,c}$ denote the number

of observations assigned to category $c \in \{0, 1\}$ for specialist $m \in \{\mathrm{PR}, \mathrm{SR}\}$, and let $N_m = n_{m,0} + n_{m,1} \leq T_{\mathrm{mem}}$. We use a symmetric Dirichlet prior with unit concentration per category, following the reference implementation of reliability based arbitration control:

$$\boldsymbol{\theta}_m \sim \mathrm{Dirichlet}(1, 1). \tag{16}$$

This yields the posterior

$$\boldsymbol{\theta}_m \mid \mathcal{D}_m \sim \mathrm{Dirichlet}\left(1 + n_{m,0}, 1 + n_{m,1}\right). \tag{17}$$

For category c, the posterior mean and variance are

$$\begin{aligned} \mu_{m,c} &= \frac{1 + n_{m,c}}{2 + N_m}, \\ \nu_{m,c} &= \frac{(1 + n_{m,c})(1 + N_m - n_{m,c})}{(2 + N_m)^2(3 + N_m)}. \end{aligned} \tag{18}$$

Following the reliability formulation of Lee et al. (2014), the reliability of specialist m is

$$\chi_m = \frac{\mu_{m,0}/\nu_{m,0}}{\mu_{m,0}/\nu_{m,0} + \mu_{m,1}/\nu_{m,1}}. \tag{19}$$

Thus, χ_m increases as the accumulated evidence more consistently supports the specialist's assumption. With an empty history, the symmetric prior yields $\chi_m = 0.5$.

Transition dynamics. The two reliabilities determine the transition rates between the specialists. Let α_t denote the transition rate from the SR-Specialist toward the PR-Specialist, and β_t the transition rate in the opposite direction. We compute

$$\begin{aligned} \alpha_t &= \frac{1}{2} \frac{A_\alpha}{1 + \exp(B_\alpha \chi_{\mathrm{SR}})}, \\ \beta_t &= \frac{1}{2} \frac{A_\beta}{1 + \exp(B_\beta \chi_{\mathrm{PR}})}, \end{aligned} \tag{20}$$

where

$$B_\alpha = \log\left(\frac{A_\alpha}{\delta_\alpha} - 1\right), \qquad B_\beta = \log\left(\frac{A_\beta}{\delta_\beta} - 1\right). \tag{21}$$

We use

$$A_\alpha = 12, \qquad A_\beta = 9, \qquad \delta_\alpha = 0.05, \qquad \delta_\beta = 0.01, \tag{22}$$

which give $B_\alpha = 5.477$ and $B_\beta = 6.801$. Because each transition rate decreases with the reliability of the specialist from which control would move away, low χ_{SR} increases the shift toward the PR-Specialist, whereas low χ_{PR} increases the shift toward the SR-Specialist. For fixed α_t and β_t over one timestep, the continuous two-state transition equation has equilibrium

$$p_t^* = \frac{\alpha_t}{\alpha_t + \beta_t}. \tag{23}$$

We use its exact unit-time update:

$$p_{t+1} = p_t^* + (p_t - p_t^*) \exp\left[-(\alpha_t + \beta_t)\right]. \tag{24}$$

We initialize $p_0 = 0.8$ to assign greater weight to the PR-Specialist, reflecting our initial assumption of cooperative partner behavior.

Update order during interaction. At the beginning of each episode, the two reliability windows are emptied, so both reliability estimates start at 0.5. The SR smoothing state $\delta_{-1}^{\mathrm{SR}}$ is reset to zero, and the arbitration weight is initialized to $p_0 = 0.8$. At timestep t, the ego and partner select their actions simultaneously. The ego action a_e^t is selected using p_t, which incorporates evidence observed only through timestep $t-1$. After the joint action (a_e^t, a_p^t) is executed, RDSA observes the partner action a_p^t, computes π_t^{BR} and the two evidence signals, assigns their corresponding categories, and inserts them into the two FIFO windows. The reliability estimates are then recomputed, and Eq. 24 updates the arbitration weight to p_{t+1}. The PR- and SR-Specialist network parameters remain fixed throughout interaction; only the two FIFO windows, the SR smoothing state, the reliability estimates, and the arbitration weight are updated.

Algorithm 1 RDSA update during interaction

Require: Fixed PR- and SR-Specialist network parameters
Empty the two FIFO windows
$\chi_{\mathrm{PR}}, \chi_{\mathrm{SR}} \leftarrow 0.5$
$\delta_{-1}^{\mathrm{SR}} \leftarrow 0$, $p_0 \leftarrow 0.8$
for $t = 0, \ldots, T-1$ **do**
    Ego and partner select actions simultaneously; ego selects a_e^t using p_t
    Execute joint action (a_e^t, a_p^t) and observe a_p^t
    Compute π_t^{BR} and the PR and SR evidence signals
    Assign the corresponding evidence categories
    Insert the categories into the two FIFO windows
    Recompute χ_{PR} and χ_{SR}
    Compute α_t and β_t
    $p_t^* \leftarrow \dfrac{\alpha_t}{\alpha_t + \beta_t}$
    $p_{t+1} \leftarrow p_t^* + (p_t - p_t^*)\exp[-(\alpha_t + \beta_t)]$
end for

A.3 Decision process impairment modeling

Perception impairment. The observation channels were grouped according to the task relevant environmental elements they encode. The standard layouts use three eligible groups corresponding to pots, onions, and dishes, whereas the risky layouts additionally include risk and subgoal related groups. Channels encoding the positions and orientations of both agents were excluded from perturbation to preserve localization information. At the beginning of each episode, one eligible channel group g was sampled uniformly and remained fixed throughout the episode. Let m_g denote the corresponding binary channel mask. At each timestep, independently sampled Gaussian noise was applied only to the selected channels:

$$\tilde{o}_t = o_t + m_g \odot \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}\left(0, \sigma_{\mathrm{perc}}^2 I\right), \tag{25}$$

where o_t and $\tilde{o}_t$ denote the clean and perturbed partner observations at timestep t, respectively, and $\odot$ denotes element wise multiplication. The parameter σ_{perc} controls the noise magnitude. The perturbation was applied only to the partner observation; the ego agent received the corresponding clean observation.

Valuation impairment. Valuation impairment was applied to the MLP of Ref's joint Q network. For each MLP weight matrix W, we randomly selected 10% of the strictly positive weights, i.e., entries satisfying $W_{ij} > 0$, and set them to zero. Bias parameters were excluded from both pruning and positive weight regularization. After pruning, the network was fine-tuned through self-play on the original task to allow compensation under the modified weight structure. Fine-tuning used the original temporal difference objective with an additional penalty on positive MLP weights:

$$\mathcal{L}_{\mathrm{ft}}(\vartheta) = \mathbb{E}_t\left[\left(Q_\vartheta(s_t, a_e^t, a_p^t) - y_t\right)^2\right] + \frac{\lambda_{\mathrm{pos}}}{2} \sum_{W \in \Theta_{\mathrm{MLP}}} \|\mathrm{ReLU}(W)\|_F^2, \tag{26}$$

where ϑ denotes the Q network parameters, y_t denotes the temporal difference target, and Θ_{MLP} denotes the set of MLP weight matrices. Here, $\mathrm{ReLU}(W)$ retains only the positive entries of W, so the second term penalizes the magnitude of positive MLP weights. We set $\lambda_{\mathrm{pos}} = 10^{-3}$. After fine-tuning, the network parameters were fixed. During inference, independently sampled Gaussian noise was added to the MLP input at each timestep:

$$\tilde{h}_t = h_t + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}\left(0, \sigma_{\mathrm{val}}^2 I\right), \tag{27}$$

where h_t and $\tilde{h}_t$ denote the clean and perturbed MLP inputs at timestep t, respectively. The pruning fraction and λ_{pos} were fixed across all valuation impaired partners, while σ_{val} determined impairment severity.

Table 4: **Fixed-severity impairment parameters.** Values used for mild, moderate, and severe impairment in the behavioral analyses. Larger σ_{perc} and σ_{val}, and smaller λ, correspond to greater impairment.

Impairment parameter	Mild	Moderate	Severe
σ_{val}	0.2	0.6	1.0
λ	5.0	1.0	0.25
σ_{perc} (standard)	0.4	1.2	2.0
σ_{perc} (risky)	0.3	0.9	1.5

Action selection impairment. Action selection impairment was applied by varying only the rationality parameter λ in Ref's quantal response procedure. At recursion step ℓ, the action distribution of agent i is computed using the seat aware joint action value

$$\bar{G}_i(s, a_i, a_{-i}) = \begin{cases} G_{\text{Ref}}(s, a_i, a_{-i}), & i = e, \\ G_{\text{Ref}}(s, a_{-i}, a_i), & i = p. \end{cases} \tag{28}$$

The corresponding action distribution is

$$\pi_i^{(\ell)}(a_i \mid s) = \frac{\exp\Big(\lambda\, \mathbb{E}_{a_{-i}\sim\pi_{-i}^{(\ell-1)}}\left[\bar{G}_i(s, a_i, a_{-i})\right]\Big)}{\sum_{a_i'\in\mathcal{A}_i} \exp\Big(\lambda\, \mathbb{E}_{a_{-i}\sim\pi_{-i}^{(\ell-1)}}\left[\bar{G}_i(s, a_i', a_{-i})\right]\Big)}, \qquad \pi_i^{(0)} = \text{Unif}(\mathcal{A}_i). \tag{29}$$

The recursion depth remained fixed at $k = 8$. Network parameters, observation processing, and valuation computation were unchanged. Decreasing λ reduces sensitivity to differences in expected joint action values, producing a more stochastic action distribution and therefore stronger action selection impairment.

Impairment severity. We parameterize impairment severity by σ_{perc} for perception, σ_{val} for valuation, and λ for action selection. Larger σ_{perc} and σ_{val} correspond to more severe impairment, whereas smaller λ corresponds to more severe action selection impairment. For the main evaluation, the partner's impairment parameters were independently sampled at the beginning of each episode and then held fixed throughout that episode. We sampled $\sigma_{\text{val}} \sim U(0, 1)$ for valuation impairment and $\lambda \sim \text{LogU}(0.25, 10)$ for action selection impairment. For perception impairment, both the affected channel group and σ_{perc} were sampled at the beginning of the episode. The channel group was sampled uniformly from the eligible groups for the layout, and $\sigma_{\text{perc}} \sim U(0, 2)$ in the standard layouts and $\sigma_{\text{perc}} \sim U(0, 1.5)$ in the risky layouts. These episode level samples were generated from a random seed determined solely by the episode index, so all evaluated methods encountered the same sequence of impairment conditions. Within an episode, the sampled impairment parameters remained fixed. For perception and valuation impairment, Gaussian noise was independently resampled at each timestep, and the partner action was sampled from the k level quantal response distribution. For each impairment type, we swept the impairment parameter in self-play and selected the severe endpoint near the point at which successful deliveries fell to approximately zero. For valuation impairment, self-play at $\sigma_{\text{val}} = 1$ produced mean delivery counts of 0.4, 0.0, and 0.0 in Cramped Room, Coordination Ring, and Counter Circuit, respectively. For action selection impairment, the corresponding counts at $\lambda = 0.25$ were 0.3, 0.0, and 0.0. For perception impairment, collapse occurred near $\sigma_{\text{perc}} = 2$ on the standard layouts and near $\sigma_{\text{perc}} = 1.5$ on the risky layouts. We used log-uniform sampling for λ because self-play performance changed approximately linearly with $\log \lambda$; uniform sampling in λ would therefore over-sample the near-intact regime. For the behavioral analyses requiring fixed impairment levels, we used three severity levels as summarized in Table 4. The mild, moderate, and severe values correspond approximately to 20%, 60%, and 100% of each impairment range in the direction of increasing severity, using a logarithmic scale for λ.

A.4 Arbitration hyperparameter search

Search setup. We searched hyperparameters using anchor based tracking error. The search covered the SR evidence threshold τ_{SR}, the SR smoothing coefficient κ, the transition-rate scales (A_α, A_β), the SR-to-PR rate boundary parameter δ_α, and the reliability memory length T_{mem}. We

retained $\delta_\beta = 0.01$, $p_0 = 0.8$, and the symmetric $\mathrm{Dirichlet}(1, 1)$ prior from the reference implementation without tuning. The PR evidence threshold remained zero, as defined in Section A.2. The ego agent used RDSA with two anchor partners. The PR anchor used the PR-Specialist as the partner in seat 1, with target weight $y_t = 1$. The SR anchor sampled uniformly from the feasible partner actions with target weight $y_t = 0$. We conducted the search in Coordination Ring and Risky Multipath, with slip probability $p_{\mathrm{slip}} = 0.4$ in Risky Multipath. Each episode lasted $T = 400$ timesteps.

Search grid and procedure. Because p_t denotes the relative weight assigned to the PR-Specialist, we set its target to $y_t = 1$ under a PR anchor and to $y_t = 0$ under an SR anchor. We then measured how closely the arbitration weight tracked the target associated with the active anchor. Under a PR anchor, the partner behavior supports the cooperative partner assumption, so the target weight is $y_t = 1$. Under an SR anchor, the behavior supports the uniform partner assumption, so the target weight is $y_t = 0$. We therefore used the deviation of p_t from the active anchor's target weight to evaluate how well the arbitration control tracked the behavior represented by each anchor.

To explore a broad range of settings without evaluating every joint combination, we searched the grid parameters and memory length in separate stages. With $T_{\mathrm{mem}} = 20$ fixed, we evaluated all 24 combinations of $\tau_{\mathrm{SR}} \in \{0.05, 0.1\}$, $\kappa \in \{0.1, 0.3\}$, $(A_\alpha, A_\beta) \in \{(4, 3), (8, 6), (12, 9)\}$, and $\delta_\alpha \in \{0.05, 0.1\}$. Each configuration was evaluated with switching interval $S = 400$, so the anchor remained fixed throughout each episode to match the evaluation setting. We ran three episodes per configuration and layout using the PR, SR, and PR anchors in that order. For each episode, tracking error was defined as the mean absolute difference between the arbitration weight p_t and the active anchor's target weight y_t:

$$\mathcal{T} = \frac{1}{T} \sum_{t=0}^{T-1} |p_t - y_t|. \tag{30}$$

For each configuration, we first averaged the three episode level tracking errors within each layout and then averaged the two layout means.

Parameter selection. Table 5 reports the search results at $\kappa = 0.1$ and $S = 400$. We omit the results for $\kappa = 0.3$ because $\kappa = 0.1$ yielded lower tracking error in all 12 matched configurations at $S = 400$. Among the reported configurations, $\tau_{\mathrm{SR}} = 0.1$, $(A_\alpha, A_\beta) = (12, 9)$, and $\delta_\alpha = 0.05$ achieved the lowest tracking error of 0.0465. For this configuration, the mean p_t was 0.993 under the PR anchor and 0.125 under the SR anchor, close to the respective target weights of 1 and 0. The anchor weight difference of 0.868 was also the largest among the reported configurations. We selected these values together with $\kappa = 0.1$ and held them fixed for the subsequent memory length search.

Reliability memory search. We evaluated $T_{\mathrm{mem}} \in \{6, 12, 20, 40, 80\}$ at $S = 400$, with $\kappa = 0.1$ and all other parameters fixed. Each configuration used five seed blocks, with three episodes per block and layout. For each seed block, we averaged the episode level tracking errors within each layout and then averaged the two layout means. As shown in Table 6, mean tracking error decreased as the memory length increased. The adopted memory length $T_{\mathrm{mem}} = 40$ achieved a mean tracking error of 0.044, compared with the lowest mean of 0.040 at $T_{\mathrm{mem}} = 80$. We adopted $T_{\mathrm{mem}} = 40$ because its tracking error was close to the lowest value while requiring less storage for the evidence history.

Fixed settings and final configuration. The PR evidence threshold was fixed at zero, corresponding to the midpoint of the probability weighted percentile. Each specialist used two evidence categories. The values $\delta_\beta = 0.01$ and $p_0 = 0.8$, together with the symmetric $\mathrm{Dirichlet}(1, 1)$ prior, were inherited from the reference implementation used for reliability based arbitration control. The initial weight $p_0 = 0.8$ assigned greater influence to the PR-Specialist, reflecting an initial assumption of cooperative partner behavior. The final configuration used $\tau_{\mathrm{SR}} = 0.1$, $\kappa = 0.1$, $T_{\mathrm{mem}} = 40$, $(A_\alpha, A_\beta) = (12, 9)$, $\delta_\alpha = 0.05$, $\delta_\beta = 0.01$, and $p_0 = 0.8$. All reported benchmark results with 1,000 episodes per condition used this configuration.

Table 5: **Hyperparameter search results.** τ_{SR} denotes the SR evidence threshold; A_α and A_β denote the SR-to-PR and PR-to-SR transition rate scales, respectively; and δ_α denotes the SR-to-PR rate boundary parameter. $\mathcal{T}$ denotes tracking error. The PR and SR anchor columns report the mean PR-Specialist weight p_t under the corresponding anchor, and the anchor weight difference is computed as the PR anchor mean minus the SR anchor mean before rounding. Results use $\kappa = 0.1$, $S = 400$, $T_{\mathrm{mem}} = 20$, and $\delta_\beta = 0.01$. Rows are ordered by tracking error, and bold indicates the lowest value.

τ_{SR}	(A_α, A_β)	δ_α	$\mathcal{T} \downarrow$	Mean p_t PR anchor	Mean p_t SR anchor	Anchor weight difference
0.1	(12,9)	0.05	**0.0465**	0.993	0.125	0.868
0.1	(8,6)	0.05	0.0559	0.991	0.150	0.841
0.1	(12,9)	0.10	0.0648	0.993	0.181	0.812
0.1	(8,6)	0.10	0.0733	0.992	0.203	0.788
0.1	(4,3)	0.05	0.0737	0.986	0.194	0.793
0.05	(8,6)	0.05	0.0766	0.993	0.215	0.778
0.05	(12,9)	0.05	0.0781	0.994	0.223	0.772
0.05	(4,3)	0.05	0.0917	0.988	0.251	0.738
0.1	(4,3)	0.10	0.0983	0.986	0.268	0.718
0.05	(8,6)	0.10	0.0998	0.993	0.285	0.708
0.05	(12,9)	0.10	0.1036	0.994	0.300	0.694
0.05	(4,3)	0.10	0.1258	0.988	0.354	0.634

Table 6: **Reliability memory search.** T_{mem} denotes the reliability memory length, and $\mathcal{T}$ denotes tracking error. Entries report the mean $\pm$ 95% confidence interval half-width across five seed-block means. Results use $S = 400$, $\kappa = 0.1$, $\tau_{\mathrm{SR}} = 0.1$, $(A_\alpha, A_\beta) = (12, 9)$, $\delta_\alpha = 0.05$, and $\delta_\beta = 0.01$. Bold indicates the lowest mean, and "adopted" marks the memory length used for the benchmark results.

T_{mem}	Tracking error $\mathcal{T} \downarrow$
6	0.068 ± 0.010
12	0.055 ± 0.004
20	0.050 ± 0.007
40 (adopted)	0.044 ± 0.006
80	$\mathbf{0.040} \pm 0.004$

B Experiment Details

B.1 Environment details

Overcooked is a two player cooperative gridworld in which both agents work together to prepare and deliver soups for a shared reward (Carroll et al., 2019). To complete a soup, agents collect onions from the dispenser, place three onions into a pot, collect the cooked soup using a dish, and deliver it to the serving location. At each timestep, each agent selects one of six actions: movement in one of four directions, stay, or interact (Carroll et al., 2019; Yu et al., 2023). The interact action is used to pick up or place objects, interact with pots, and deliver completed soups. We used five layouts that impose different coordination demands.

Cramped Room. Cramped Room contains a small and highly constrained workspace in which the agents frequently share movement paths. Effective cooperation therefore requires close movement coordination to avoid blocking one another while collecting ingredients, dishes, and completed soups (Carroll et al., 2019).

Coordination Ring. Coordination Ring organizes the traversable space around a central obstacle. Agents must coordinate their movement direction when traveling between task relevant locations, since incompatible movement choices can cause the agents to block one another (Yu et al., 2023). This layout therefore emphasizes coordinated navigation in addition to task execution.

Counter Circuit. Counter Circuit allows agents to complete the task by navigating around the circuit, but also provides an efficient coordinated strategy in which objects are passed through the middle counter (Yu et al., 2023). For example, one agent can place onions on the central counter while the other retrieves them and transfers them to the pots. The layout therefore requires coordination in both movement and subtask handoffs.

Risky Dualpath. Risky Dualpath contains a central obstacle that creates two alternative routes between task relevant locations. A blue puddle introduces a slip hazard along one route, while the yellow subgoal is located along the alternative route. Entering a puddle may cause an agent to lose the object it is carrying, whereas completing the subgoal removes the slip hazards for 60 timesteps. To prevent accidental activation, the subgoal is completed only when an agent performs the interact action at its location.

Risky Multipath. Risky Multipath contains a more complex central obstacle that provides multiple routes between task relevant locations. Several blue puddles are distributed across these routes, creating different levels of exposure to slip hazards depending on the path taken. As in Risky Dualpath, completing the yellow subgoal removes the hazards for 60 timesteps and requires an explicit interact action at the subgoal location to prevent accidental activation.

Meaning of Subgoal The combination of puddles and the subgoal creates situations in which an agent must evaluate whether its partner is likely to provide the needed support. Depending on the partner's behavior, the agent must choose whether to rely on the partner to activate the subgoal and proceed with soup delivery, or to activate the subgoal itself before continuing the cooking task. These risky layouts therefore provide a setting for evaluating how effectively an agent cooperates with partners affected by decision process impairment. Effective cooperation in this setting involves adapting the agent's cooperative actions based on both the partner's observed behavior and its inferred ability to provide support in the current situation.

B.2 Training and implementation details

We followed the training setup of Smith & Zhang (2025), including its backward curriculum, reward shaping scheme, and one step next state prospection, as a common training basis for Ref and the two specialists. This setup was originally developed for deep joint Q learning in Risky Overcooked and provides a consistent training procedure across all five layouts.

Ref was trained with a joint DQN, while the PR- and SR-Specialists were trained with joint dueling DQNs. Because Ref is intended to represent an unimpaired, risk neutral base partner, we implemented the neutral setting defined in Smith & Zhang (2025). Dueling architectures are useful in states where several actions have similar values and distinguishing individual action values is therefore less important (Wang et al., 2016). Such states occur frequently in Overcooked; for example, while agents are navigating before reaching an interactable object, multiple joint actions can have similar effects on immediate task progress. We therefore used a dueling architecture for the specialists, decomposing each joint action value into a state value and a relative joint action advantage.

Unlike Smith & Zhang (2025), our objective is not to model risk sensitive behavior. We therefore set all cumulative prospect theory (CPT) parameters to their risk neutral values, for which the CPT transformation becomes the identity. For a joint action $a = (a_e, a_p)$, let $\{(s'_i, p_i)\}_i$ denote the possible one step next states and their transition probabilities, where $p_i = P(s'_i \mid s, a)$. For model $m \in \{\mathrm{Ref}, \mathrm{PR}, \mathrm{SR}\}$, the corresponding TD-target prospects are

$$\tau_i^m = R(s,a) + \gamma V_m(s'_i), \qquad p_i = P(s'_i \mid s, a), \tag{31}$$

where $V_m(s'_i)$ denotes the expected joint action value at the prospective next state under model m.

Table 7: **Optimization and replay settings for Ref and the two specialists.**

Setting	Value
Optimizer	AdamW with AMSGrad
Learning rate	10^{-3}
Weight decay	0
Gradient-norm clipping	0.75
Discount factor γ	0.95
Target-update coefficient τ	5×10^{-3}
Replay sampling	Uniform
Replay capacity	10^5
Minibatch size	256
Update interval	4 environment timesteps
Replay warm-up	20,000 transitions
Training episode horizon	200 timesteps

Under the risk neutral CPT parameters, the prospect-value transformation reduces to

$$\rho(\{\tau_i^m, p_i\}_i) = \sum_i p_i \tau_i^m, \tag{32}$$

and the resulting one step target becomes

$$y_m(s,a) = R(s,a) + \gamma \sum_i P(s_i' \mid s,a) V_m(s_i'). \tag{33}$$

Thus, the risk sensitive transformation in the original training procedure is removed, yielding a risk neutral expected TD backup analogous in form to Expected SARSA.

Optimization and replay. We trained Ref, the PR-Specialist, and the SR-Specialist separately through self-play. Within each training run, a single joint Q network controlled both agents. Training episodes lasted 200 timesteps, whereas evaluation episodes lasted 400 timesteps. Training minimized the mean squared temporal difference error. The target value used the action distributions computed from the target network using the corresponding action distribution procedure. We used AdamW with AMSGrad and sampled transitions uniformly from the replay buffer. Updates were performed every four environment timesteps after a warm-up of 20,000 stored transitions. The target-network parameters were updated as

$$\bar{\vartheta} \leftarrow \tau \vartheta + (1-\tau)\bar{\vartheta}, \tag{34}$$

where ϑ and $\bar{\vartheta}$ denote the online- and target-network parameters, respectively, and τ is the target-update coefficient. Table 7 summarizes the optimization and replay settings.

Exploration schedule. The exploration coefficient ϵ decreased from 1.0 to 0.1 using an exponential schedule and remained at 0.1 thereafter. Episodes revisiting completed curriculum stages used half of the current exploration coefficient.

Reward shaping. To improve learning efficiency, the training reward combined the environment reward with a scaled shaping reward:

$$r_{i,t}^{\text{train}} = r_t^{\text{env}} + \phi_t r_{i,t}^{\text{shape}}, \qquad i \in \{e, p\}, \tag{35}$$

where r_t^{env} denotes the environment reward, $r_{i,t}^{\text{shape}}$ denotes the shaping reward for agent i, and ϕ_t is the shaping coefficient. The environment reward included a shared reward of $+20$ per delivered soup and a time cost of -0.3 per timestep. The shaping reward included three event based terms: placing an onion in a pot ($+3$), picking up a dish from a dispenser or counter ($+1$), and collecting cooked soup with a dish ($+3$). These rewards were credited only to the acting agent. A distance progress term was computed for each agent at every timestep in both standard and risky layouts. The subgoal press bonus was $+2$. Table 8 summarizes the reward components.

The distance-progress term used an undiscounted potential difference:

$$F_{i,t} = 0.3\left[\Phi_i(s_{t+1}) - \Phi_i(s_t)\right], \tag{36}$$

Table 8: **Environment reward and shaping terms.** Rows list the environment reward components and reward shaping terms used during training. Shaping values are reported before multiplication by ϕ_t. $F_{i,t}$ denotes the distance progress term defined in Eq. 36, and the subgoal press bonus applies only in layouts containing a subgoal tile.

Component	Event or term	Value
Environment	Deliver soup	$+20$
Environment	Each timestep	-0.3
Shaping	Place an onion in a pot	$+3$
Shaping	Pick up a dish	$+1$
Shaping	Collect cooked soup with a dish	$+3$
Shaping	Distance progress	$F_{i,t}$
Shaping	Press a subgoal tile	$+2$

Table 9: **Backward curriculum and stage specific initialization.** Rows list the nine curriculum stages, their corresponding subtasks, and the initialization used at each stage, following Smith & Zhang (2025). Onion numbers indicate the order in which ingredients are added to a pot.

Stage	Subtask	Initialization
1	Deliver soup	Holding cooked soup
2	Pick up cooked soup	Holding a dish
3	Pick up a dish	Empty-handed; soup cooking
4	Place the third onion in a pot	Holding the third onion
5	Pick up the third onion	Empty-handed; two onions in a pot
6	Place the second onion in a pot	Holding the second onion
7	Pick up the second onion	Empty-handed; one onion in a pot
8	Place the first onion in a pot	Holding the first onion
9	Full task	Standard initial state

where $\Phi_i(s)$ is the negative walking distance from agent i to its current target. The target depended on the object carried by the agent. An empty-handed agent targeted the dispenser of the next item required by the team. An agent holding an onion targeted a pot that could accept another onion. An agent holding a dish targeted a pot containing cooking or cooked soup. An agent holding soup targeted a serving location. An empty-handed agent with no required target was assigned a potential of zero. For agents carrying an object, slip hazards were treated as blocked when computing the walking distance. All shaping terms were multiplied by ϕ_t. During each curriculum stage, the coefficient decreased linearly from 1.0 to 0.1 and remained at 0.1 thereafter.

Curriculum learning. The backward curriculum started with soup delivery and progressively introduced earlier subtasks. Each stage initialized the agents in an intermediate task state so that they could practice the remaining steps toward delivery. Table 9 lists the eight subtask stages and the final full task stage. The full task stage used the standard initial state. A stage advanced when the mean delivery count over the most recent 10 episodes met the stage-specific progression criterion and at least 100 episodes had been completed at that stage. The recent delivery count history was cleared after each stage transition to prevent performance from the preceding stage from triggering successive transitions. Completed stages remained in the sampling pool with geometrically decaying sampling weights and a decay factor of 0.5. These stages were revisited with the reduced exploration coefficient described above to maintain previously learned subtask skills.

B.3 Details of behavioral differences beyond skill level variation

Probe-state construction. For each layout, the fixed probe state contained 308 states: 25 states from each of the eight backward curriculum stages and 108 states from Ref rollouts. Stages 1–8 follow the order in Table 9. The rollout states formed a separate group and were not a curriculum stage. All partners within each layout were evaluated on the same probe set.

Table 10: **Stage-wise JS divergence under mild impairment.** Entries report JS divergence, in bits, between each impaired partner and its return matched intermediate Ref checkpoint for each probe-state group. Stages 1–8 correspond to the backward curriculum stages in Table 9, with 25 probe states per stage. The Ref rollout column reports results over 108 states sampled from Ref self-play trajectories and does not correspond to a curriculum stage.

Layout	Impairment	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	Stage 6	Stage 7	Stage 8	Ref rollout
Cramped Room	Valuation	0.244	0.341	0.209	0.283	0.370	0.343	0.260	0.176	0.274
	Action Selection	0.257	0.124	0.077	0.155	0.232	0.188	0.244	0.211	0.113
	Perception	0.242	0.170	0.064	0.104	0.250	0.159	0.202	0.186	0.083
Coordination Ring	Valuation	0.197	0.146	0.184	0.156	0.182	0.289	0.284	0.259	0.181
	Action Selection	0.201	0.242	0.127	0.211	0.243	0.315	0.233	0.244	0.178
	Perception	0.283	0.334	0.220	0.259	0.240	0.370	0.366	0.537	0.246
Counter Circuit	Valuation	0.163	0.176	0.192	0.208	0.161	0.228	0.151	0.200	0.151
	Action Selection	0.158	0.211	0.245	0.276	0.175	0.248	0.172	0.232	0.173
	Perception	0.263	0.295	0.251	0.407	0.268	0.356	0.248	0.289	0.199
Risky Dualpath	Valuation	0.413	0.191	0.195	0.144	0.149	0.201	0.203	0.163	0.283
	Action Selection	0.220	0.204	0.222	0.408	0.222	0.240	0.168	0.176	0.230
	Perception	0.194	0.200	0.174	0.312	0.161	0.192	0.143	0.129	0.212
Risky Multipath	Valuation	0.413	0.348	0.226	0.223	0.109	0.253	0.155	0.254	0.234
	Action Selection	0.213	0.157	0.161	0.236	0.204	0.274	0.195	0.240	0.277
	Perception	0.218	0.174	0.129	0.195	0.158	0.241	0.148	0.231	0.253

JS-divergence computation and aggregation. Let $\pi_{\text{imp}}(\cdot \mid s)$ and $\pi_{\text{match}}(\cdot \mid s)$ denote the action distributions of the impaired partner and its return matched intermediate checkpoint, respectively. Before computing the divergence, infeasible partner actions were masked to zero in both distributions and the remaining probabilities were renormalized over the feasible partner action set. Their equal weight mixture is $\bar{\pi}(\cdot \mid s) = \frac{1}{2}\big[\pi_{\text{imp}}(\cdot \mid s) + \pi_{\text{match}}(\cdot \mid s)\big]$. At each probe state, we computed

$$D_{\text{JS}}(s) = \frac{1}{2} \sum_{a_p \in \mathcal{A}_p} \pi_{\text{imp}}(a_p \mid s) \log_2 \frac{\pi_{\text{imp}}(a_p \mid s)}{\bar{\pi}(a_p \mid s)} + \frac{1}{2} \sum_{a_p \in \mathcal{A}_p} \pi_{\text{match}}(a_p \mid s) \log_2 \frac{\pi_{\text{match}}(a_p \mid s)}{\bar{\pi}(a_p \mid s)}. \tag{37}$$

Base-2 logarithms give divergence in bits, and zero-probability terms contribute zero. The values in Table 13 were averaged over all 308 states.

Tables 10, 11, and 12 report the corresponding group means under mild, moderate, and severe impairment. The comparisons use the same return matched pairs as Table 13.

B.4 Details of situation dependent behavioral variation under fixed impairment

Figure 5 shows the results for the remaining layouts from the experiment conducted in Section 5.3.

B.5 Statistical analysis of RDSA performance and static mixtures

Statistical testing. We compared RDSA with each baseline and static mixture using two-sided paired t-tests on episode return. Each condition was evaluated over 1,000 episodes with distinct random seeds. We applied Holm's correction separately to two sets of comparisons. The baseline set contained 105 tests across 15 conditions and seven comparison methods: the two specialists and five ZSC baselines. The static mixture set contained 165 tests across the same 15 conditions and 11 mixture weights, $\alpha \in \{0, 0.1, \ldots, 1\}$. For each comparison, we computed the paired difference in episode return between RDSA and the comparison method, with positive values indicating higher episode return for RDSA. We reported the mean paired difference and its 95% confidence interval. Statistical significance was determined by Holm-adjusted $p < 0.05$, using the same criterion for positive and negative differences. Let d denote the paired episode return difference. The null hypothesis was $H_0 : \mathbb{E}[d] = 0$, and the alternative hypothesis was $H_1 : \mathbb{E}[d] \neq 0$.

Table 11: **Stage-wise JS divergence under moderate impairment.** Entries report JS divergence, in bits, between each impaired partner and its return matched intermediate Ref checkpoint for each probe-state group. Stages 1–8 and the Ref rollout group are defined as in Table 10.

Layout	Impairment	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	Stage 6	Stage 7	Stage 8	Ref rollout
Cramped Room	Valuation	0.558	0.556	0.441	0.428	0.453	0.495	0.461	0.406	0.436
	Action Selection	0.489	0.501	0.373	0.438	0.445	0.440	0.427	0.406	0.412
	Perception	0.390	0.499	0.349	0.385	0.414	0.408	0.430	0.483	0.404
Coordination Ring	Valuation	0.468	0.463	0.313	0.394	0.338	0.440	0.340	0.526	0.383
	Action Selection	0.444	0.448	0.297	0.400	0.345	0.390	0.368	0.478	0.357
	Perception	0.458	0.478	0.296	0.456	0.334	0.387	0.371	0.491	0.393
Counter Circuit	Valuation	0.314	0.390	0.379	0.522	0.361	0.489	0.453	0.573	0.389
	Action Selection	0.413	0.470	0.465	0.463	0.331	0.500	0.271	0.542	0.425
	Perception	0.368	0.541	0.494	0.470	0.541	0.503	0.482	0.534	0.511
Risky Dualpath	Valuation	0.371	0.234	0.265	0.270	0.233	0.246	0.174	0.196	0.310
	Action Selection	0.387	0.342	0.402	0.421	0.396	0.366	0.311	0.331	0.386
	Perception	0.381	0.409	0.369	0.372	0.192	0.374	0.257	0.382	0.391
Risky Multipath	Valuation	0.413	0.361	0.294	0.309	0.268	0.398	0.281	0.382	0.322
	Action Selection	0.405	0.355	0.422	0.328	0.199	0.344	0.150	0.296	0.358
	Perception	0.320	0.272	0.244	0.240	0.227	0.350	0.241	0.339	0.281

Table 12: **Stage-wise JS divergence under severe impairment.** Entries report JS divergence, in bits, between each impaired partner and its return matched intermediate Ref checkpoint for each probe-state group. Stages 1–8 and the Ref rollout group are defined as in Table 10.

Layout	Impairment	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	Stage 6	Stage 7	Stage 8	Ref rollout
Cramped Room	Valuation	0.678	0.595	0.441	0.475	0.513	0.578	0.515	0.456	0.468
	Action Selection	0.550	0.551	0.399	0.465	0.467	0.447	0.443	0.412	0.441
	Perception	0.498	0.531	0.383	0.435	0.454	0.411	0.440	0.476	0.433
Coordination Ring	Valuation	0.449	0.538	0.434	0.523	0.336	0.541	0.307	0.482	0.396
	Action Selection	0.510	0.486	0.449	0.542	0.424	0.508	0.475	0.537	0.502
	Perception	0.490	0.511	0.328	0.480	0.364	0.417	0.369	0.510	0.423
Counter Circuit	Valuation	0.428	0.502	0.548	0.579	0.442	0.608	0.328	0.512	0.458
	Action Selection	0.538	0.557	0.534	0.488	0.551	0.527	0.542	0.561	0.527
	Perception	0.444	0.557	0.565	0.551	0.488	0.552	0.529	0.544	0.495
Risky Dualpath	Valuation	0.394	0.491	0.316	0.452	0.387	0.418	0.410	0.367	0.356
	Action Selection	0.405	0.466	0.456	0.451	0.287	0.477	0.248	0.361	0.408
	Perception	0.119	0.171	0.151	0.283	0.181	0.214	0.147	0.155	0.238
Risky Multipath	Valuation	0.420	0.213	0.230	0.269	0.272	0.312	0.194	0.264	0.271
	Action Selection	0.459	0.400	0.435	0.351	0.203	0.354	0.157	0.303	0.378
	Perception	0.365	0.186	0.170	0.208	0.188	0.247	0.091	0.192	0.205

Baseline comparisons. Table 14 reports the full episode return results and statistical comparisons between RDSA and the PR-Specialist, SR-Specialist, FCP, HSP, MEP, GAMMA, and E3T. RDSA was significantly better in 94 of the 105 baseline comparisons and significantly worse in 6, with 5 non-significant comparisons.

Static mixture comparisons. Tables 15 and 16 report the statistical comparisons between RDSA and all 11 static mixtures. The mixture weight α denotes the weight assigned to the PR-Specialist. RDSA significantly outperformed the corresponding static mixture in 155 of the 165 comparisons and did not significantly underperform in any comparison.

B.6 Full results for baseline and static mixture comparisons

Table 17 reports the full results corresponding to Section 5.4.

Tables 18 and 19 report the full results corresponding to Section 5.5.

Table 13: **Behavioral differences from return matched checkpoints.** Entries report JS divergence, in bits, between each impaired partner and its return matched intermediate checkpoint over the same 308 probe states. The Baseline column reports the mean JS divergence among return matched Ref checkpoints within each layout. The Average column averages across impairment types and severity levels within each layout, while the Average row averages across layouts and impairment types for each severity level.

Layout	Impairment	Mild	Moderate	Severe	Average	Baseline
Cramped Room	Valuation	0.277	0.461	0.509	0.366	0.090
	Action Selection	0.160	0.430	0.458		
	Perception	0.141	0.414	0.446		
Coordination Ring	Valuation	0.201	0.401	0.432	0.361	0.120
	Action Selection	0.210	0.382	0.495		
	Perception	0.298	0.403	0.430		
Counter Circuit	Valuation	0.173	0.419	0.481	0.390	0.192
	Action Selection	0.200	0.429	0.534		
	Perception	0.263	0.498	0.517		
Risky Dualpath	Valuation	0.234	0.270	0.387	0.295	0.182
	Action Selection	0.232	0.375	0.399		
	Perception	0.196	0.359	0.199		
Risky Multipath	Valuation	0.243	0.332	0.272	0.273	0.183
	Action Selection	0.234	0.329	0.349		
	Perception	0.210	0.280	0.205		
Average		**0.218**	**0.385**	**0.408**		

Table 14: **Paired episode level reward differences.** Entries report the mean paired difference in episode return between RDSA and each comparison method over 1,000 matched episodes per condition, with $\pm$ denoting the half-width of an unadjusted 95% confidence interval. Positive values favor RDSA, whereas negative values favor the comparison method. ** indicates Holm-adjusted $p < 0.01$ from two-sided paired t-tests across all 105 comparisons (15 layout–impairment conditions $\times$ 7 comparison methods).

Layout	Impairment	PR-Specialist	SR-Specialist	FCP	HSP	MEP	GAMMA	E3T
Cramped Room	Valuation	$+31.1 \pm 3.0^{**}$	$+16.7 \pm 3.5^{**}$	$+15.1 \pm 2.5^{**}$	$+13.8 \pm 2.3^{**}$	$+21.4 \pm 3.1^{**}$	$+51.2 \pm 3.4^{**}$	$+26.7 \pm 2.8^{**}$
	Action selection	$+37.3 \pm 3.0^{**}$	$+20.9 \pm 3.3^{**}$	$+11.4 \pm 1.6^{**}$	$+10.8 \pm 1.7^{**}$	$+33.0 \pm 3.5^{**}$	$+43.6 \pm 2.9^{**}$	$+18.2 \pm 2.1^{**}$
	Perception	$+25.2 \pm 2.8^{**}$	$+14.1 \pm 3.3^{**}$	$+14.1 \pm 2.1^{**}$	$+8.9 \pm 2.0^{**}$	$+29.5 \pm 3.2^{**}$	$+41.4 \pm 3.4^{**}$	$+9.2 \pm 2.0^{**}$
Coordination Ring	Valuation	$+10.0 \pm 2.0^{**}$	$+36.0 \pm 4.0^{**}$	$+25.6 \pm 2.6^{**}$	$+25.5 \pm 2.9^{**}$	$+20.2 \pm 2.9^{**}$	$+52.6 \pm 3.0^{**}$	$+44.5 \pm 2.4^{**}$
	Action selection	$+16.1 \pm 2.3^{**}$	$+19.5 \pm 3.0^{**}$	$+15.2 \pm 2.2^{**}$	$+14.4 \pm 1.9^{**}$	$+11.5 \pm 2.3^{**}$	$+34.7 \pm 2.0^{**}$	$+34.5 \pm 2.4^{**}$
	Perception	$+18.1 \pm 2.2^{**}$	$+14.2 \pm 2.8^{**}$	$+12.9 \pm 2.0^{**}$	$+11.1 \pm 2.0^{**}$	$+8.9 \pm 2.2^{**}$	$+32.4 \pm 2.0^{**}$	$+33.8 \pm 2.7^{**}$
Counter Circuit	Valuation	$+32.8 \pm 3.3^{**}$	$+19.9 \pm 3.5^{**}$	$+46.3 \pm 3.0^{**}$	$+78.3 \pm 2.4^{**}$	$+89.4 \pm 2.5^{**}$	$+79.2 \pm 2.6^{**}$	$+48.9 \pm 3.1^{**}$
	Action selection	$+50.9 \pm 3.3^{**}$	$+11.0 \pm 2.8^{**}$	$+36.1 \pm 2.3^{**}$	$+72.8 \pm 2.0^{**}$	$+88.3 \pm 2.2^{**}$	$+78.2 \pm 2.1^{**}$	$+65.1 \pm 3.0^{**}$
	Perception	$+67.2 \pm 3.1^{**}$	$+8.0 \pm 2.4^{**}$	$+56.3 \pm 2.5^{**}$	$+80.9 \pm 2.0^{**}$	$+100.3 \pm 2.1^{**}$	$+91.4 \pm 2.2^{**}$	$+85.1 \pm 2.8^{**}$
Risky Dualpath	Valuation	-0.2 ± 3.0	$+7.4 \pm 2.4^{**}$	$-22.9 \pm 2.5^{**}$	$-19.8 \pm 2.4^{**}$	$-12.9 \pm 2.8^{**}$	$+11.4 \pm 2.4^{**}$	$+18.8 \pm 3.7^{**}$
	Action selection	$+15.2 \pm 2.9^{**}$	$+29.6 \pm 3.6^{**}$	$+15.8 \pm 3.0^{**}$	$+18.9 \pm 2.7^{**}$	$+14.0 \pm 3.2^{**}$	$+44.3 \pm 2.6^{**}$	$+52.3 \pm 4.5^{**}$
	Perception	$+11.1 \pm 2.8^{**}$	$+55.3 \pm 5.0^{**}$	$+29.0 \pm 3.4^{**}$	$+29.9 \pm 3.2^{**}$	$+25.1 \pm 3.4^{**}$	$+50.2 \pm 3.4^{**}$	$+82.1 \pm 5.3^{**}$
Risky Multipath	Valuation	$+8.5 \pm 2.8^{**}$	$+0.8 \pm 2.3$	$-12.0 \pm 2.3^{**}$	$-6.0 \pm 2.3^{**}$	-3.0 ± 2.6	$+14.1 \pm 2.4^{**}$	$+35.3 \pm 2.9^{**}$
	Action selection	$+8.6 \pm 2.4^{**}$	$+13.6 \pm 2.9^{**}$	$-6.4 \pm 2.0^{**}$	-1.3 ± 2.2	$+0.1 \pm 2.3$	$+19.0 \pm 1.9^{**}$	$+38.7 \pm 2.8^{**}$
	Perception	$+12.7 \pm 2.6^{**}$	$+27.1 \pm 3.2^{**}$	$+4.6 \pm 2.3^{**}$	$+17.2 \pm 2.5^{**}$	$+10.9 \pm 2.5^{**}$	$+29.7 \pm 2.2^{**}$	$+58.7 \pm 2.9^{**}$

Table 15: **Paired episode level reward differences for static mixtures ($\alpha = 0.0$–0.5).** Entries report the mean paired difference in episode return between RDSA and each static mixture over 1,000 matched episodes per condition, with $\pm$ denoting the half-width of an unadjusted 95% confidence interval. Positive values favor RDSA, whereas negative values favor the static mixture. The mixture weight α denotes the weight on the PR-Specialist, with $\alpha = 0$ corresponding to the SR-Specialist and $\alpha = 1$ to the PR-Specialist. ** indicates Holm-adjusted $p < 0.01$ from two-sided paired t-tests across all 165 comparisons (15 layout–impairment conditions $\times$ 11 mixture weights).

Layout	Impairment	$\alpha = 0.0$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$
Cramped Room	Valuation	$+16.7 \pm 3.5^{**}$	$+12.8 \pm 3.2^{**}$	$+9.6 \pm 2.8^{**}$	$+5.0 \pm 2.6^{**}$	$+18.0 \pm 2.9^{**}$	$+16.7 \pm 2.5^{**}$
	Action selection	$+20.9 \pm 3.3^{**}$	$+12.4 \pm 2.6^{**}$	$+8.2 \pm 2.2^{**}$	$+4.7 \pm 1.7^{**}$	$+15.7 \pm 2.3^{**}$	$+15.1 \pm 2.1^{**}$
	Perception	$+14.1 \pm 3.3^{**}$	$+9.1 \pm 2.8^{**}$	$+6.8 \pm 2.6^{**}$	$+4.3 \pm 2.1^{**}$	$+7.6 \pm 2.2^{**}$	$+7.8 \pm 2.1^{**}$
Coordination Ring	Valuation	$+36.0 \pm 4.0^{**}$	$+34.8 \pm 3.9^{**}$	$+35.7 \pm 3.9^{**}$	$+32.3 \pm 3.5^{**}$	$+34.4 \pm 3.6^{**}$	$+29.9 \pm 3.2^{**}$
	Action selection	$+19.5 \pm 3.0^{**}$	$+18.6 \pm 2.9^{**}$	$+19.0 \pm 3.0^{**}$	$+15.5 \pm 2.5^{**}$	$+17.8 \pm 2.2^{**}$	$+28.7 \pm 2.7^{**}$
	Perception	$+14.2 \pm 2.8^{**}$	$+14.1 \pm 2.9^{**}$	$+15.4 \pm 2.9^{**}$	$+11.1 \pm 2.5^{**}$	$+13.8 \pm 2.4^{**}$	$+21.3 \pm 2.6^{**}$
Counter Circuit	Valuation	$+19.9 \pm 3.5^{**}$	$+19.5 \pm 3.4^{**}$	$+21.8 \pm 3.5^{**}$	$+32.4 \pm 3.7^{**}$	$+55.3 \pm 3.5^{**}$	$+73.3 \pm 2.8^{**}$
	Action selection	$+11.0 \pm 2.8^{**}$	$+10.5 \pm 2.7^{**}$	$+16.5 \pm 3.1^{**}$	$+26.8 \pm 3.4^{**}$	$+43.1 \pm 3.2^{**}$	$+69.5 \pm 2.5^{**}$
	Perception	$+8.0 \pm 2.4^{**}$	$+10.7 \pm 2.6^{**}$	$+13.7 \pm 2.9^{**}$	$+20.6 \pm 3.1^{**}$	$+39.1 \pm 3.1^{**}$	$+71.3 \pm 2.7^{**}$
Risky Dualpath	Valuation	$+7.4 \pm 2.4^{**}$	$+7.3 \pm 2.4^{**}$	$+3.0 \pm 2.5$	$+2.4 \pm 2.6$	$+6.8 \pm 2.8^{**}$	$+22.6 \pm 3.5^{**}$
	Action selection	$+29.6 \pm 3.6^{**}$	$+29.6 \pm 3.7^{**}$	$+30.0 \pm 3.7^{**}$	$+30.3 \pm 3.6^{**}$	$+35.2 \pm 3.5^{**}$	$+45.6 \pm 3.5^{**}$
	Perception	$+55.3 \pm 5.0^{**}$	$+52.8 \pm 4.8^{**}$	$+53.2 \pm 4.8^{**}$	$+51.8 \pm 4.8^{**}$	$+53.7 \pm 4.8^{**}$	$+58.2 \pm 4.0^{**}$
Risky Multipath	Valuation	$+0.8 \pm 2.3$	-1.1 ± 2.2	-0.9 ± 2.3	$+1.5 \pm 2.3$	$+4.7 \pm 2.4^{**}$	$+4.1 \pm 2.3^{**}$
	Action selection	$+13.6 \pm 2.9^{**}$	$+12.8 \pm 2.6^{**}$	$+11.5 \pm 2.5^{**}$	$+12.7 \pm 2.4^{**}$	$+17.7 \pm 2.5^{**}$	$+17.2 \pm 2.6^{**}$
	Perception	$+27.1 \pm 3.2^{**}$	$+24.8 \pm 3.2^{**}$	$+23.6 \pm 3.0^{**}$	$+25.5 \pm 3.0^{**}$	$+26.4 \pm 3.1^{**}$	$+28.0 \pm 3.1^{**}$

Table 16: **Paired episode level reward differences for static mixtures ($\alpha = 0.6$–1.0).** Entries report the mean paired difference in episode return between RDSA and each static mixture over 1,000 matched episodes per condition, with $\pm$ denoting the half-width of an unadjusted 95% confidence interval. Positive values favor RDSA, whereas negative values favor the static mixture. The mixture weight α denotes the weight on the PR-Specialist, with $\alpha = 0$ corresponding to the SR-Specialist and $\alpha = 1$ to the PR-Specialist. ** indicates Holm-adjusted $p < 0.01$ from two-sided paired t-tests across all 165 comparisons (15 layout–impairment conditions $\times$ 11 mixture weights).

Layout	Impairment	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$	$\alpha = 1.0$
Cramped Room	Valuation	$+20.1 \pm 2.5^{**}$	$+20.5 \pm 2.5^{**}$	$+21.3 \pm 2.6^{**}$	$+24.8 \pm 2.7^{**}$	$+31.1 \pm 3.0^{**}$
	Action selection	$+18.2 \pm 2.1^{**}$	$+22.0 \pm 2.2^{**}$	$+25.3 \pm 2.4^{**}$	$+27.7 \pm 2.6^{**}$	$+37.3 \pm 3.0^{**}$
	Perception	$+12.6 \pm 2.3^{**}$	$+14.6 \pm 2.3^{**}$	$+16.9 \pm 2.4^{**}$	$+21.3 \pm 2.7^{**}$	$+25.2 \pm 2.8^{**}$
Coordination Ring	Valuation	$+26.6 \pm 2.8^{**}$	$+15.2 \pm 2.7^{**}$	$+16.5 \pm 2.8^{**}$	$+12.9 \pm 2.5^{**}$	$+10.0 \pm 2.0^{**}$
	Action selection	$+24.9 \pm 2.4^{**}$	$+19.5 \pm 2.4^{**}$	$+19.3 \pm 2.6^{**}$	$+16.3 \pm 2.3^{**}$	$+16.1 \pm 2.3^{**}$
	Perception	$+18.9 \pm 2.4^{**}$	$+14.8 \pm 2.3^{**}$	$+18.7 \pm 2.6^{**}$	$+18.3 \pm 2.4^{**}$	$+18.1 \pm 2.2^{**}$
Counter Circuit	Valuation	$+46.8 \pm 3.1^{**}$	$+40.9 \pm 3.0^{**}$	$+34.0 \pm 3.2^{**}$	$+32.5 \pm 3.3^{**}$	$+32.8 \pm 3.3^{**}$
	Action selection	$+55.8 \pm 2.9^{**}$	$+49.3 \pm 3.2^{**}$	$+48.7 \pm 3.2^{**}$	$+49.7 \pm 3.3^{**}$	$+50.9 \pm 3.3^{**}$
	Perception	$+72.3 \pm 2.7^{**}$	$+63.8 \pm 2.8^{**}$	$+63.2 \pm 2.9^{**}$	$+65.3 \pm 3.0^{**}$	$+67.2 \pm 3.1^{**}$
Risky Dualpath	Valuation	$+8.4 \pm 3.3^{**}$	-2.3 ± 2.9	-3.4 ± 2.9	-0.5 ± 3.0	-0.2 ± 3.0
	Action selection	$+37.4 \pm 3.1^{**}$	$+15.0 \pm 2.8^{**}$	$+11.7 \pm 2.7^{**}$	$+10.2 \pm 2.7^{**}$	$+15.2 \pm 2.9^{**}$
	Perception	$+42.4 \pm 3.5^{**}$	$+13.7 \pm 2.8^{**}$	$+8.9 \pm 2.8^{**}$	$+9.4 \pm 2.9^{**}$	$+11.1 \pm 2.8^{**}$
Risky Multipath	Valuation	$+6.8 \pm 2.5^{**}$	$+5.4 \pm 2.5^{**}$	$+7.1 \pm 2.7^{**}$	$+7.4 \pm 2.7^{**}$	$+8.5 \pm 2.8^{**}$
	Action selection	$+8.1 \pm 2.0^{**}$	$+4.1 \pm 2.1^{**}$	$+4.7 \pm 2.3^{**}$	$+6.5 \pm 2.4^{**}$	$+8.6 \pm 2.4^{**}$
	Perception	$+14.5 \pm 2.4^{**}$	$+6.8 \pm 2.2^{**}$	$+8.5 \pm 2.4^{**}$	$+10.9 \pm 2.5^{**}$	$+12.7 \pm 2.6^{**}$

Table 17: **Performance across layout–impairment conditions.** Entries report mean episode returns ± 95% normal CI over 1,000 episodes; the Average row reports the mean across 15 condition means ± 95% normal CI across those means. Bold indicates the highest mean per condition, and underlines mark the two specialists and best performing ZSC baseline when both specialists underperform that baseline but RDSA surpasses it.

Layout	Impairment	RDSA (Ours)	PR-Specialist	SR-Specialist	FCP	HSP	MEP	GAMMA	E3T
Cramped Room	Valuation	**192.9±1.8**	161.8±3.2	176.2±2.9	177.8±2.0	179.1±2.1	171.4±2.5	141.7±3.0	166.2±2.6
	Action selection	**196.7±1.2**	159.4±3.4	175.8±2.9	185.4±1.4	185.9±1.8	163.7±3.1	153.1±2.8	178.5±2.1
	Perception	**196.4±1.6**	171.2±3.1	182.3±2.7	182.3±2.1	187.4±2.0	166.9±3.0	155.0±3.2	187.2±2.0
Coordination Ring	Valuation	**181.7±4.1**	171.7±4.7	145.8±2.4	156.1±2.5	156.2±2.4	161.6±2.7	129.2±2.3	137.2±3.6
	Action selection	**158.9±2.4**	142.8±3.5	139.4±2.0	143.8±1.8	144.6±1.8	147.4±1.9	124.3±1.6	124.4±2.6
	Perception	**156.3±2.4**	138.2±3.4	142.1±1.8	143.4±1.7	145.2±1.9	147.4±1.8	123.9±1.6	122.5±2.8
Counter Circuit	Valuation	**115.5±1.9**	82.7±3.5	95.6±2.4	69.3±1.8	37.2±1.2	26.1±1.2	36.3±1.4	66.7±3.3
	Action selection	**113.3±1.8**	62.5±3.6	102.4±1.9	77.2±1.3	40.5±1.0	25.1±1.0	35.1±1.2	48.2±3.3
	Perception	**119.2±1.5**	52.0±3.5	111.2±1.5	62.8±1.9	38.3±1.2	18.9±1.3	27.8±1.6	34.0±3.3
Risky Dualpath	Valuation	171.9±2.0	172.2±2.6	164.5±2.0	**194.8±1.7**	191.7±1.9	184.9±2.2	160.6±1.5	153.2±3.2
	Action selection	**191.5±2.6**	176.3±4.1	161.9±1.7	175.7±2.1	172.6±1.6	177.5±2.2	147.2±1.7	139.2±3.2
	Perception	**190.7±4.5**	179.6±5.7	135.4±2.6	161.7±3.0	160.8±2.5	165.6±3.2	140.5±2.6	108.6±3.6
Risky Multipath	Valuation	112.6±1.8	104.1±2.4	111.8±1.6	**124.6±1.7**	118.7±1.7	115.6±2.1	98.5±1.7	77.4±2.3
	Action selection	134.3±1.6	125.7±2.3	120.7±2.1	**140.7±1.5**	135.6±1.6	134.2±1.6	115.3±1.3	95.6±2.2
	Perception	**122.2±2.7**	109.5±4.0	95.2±2.3	117.6±2.6	105.1±2.3	111.3±2.4	92.5±2.1	63.6±2.7
Average		**157.0±17.3**	134.0±21.7	137.4±15.3	140.9±21.8	133.3±27.8	127.8±29.3	112.1±22.9	113.5±24.3

Table 18: **Full static mixture results** for $\alpha = 0.0$–0.5. Results are reported across 15 layout–impairment conditions. For each condition, entries report the mean episode return, with $\pm$ denoting the 95% confidence interval across 1,000 evaluation episodes. The average row reports the mean across the 15 conditions, with $\pm$ denoting the 95% confidence interval across condition means. The mixture weight α denotes the weight assigned to the PR-Specialist, with $\alpha = 0$ corresponding to the SR-Specialist and $\alpha = 1$ to the PR-Specialist. Bold marks the condition wise best static mixture over the full $\alpha \in [0, 1]$ range when it appears in this table, and also the RDSA mean when it is higher than all static mixture means. Bold does not indicate statistical significance; statistical comparisons are reported in Table 15.

Layout	Impairment	$\alpha = 0.0$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	RDSA
Cramped Room	Valuation	176.2 ± 2.9	180.1 ± 2.7	183.3 ± 2.4	$\mathbf{187.9 \pm 1.9}$	174.8 ± 2.5	176.2 ± 2.3	$\mathbf{192.9 \pm 1.8}$
	Action selection	175.8 ± 2.9	184.3 ± 2.3	188.5 ± 1.9	$\mathbf{192.0 \pm 1.4}$	181.0 ± 2.1	181.6 ± 2.1	$\mathbf{196.7 \pm 1.2}$
	Perception	182.3 ± 2.7	187.2 ± 2.4	189.6 ± 2.1	$\mathbf{192.1 \pm 1.7}$	188.8 ± 1.9	188.5 ± 2.0	$\mathbf{196.4 \pm 1.6}$
Coordination Ring	Valuation	145.8 ± 2.4	146.9 ± 2.4	146.0 ± 2.4	149.4 ± 2.3	147.4 ± 2.5	151.8 ± 2.9	$\mathbf{181.7 \pm 4.1}$
	Action selection	139.4 ± 2.0	140.3 ± 2.0	139.9 ± 2.1	$\mathbf{143.4 \pm 1.8}$	141.1 ± 1.8	130.2 ± 2.4	$\mathbf{158.9 \pm 2.4}$
	Perception	142.1 ± 1.8	142.2 ± 1.8	140.9 ± 1.9	$\mathbf{145.3 \pm 1.8}$	142.5 ± 1.8	135.0 ± 2.4	$\mathbf{156.3 \pm 2.4}$
Counter Circuit	Valuation	95.6 ± 2.4	$\mathbf{96.0 \pm 2.4}$	93.7 ± 2.5	83.2 ± 2.7	60.2 ± 2.5	42.2 ± 1.8	$\mathbf{115.5 \pm 1.9}$
	Action selection	102.4 ± 1.9	$\mathbf{102.8 \pm 1.9}$	96.8 ± 2.2	86.5 ± 2.4	70.2 ± 2.2	43.8 ± 1.7	$\mathbf{113.3 \pm 1.8}$
	Perception	$\mathbf{111.2 \pm 1.5}$	108.5 ± 1.8	105.5 ± 1.9	98.5 ± 2.1	80.1 ± 2.2	47.9 ± 2.0	$\mathbf{119.2 \pm 1.5}$
Risky Dualpath	Valuation	164.5 ± 2.0	164.7 ± 2.1	168.9 ± 1.9	169.6 ± 1.8	165.1 ± 2.1	149.3 ± 3.1	171.9 ± 2.0
	Action selection	161.9 ± 1.7	161.9 ± 1.8	161.5 ± 1.9	161.2 ± 1.8	156.4 ± 1.7	145.9 ± 2.6	$\mathbf{191.5 \pm 2.6}$
	Perception	135.4 ± 2.6	137.9 ± 2.6	137.5 ± 2.6	138.9 ± 2.6	137.0 ± 2.6	132.5 ± 3.2	$\mathbf{190.7 \pm 4.5}$
Risky Multipath	Valuation	111.8 ± 1.6	$\mathbf{113.7 \pm 1.5}$	113.5 ± 1.6	111.1 ± 1.6	107.9 ± 1.8	108.5 ± 1.8	112.6 ± 1.8
	Action selection	120.7 ± 2.1	121.5 ± 1.9	122.8 ± 1.8	121.6 ± 1.8	116.6 ± 1.9	117.1 ± 1.9	$\mathbf{134.3 \pm 1.6}$
	Perception	95.2 ± 2.3	97.5 ± 2.3	98.6 ± 2.4	96.8 ± 2.3	95.9 ± 2.4	94.2 ± 2.6	$\mathbf{122.2 \pm 2.7}$
Average		137.4 ± 15.3	139.0 ± 16.1	$\mathbf{139.1 \pm 17.1}$	138.5 ± 19.1	131.0 ± 20.7	123.0 ± 24.3	$\mathbf{157.0 \pm 17.3}$

Table 19: **Full static mixture results** for $\alpha = 0.6$–1.0. Results are reported across 15 layout–impairment conditions. For each condition, entries report the mean episode return, with $\pm$ denoting the 95% confidence interval across 1,000 evaluation episodes. The average row reports the mean across the 15 conditions, with $\pm$ denoting the 95% confidence interval across condition means. The mixture weight α denotes the weight assigned to the PR-Specialist, with $\alpha = 0$ corresponding to the SR-Specialist and $\alpha = 1$ to the PR-Specialist. Bold marks the condition wise best static mixture over the full $\alpha \in [0, 1]$ range when it appears in this table, and also the RDSA mean when it is higher than all static mixture means. Bold does not indicate statistical significance; statistical comparisons are reported in Table 16.

Layout	Impairment	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$	$\alpha = 1.0$	RDSA
Cramped Room	Valuation	172.8 ± 2.5	172.4 ± 2.6	171.6 ± 2.7	168.1 ± 2.9	161.8 ± 3.2	$\mathbf{192.9 \pm 1.8}$
	Action selection	178.5 ± 2.2	174.7 ± 2.5	171.4 ± 2.7	169.1 ± 2.9	159.4 ± 3.4	$\mathbf{196.7 \pm 1.2}$
	Perception	183.7 ± 2.5	181.8 ± 2.6	179.4 ± 2.7	175.0 ± 3.0	171.2 ± 3.1	$\mathbf{196.4 \pm 1.6}$
Coordination Ring	Valuation	155.1 ± 4.2	166.5 ± 4.7	165.2 ± 5.2	168.9 ± 5.0	$\mathbf{171.7 \pm 4.7}$	$\mathbf{181.7 \pm 4.1}$
	Action selection	134.0 ± 3.2	139.4 ± 3.4	139.7 ± 3.6	142.7 ± 3.5	142.8 ± 3.5	$\mathbf{158.9 \pm 2.4}$
	Perception	137.4 ± 3.0	141.5 ± 3.2	137.6 ± 3.5	138.0 ± 3.5	138.2 ± 3.4	$\mathbf{156.3 \pm 2.4}$
Counter Circuit	Valuation	68.7 ± 2.8	74.6 ± 3.0	81.5 ± 3.3	83.0 ± 3.4	82.7 ± 3.5	$\mathbf{115.5 \pm 1.9}$
	Action selection	57.6 ± 2.6	64.1 ± 3.3	64.6 ± 3.4	63.7 ± 3.6	62.5 ± 3.6	$\mathbf{113.3 \pm 1.8}$
	Perception	46.9 ± 2.4	55.4 ± 3.2	56.0 ± 3.4	53.9 ± 3.5	52.0 ± 3.5	$\mathbf{119.2 \pm 1.5}$
Risky Dualpath	Valuation	163.6 ± 2.8	174.3 ± 2.5	$\mathbf{175.4 \pm 2.5}$	172.4 ± 2.6	172.2 ± 2.6	171.9 ± 2.0
	Action selection	154.1 ± 3.4	176.5 ± 3.7	179.8 ± 3.9	$\mathbf{181.3 \pm 3.9}$	176.3 ± 4.1	$\mathbf{191.5 \pm 2.6}$
	Perception	148.3 ± 4.3	177.0 ± 5.3	$\mathbf{181.8 \pm 5.6}$	181.3 ± 5.7	179.6 ± 5.7	$\mathbf{190.7 \pm 4.5}$
Risky Multipath	Valuation	105.8 ± 2.1	107.2 ± 2.2	105.5 ± 2.4	105.2 ± 2.4	104.1 ± 2.4	112.6 ± 1.8
	Action selection	126.2 ± 1.6	$\mathbf{130.2 \pm 1.9}$	129.6 ± 2.0	127.8 ± 2.2	125.7 ± 2.3	$\mathbf{134.3 \pm 1.6}$
	Perception	107.7 ± 2.9	$\mathbf{115.4 \pm 3.4}$	113.7 ± 3.7	111.3 ± 3.9	109.5 ± 4.0	$\mathbf{122.2 \pm 2.7}$
Average		129.4 ± 22.1	136.7 ± 22.4	136.9 ± 22.1	136.1 ± 22.1	134.0 ± 21.7	$\mathbf{157.0 \pm 17.3}$

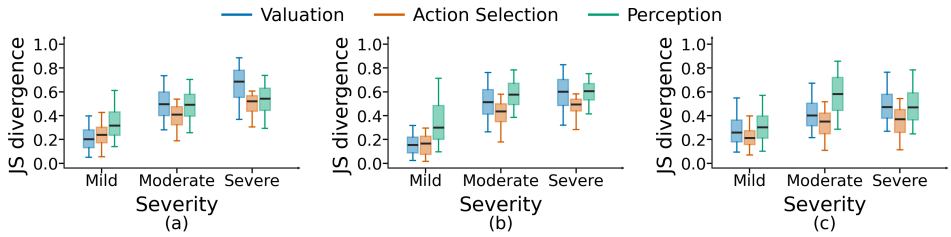


Figure 5: **State dependent behavioral differences across additional layouts.** JS divergence, in bits, between each impaired partner and its return matched intermediate Ref checkpoint across states visited during cross-play with Ref. Results are grouped by impairment severity and impairment type and are shown after a 10-step moving average within each episode. Boxes show the interquartile range, center lines the median, and whiskers the 5th and 95th percentiles. (a), (b), and (c) show the results for Coordination Ring, Counter Circuit, and Risky Multipath, respectively.

C Extended Experiments

C.1 Robustness to impaired behavioral cloning partners

Behavioral cloning (BC) models trained from human gameplay data have been widely used as human proxies for evaluating zero shot human–AI coordination in Overcooked (Carroll et al., 2019; Strouse et al., 2021; Zhao et al., 2023; Yu et al., 2023; Yan et al., 2023). To evaluate coordination with human-like partners, prior work has used the evaluation protocol proposed by Carroll et al. (Carroll et al., 2019), in which a human proxy model is trained through behavioral cloning on collected human gameplay data and used as an evaluation partner. Following this protocol, we conducted an additional experiment using an independently trained BC human proxy and introduced decision process impairment into computational processes that have meaningful counterparts in the BC policy. This evaluation examines whether RDSA can effectively cooperate with impaired partners constructed from a human proxy model commonly used in ZSC evaluation.

The BC partner receives a 96-dimensional feature representation of the current state and directly produces action logits, from which its action is sampled. We applied decision process impairment to computational processes that can be directly identified and modified in the pretrained BC policy. For *Perception*, Gaussian noise was added to the BC input features. For *Action selection (A)*, the action logits were scaled with $\lambda < 1$, flattening the action distribution and increasing action stochasticity. For *Action selection (B)*, the logits were scaled with $\lambda > 1$, sharpening the action distribution and reducing action flexibility.

We calibrated the perturbation ranges using self-play of the pretrained BC policy in Cramped Room before the main evaluation. Specifically, we performed pilot sweeps from the unperturbed policy toward increasingly severe perturbations using $\sigma \in [0, 4]$ for Perception, $\lambda \in [0.05, 1]$ for Action selection (A), and $\lambda \in [1, 20]$ for Action selection (B), with 60 self-play episodes for each calibration setting. The two endpoints represent the unperturbed policy and a severely perturbed policy whose BC self-play performance substantially deteriorated. Based on these sweeps, we used $\sigma \sim U(0, 2)$ for Perception, $\lambda \sim \mathrm{LogU}(0.05, 1)$ for Action selection (A), and $\lambda \sim \mathrm{LogU}(1, 10)$ for Action selection (B). These ranges were fixed after calibration and applied unchanged across all evaluated layouts, avoiding layout specific adjustment of perturbation severity.

The perturbation severity was sampled once at the beginning of each episode and remained fixed throughout the episode, while feature noise was resampled at each timestep. All compared methods were evaluated using the same episode indices, sampled perturbation severities, and random conditions, with 1,000 episodes per condition. We did not construct a valuation impairment for the BC partner because this evaluation directly uses an already trained BC policy rather than retraining or fine-tuning the partner model for each impairment.

Tables 20 and 21 summarize the results. RDSA achieved the highest mean episode return in all nine evaluated conditions across the three layouts and three BC perturbations. We additionally compared RDSA with each ZSC baseline using paired t-tests under the matched episode conditions described above and applied Holm correction across the 45 pairwise comparisons. Of these comparisons, 41

Table 20: **Performance with impaired behavioral cloning partners.** Entries report mean episode return and standard deviation over 1,000 episodes. Perception adds Gaussian noise to the BC input features. Action selection (A) flattens the action distribution, whereas Action selection (B) sharpens it. Bold values indicate the highest mean episode return in each condition.

Layout	Impairment	RDSA (Ours)	MEP	FCP	HSP	GAMMA	E3T
Cramped Room	Perception	**151.7±32.0**	146.7±36.4	149.9±27.0	145.7±27.9	117.8±35.0	115.3±38.9
	Action selection (A)	**162.3±33.6**	156.3±38.1	160.0±28.3	160.0±28.7	125.8±43.6	141.9±37.5
	Action selection (B)	**132.4±61.0**	117.4±60.3	120.7±55.1	124.6±50.3	87.5±51.3	63.4±54.8
Coordination Ring	Perception	**127.8±16.6**	121.5±19.7	117.7±22.1	119.1±20.7	109.4±19.1	77.8±27.2
	Action selection (A)	**133.2±21.1**	129.2±20.9	123.5±23.6	125.6±23.2	110.7±24.0	92.9±34.3
	Action selection (B)	**32.0±42.7**	31.6±40.5	25.4±34.1	28.0±37.5	27.9±37.7	10.2±25.1
Counter Circuit	Perception	**89.5±35.0**	6.4±13.0	42.6±25.1	16.4±16.5	7.3±11.7	0.3±2.4
	Action selection (A)	**83.9±41.8**	19.1±23.2	57.1±28.2	22.5±22.5	15.2±19.3	1.7±5.6
	Action selection (B)	**10.7±21.5**	4.3±10.9	5.3±13.7	3.2±10.5	0.6±3.7	0.1±1.5

Table 21: **Paired reward differences with impaired behavioral cloning partners.** Paired reward differences between RDSA and the ZSC baselines when interacting with impaired behavioral cloning partners. Each entry reports the mean paired difference $\Delta = R_{\text{RDSA}} - R_{\text{comparison}}$ with $\pm$ 95% confidence interval. Positive values favor RDSA. Statistical significance was assessed using paired t-tests with Holm correction across comparisons. $^{**}p_{\text{Holm}} < 0.01$.

Layout	Impairment	MEP	FCP	HSP	GAMMA	E3T
Cramped Room	Perception	$+5.0 \pm 2.7^{**}$	$+1.8 \pm 2.4$	$+6.0 \pm 2.5^{**}$	$+33.9 \pm 2.8^{**}$	$+36.4 \pm 3.3^{**}$
	Action selection (A)	$+6.0 \pm 2.8^{**}$	$+2.3 \pm 2.5$	$+2.3 \pm 2.7$	$+36.5 \pm 3.0^{**}$	$+20.4 \pm 3.3^{**}$
	Action selection (B)	$+15.0 \pm 4.4^{**}$	$+11.7 \pm 4.3^{**}$	$+7.8 \pm 4.3^{**}$	$+44.9 \pm 4.0^{**}$	$+69.0 \pm 6.1^{**}$
Coordination Ring	Perception	$+6.3 \pm 1.6^{**}$	$+10.1 \pm 1.6^{**}$	$+8.7 \pm 1.5^{**}$	$+18.4 \pm 1.5^{**}$	$+50.0 \pm 2.0^{**}$
	Action selection (A)	$+4.0 \pm 1.8^{**}$	$+9.7 \pm 1.8^{**}$	$+7.6 \pm 1.8^{**}$	$+22.5 \pm 1.9^{**}$	$+40.3 \pm 2.5^{**}$
	Action selection (B)	$+0.4 \pm 2.3$	$+6.5 \pm 2.2^{**}$	$+3.9 \pm 2.2^{**}$	$+4.0 \pm 2.2^{**}$	$+21.8 \pm 2.3^{**}$
Counter Circuit	Perception	$+83.1 \pm 2.4^{**}$	$+46.9 \pm 2.5^{**}$	$+73.1 \pm 2.2^{**}$	$+82.2 \pm 2.2^{**}$	$+89.2 \pm 2.2^{**}$
	Action selection (A)	$+64.8 \pm 3.0^{**}$	$+26.8 \pm 2.4^{**}$	$+61.4 \pm 2.7^{**}$	$+68.7 \pm 2.5^{**}$	$+82.2 \pm 2.6^{**}$
	Action selection (B)	$+6.5 \pm 1.4^{**}$	$+5.4 \pm 1.5^{**}$	$+7.5 \pm 1.4^{**}$	$+10.1 \pm 1.3^{**}$	$+10.6 \pm 1.3^{**}$

Table 22: **Arbitration weight across behavioral difference intervals.** Mean arbitration weight p_t assigned to the PR-Specialist across four behavioral difference intervals. Larger interval values indicate greater behavioral difference from the fully trained Ref. Each entry reports the mean of episode level p_t values across the 30 episodes and three impairment types for each layout. Only episodes containing at least one timestep in the corresponding interval contribute to that interval.

Layout	0–.25	.25–.5	.5–.75	.75–1
Cramped Room	0.810	0.418	0.373	0.273
Coordination Ring	0.798	0.522	0.449	0.346
Counter Circuit	0.824	0.379	0.280	0.224
Risky Dualpath	0.700	0.403	0.287	0.249
Risky Multipath	0.715	0.381	0.320	0.270

showed a significant positive difference in favor of RDSA at $p_{\mathrm{Holm}} < 0.01$. The four nonsignificant comparisons were against FCP under Perception, FCP and HSP under Action selection (A) in Cramped Room, and against MEP under Action selection (B) in Coordination Ring. These results further support the robustness of RDSA across different partner models and impairment settings.

C.2 Behavior dependent adaptation of arbitration control

To evaluate whether RDSA adapted the specialists' relative influence to the partner's current behavior, we examined the relationship between the arbitration weight p_t and the impaired partner's behavioral difference from the fully trained Ref. We evaluated 15 layout–impairment conditions using 30 episodes of 400 timesteps per condition. Impairment severity remained fixed within each episode. Unlike the JS divergence analyses above, which used return matched intermediate checkpoints, this analysis used the fully trained Ref. At each timestep, we computed a behavioral difference score from the probability that the fully trained Ref assigned to the observed partner action. Specifically, we used normalized clipped surprisal:

$$d_t = \min\left(\frac{-\log \pi_{\mathrm{Ref}}(a_p^t \mid s_t)}{3}, 1\right), \tag{38}$$

where larger values indicate that the observed partner action is less likely under Ref. The score is zero when Ref assigns probability 1 to the observed action and is clipped at 1 when the probability is at most e^{-3}.

We applied an 8-step trailing moving average to the behavioral difference score, using the available history during the first seven timesteps. The smoothed score was divided into four intervals: $[0, 0.25)$, $[0.25, 0.5)$, $[0.5, 0.75)$, and $[0.75, 1]$. For each episode, we averaged p_t over the timesteps assigned to each interval. An episode contributed to an interval only if at least one timestep fell within that interval. We then averaged these episode level values across the 30 episodes and three impairment types for each layout. Because not every episode visited every behavioral difference interval, the number of contributing episodes differed across intervals, with the lowest difference interval occurring less frequently than the middle intervals.

Table 22 shows that the mean p_t decreased monotonically with behavioral difference in all five layouts. It ranged from 0.700 to 0.824 in the lowest interval and from 0.224 to 0.346 in the highest interval. These results indicate that RDSA adjusts the specialists' relative influence within an episode according to the partner's current behavior.

C.3 Arbitration control response to impairment severity

To examine whether arbitration control responds systematically to the severity of partner impairment, we analyzed the relationship between the injected impairment severity and the episode mean arbitration weight p_t. We evaluated 15 layout–impairment conditions, consisting of five layouts and three impairment types, with 30 episodes of 400 timesteps per condition. For each episode, impairment severity was sampled independently of the policy at the beginning of the episode and remained fixed throughout the episode. The injected severity was recovered from the corresponding episode seed. We then averaged p_t over all 400 timesteps, such that each episode was represented by a

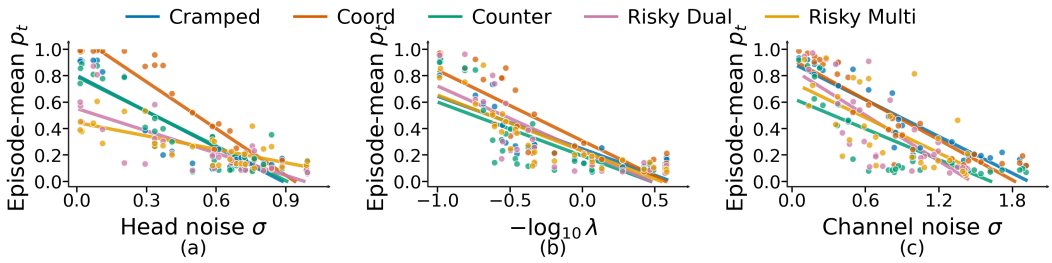


Figure 6: **Relationship between impairment severity and arbitration weight.** Relationship between the injected impairment severity and the episode mean arbitration weight p_t for (a) valuation, (b) action selection, and (c) perception impairment. Each point represents one episode, with 30 episodes evaluated for each layout–impairment condition. Lines show ordinary least squares fits separately for each layout. For valuation and perception impairment, severity is represented by σ_{val} and σ_{perc}, respectively. For action selection impairment, the x-axis shows $-\log_{10}\lambda$, such that larger values correspond to more severe impairment.

single impairment severity and a single episode mean arbitration weight. Severity was represented according to the parameter directly used to construct each impairment. For valuation impairment, we used the head noise level σ_{val}, and for perception impairment, we used the channel noise level σ_{perc}. Larger values of both parameters indicate more severe impairment. For action selection impairment, smaller values of the QRE rationality parameter λ indicate more severe impairment. We therefore used $-\log_{10}\lambda$ so that, for all three impairment types, increasing values correspond to increasing impairment severity. Within each layout–impairment condition, we examined both Pearson correlation and Spearman correlation between the injected severity and the episode mean p_t.

Figure 6 shows the episode level relationship for valuation, action selection, and perception impairment. Each point represents one episode, and the lines show ordinary least squares fits separately for each layout. Across the three impairment types, the fitted lines consistently decrease as impairment severity increases. This pattern appears across all five layouts despite differences in environment structure and in the mechanism through which each impairment affects the partner. In particular, higher valuation noise, lower action selection rationality, and higher perception noise were all associated with lower episode mean p_t. Table 23 quantifies these relationships for each of the 15 layout–impairment conditions. Pearson correlations were negative in every condition, and all 95% confidence intervals remained below zero. The Fisher z averaged Pearson correlation across the 15 conditions was $r = -0.837$. Spearman correlations showed the same negative pattern, with an arithmetic mean of $\rho = -0.835$, indicating that the decrease in p_t was also consistently monotonic across conditions. The agreement between Pearson and Spearman correlations further shows that the observed relationship was not limited to a small number of extreme episodes but was broadly present across the range of impairment severity.

Because p_t determines the relative influence of the two specialists, a lower value assigns less influence to the PR-Specialist and more influence to the SR-Specialist. The consistent negative relationship therefore indicates that RDSA systematically shifts the specialists' relative influence toward the SR-Specialist as partner impairment becomes more severe. This pattern was observed across all five layouts and all three impairment types, suggesting that the response of arbitration control is not specific to a particular environment or form of decision process impairment. Overall, these results support the intended behavior of arbitration control: progressively reducing the influence of the strategy that assumes cooperative partner contribution when the partner exhibits more severe decision process impairment.

C.4 Linear mixed-effects analysis of arbitration control gain

To examine how RDSA's arbitration control gain varies across baselines, specialists, and evaluation conditions, we analyzed the gain using a linear mixed-effects model (LMM) (Laird & Ware, 1982). The data used for this analysis are the episode level evaluation results from Section 5.4, where RDSA is compared with five ZSC baselines and the fixed PR- and SR-Specialists under the same conditions. Throughout this analysis, performance is measured by the total shared reward accumulated over an

Table 23: **Association between impairment severity and arbitration weight.** Episode level association between the injected impairment severity and the episode mean arbitration weight p_t across layout–impairment conditions. Each condition contains 30 episodes of 400 timesteps. The table reports Pearson r with 95% confidence intervals and Spearman ρ for each condition. The confidence intervals for Pearson r are computed using the Fisher z transformation. The average row reports the Fisher z averaged Pearson correlation and the arithmetic mean Spearman correlation across the 15 conditions.

Layout	Impairment	n	Pearson r [95% CI]	Spearman ρ
Cramped Room	Valuation	30	$-0.872\ [-0.938, -0.746]$	-0.744
	Action Selection		$-0.803\ [-0.902, -0.624]$	-0.908
	Perception		$-0.907\ [-0.955, -0.812]$	-0.943
Coordination Ring	Valuation	30	$-0.952\ [-0.977, -0.900]$	-0.947
	Action Selection		$-0.843\ [-0.923, -0.693]$	-0.905
	Perception		$-0.907\ [-0.955, -0.812]$	-0.905
Counter Circuit	Valuation	30	$-0.902\ [-0.953, -0.802]$	-0.778
	Action Selection		$-0.736\ [-0.866, -0.511]$	-0.853
	Perception		$-0.780\ [-0.890, -0.583]$	-0.775
Risky Dualpath	Valuation	30	$-0.770\ [-0.885, -0.568]$	-0.736
	Action Selection		$-0.803\ [-0.902, -0.622]$	-0.892
	Perception		$-0.785\ [-0.893, -0.593]$	-0.807
Risky Multipath	Valuation	30	$-0.726\ [-0.861, -0.495]$	-0.712
	Action Selection		$-0.836\ [-0.919, -0.681]$	-0.932
	Perception		$-0.669\ [-0.829, -0.406]$	-0.693
Average		–	$\mathbf{-0.837}$	$\mathbf{-0.835}$

episode. For each comparison, we define the arbitration control gain Δ as the difference in episode return between RDSA and the corresponding baseline or specialist, with positive values indicating higher reward for RDSA.

All gain values are computed within the same evaluation blocks (layout $\times$ type $\times$ episode). To account for this within block dependence, we included a random intercept for each evaluation block in the LMM.

Under this common analysis setting, we examined arbitration control gain from two perspectives. The first analysis characterizes how the gain varies across experimental conditions, including layout, impairment type, and severity. The second analysis investigates the conditions under which the arbitration control gain becomes larger. Accordingly, the following two sections address two core questions:

- **How do layout, impairment type, and severity modulate arbitration control gain?**
- **Under what conditions does the arbitration control gain increase?**

C.5 How do layout, impairment type, and severity modulate arbitration control gain?

We first examine whether RDSA outperforms the baselines and specialists and how its arbitration control gain varies with layout, impairment type, and severity. For this analysis, we fit the following linear mixed-effects model (LMM), which we refer to as the *primary model*:

$$\Delta \sim 1 + \text{baseline} + \text{layout} + \text{type} \times \text{sev}_z + (1 \mid \text{eval_block}).$$

The primary model was fitted to the episode level evaluation results from Section 5.4. Each observation used the observed arbitration control gain Δ as the dependent variable, computed as the difference in episode return between RDSA and the corresponding baseline or specialist. Baseline, layout, and impairment type were included as categorical fixed effects, while impairment severity

was included as a continuous fixed effect. These effects are estimated jointly, allowing the relationship between each experimental factor and arbitration control gain to be evaluated while accounting for systematic variation associated with the other factors. A random intercept is included for each evaluation block to account for the dependence among gain values obtained from the same $\mathrm{layout} \times \mathrm{type} \times \mathrm{episode}$ condition.

Impairment severity was standardized separately within each impairment type. Accordingly, a one-unit increase in sev_z corresponds to a one-standard-deviation increase in severity within that type. The interaction between impairment type and severity examines whether the change in arbitration control gain associated with increasing severity differs across Valuation, Action Selection, and Perception impairments. This allows us to evaluate whether the benefit of arbitration control increases, decreases, or remains stable as impairment becomes more severe, and whether this pattern differs across impairment types.

After fitting the primary model, we used the fitted model to estimate the expected arbitration control gain under different experimental conditions. For each analysis, we summarize these predictions using model-estimated marginal means. A marginal mean for a factor of interest can be interpreted as the expected gain under average conditions across the other experimental factors (Searle et al., 1980). Unless otherwise stated, severity is evaluated at its mean level within each impairment type. The resulting mean therefore represents the expected episode-reward difference between RDSA and the corresponding baseline or specialist under an overall average combination of the experimental conditions included in the primary model.

For example, suppose a baseline has a model estimated marginal mean gain of $+16.1$ with a 95% confidence interval of $[15.3,\ 16.8]$. Under average conditions across the other experimental factors, the primary model estimates that RDSA achieves an episode return that is, on average, 16.1 higher than that baseline. The confidence interval quantifies uncertainty in this estimate, and an interval entirely above zero supports a positive average arbitration control gain.

The same principle applies to the other factors. For example, a positive mean gain for a layout indicates that, under average conditions across the other experimental factors, the primary model estimates higher episode return for RDSA in that layout. A severity slope of -5, for instance, indicates that the expected arbitration control gain decreases by approximately 5 reward units for each one-standard-deviation increase in impairment severity within the corresponding impairment type. Thus, the estimated means and severity effects describe both the magnitude of RDSA's arbitration control gain and how that gain changes across experimental conditions.

The LMM separates systematic differences in arbitration control gain from variation shared by observations collected under the same evaluation condition. The fixed effect terms estimate how the expected gain differs across baselines and specialists, layouts, and impairment types, and how it changes as impairment severity increases. In particular, the severity effect indicates whether the benefit of arbitration control tends to increase, decrease, or remain stable as impairment becomes more severe. The interaction between impairment type and severity evaluates whether this severity-related change differs across impairment types. This analysis therefore examines not only whether RDSA provides a gain at an average severity level, but also whether that gain is maintained or changes as each type of impairment becomes more severe. The random intercept for $\mathrm{eval_block}$ accounts for block-specific variation shared by observations from the same $\mathrm{layout} \times \mathrm{type} \times \mathrm{episode}$ condition. After fitting the model, we summarize the estimated effects using mean arbitration control gains and 95% confidence intervals. A positive mean gain indicates that RDSA achieves higher expected episode return than the corresponding baseline or specialist after accounting for the other terms in the model, while the magnitude of the gain represents the estimated difference in episode return. The 95% confidence interval reflects the uncertainty of this estimate. An interval entirely above zero supports a positive average gain, whereas an interval containing zero indicates that the model does not clearly distinguish the average gain from zero.

Comparison with baselines and specialists. We first evaluate whether RDSA outperforms the baselines and specialists, and whether its gain can be explained by either specialist alone. We compare RDSA with five ZSC baselines and the two individual specialists (PR and SR), estimating the mean arbitration control gain with 95% confidence intervals. To examine whether the results depend on the random-effects specification, we also fit a sensitivity model that adds a random intercept for *run*, defined by each layout $\times$ baseline-or-specialist combination. Table 24 reports these estimates,

Table 24: **Estimated arbitration control gains of RDSA over the baselines and specialists under the primary and sensitivity models.** Mean gains are model-estimated marginal means, averaged over layouts and impairment types with severity evaluated at the mean level within each type. The sensitivity model additionally includes a random intercept for each layout × comparison-target combination. The 95% confidence intervals are reported for both models. Positive mean gains indicate higher expected episode return for RDSA.

Comparison method	Mean gain	95% CI	
		Primary model	Sensitivity model
FCP	$+16.1$	$[15.3, 16.8]$	$[2.8, 29.3]$
SR	$+19.6$	$[18.8, 20.4]$	$[6.3, 32.9]$
PR	$+23.0$	$[22.2, 23.7]$	$[9.7, 36.2]$
HSP	$+23.7$	$[22.9, 24.5]$	$[10.4, 37.0]$
MEP	$+29.1$	$[28.4, 29.9]$	$[15.8, 42.4]$
E3T	$+43.5$	$[42.7, 44.2]$	$[30.2, 56.7]$
GAMMA	$+44.9$	$[44.1, 45.6]$	$[31.6, 58.2]$

Table 25: **RDSA arbitration control gains by partner impairment type and severity.** For each impairment type, the mean gain is the model-estimated marginal mean with severity fixed at the within-type mean ($\mathrm{sev}_z = 0$), averaged over the baselines, specialists, and layouts. Severity slopes indicate the change in arbitration control gain associated with a one-standard-deviation increase in severity within each impairment type. Brackets indicate 95% confidence intervals.

Impairment type	Mean gain [95% CI]	Severity slope [95% CI]
Valuation	$+23.0$ $[22.1, 23.9]$	-3.24 $[-4.12, -2.36]$
Action Selection	$+28.2$ $[27.3, 29.1]$	-5.24 $[-6.13, -4.36]$
Perception	$+34.5$ $[33.6, 35.4]$	-5.08 $[-5.97, -4.19]$

which represent RDSA's model-estimated arbitration control gain over each baseline or specialist after accounting for variation associated with the other experimental conditions. As shown in the table, the estimated mean gain is strictly positive for all baselines and specialists. Although the 95% confidence intervals become wider in the sensitivity model, they remain entirely above zero across all comparisons. Thus, the positive gain of RDSA is maintained even when additional run-level variation is accounted for. Crucially, RDSA achieves positive gains over both PR and SR, indicating that its performance improvement is not explained by either specialist alone, further supporting the benefit of combining them through arbitration control.

Impairment type and severity. We next examine how arbitration control gain varies across impairment types and how it changes as impairment severity increases. For each impairment type, the mean gain represents the expected arbitration control gain estimated by the primary model under average conditions across the baselines, specialists, and layouts, with severity evaluated at the mean level within that type. The severity slope represents the change in expected gain associated with a one-standard-deviation increase in severity within the corresponding impairment type. As shown in Table 25, the 95% confidence intervals of the mean gains are entirely above zero for all three impairment types, indicating positive average arbitration control gains at mean severity. The estimated mean gain is largest for Perception ($+34.5$), followed by Action Selection ($+28.2$) and Valuation ($+23.0$). The severity slopes are negative for all three impairment types: -3.24 for Valuation, -5.24 for Action Selection, and -5.08 for Perception. Thus, a one-standard-deviation increase in severity within each impairment type is associated with a corresponding decrease in the expected arbitration control gain. The 95% confidence intervals of all three severity slopes remain below zero, supporting a consistent decrease in gain as impairment becomes more severe. This pattern may reflect increasingly less predictable partner behavior at higher severity, which shifts the relative advantage toward the SR-Specialist and reduces the additional gain obtained from arbitration control between PR and SR. Overall, RDSA shows positive arbitration control gains across all three impairment types, while the magnitude of the gain varies by impairment type and decreases with increasing severity.

Table 26: **Estimated mean gains of RDSA by layout.** Estimates are marginalized over comparison methods and impairment types at mean within-type severity.

Layout	Mean gain	95% CI
Cramped Room	$+23.8$	$[22.7, 25.0]$
Coordination Ring	$+23.7$	$[22.6, 24.9]$
Counter Circuit	$+61.6$	$[60.4, 62.7]$
Risky Dualpath	$+21.2$	$[20.0, 22.3]$
Risky Multipath	$+12.4$	$[11.3, 13.5]$

Layout. We next examine how arbitration control gain varies across layouts. For each layout, the mean gain represents the expected arbitration control gain estimated by the primary model under average conditions across the baselines, specialists, and impairment types, with severity evaluated at the mean level within each impairment type. As shown in Table 26, the mean gains and their 95% confidence intervals are entirely above zero for all five layouts, indicating positive average arbitration control gains across layouts. The estimated mean gain is largest in Counter Circuit ($+61.6$), followed by Cramped Room ($+23.8$), Coordination Ring ($+23.7$), Risky Dualpath ($+21.2$), and Risky Multipath ($+12.4$). In Counter Circuit, efficient coordination often involves passing items across the counter, so coordination failures can substantially delay task completion. The larger gain in this layout therefore suggests that the advantage of arbitration control may be amplified when coordination errors are costly. Overall, the positive gains across all layouts indicate that RDSA's arbitration control advantage is not specific to a particular layout.

C.6 Under what conditions does the arbitration control gain increase?

We next investigated the conditions under which RDSA achieves larger arbitration control gains over existing ZSC baselines. We considered two factors: Visibility and Head. Following the behavioral-difference analysis in Section 5.2, Visibility is measured using the JS divergence between the action distributions of an impaired partner and a return matched intermediate checkpoint of Ref. Using a Ref checkpoint with comparable task performance reduces the influence of general skill level differences and allows Visibility to reflect behavioral differences associated with impairment more directly. Higher Visibility therefore indicates a larger behavioral difference between the impaired partner and a Ref checkpoint with similar task performance. Head is defined as the absolute difference in episode return between the PR-Specialist and the SR-Specialist. A larger Head indicates that the choice between the two specialists has a greater effect on performance in that episode.

Using the arbitration control gain Δ defined above as the dependent variable, we standardized Visibility and Head using z-scores and fitted the following LMM:

$$\Delta \sim 1 + \text{baseline} + \text{vis}_z + \text{head}_z + \text{vis}_z \times \text{head}_z + (1 \mid \text{eval_block}).$$

Baseline is included as a fixed effect to account for systematic differences in arbitration control gain across the ZSC baselines. A random intercept is included for each evaluation block to account for the dependence among gain values obtained under the same evaluation condition. The interaction between Visibility and Head evaluates whether the relationship between Visibility and arbitration control gain changes with Head. More specifically, it evaluates whether the association between greater behavioral difference and larger arbitration control gain becomes stronger when the performance difference between the PR-Specialist and SR-Specialist is larger.

Selection of the Visibility reference model. We compute Visibility following the behavioral-difference analysis in Section 5.2. Here, we compared two return-matching criteria for selecting the intermediate checkpoint of Ref used as the Visibility reference. Self-matching uses self-play return, defined as the episode return obtained when a policy cooperates with itself. The impaired partner is therefore matched to the Ref intermediate checkpoint with the most similar self-play return, providing a comparison based on each policy's task performance under its own behavior. Cross-matching instead uses cross-play return against the same fully trained Ref. Specifically, the impaired partner and each Ref intermediate checkpoint are separately paired with the fully trained Ref, and their resulting episode returns are compared. Using the same fixed partner provides a common reference

for comparing the coordination performance of the candidate checkpoints. The self-play returns of the Ref intermediate checkpoints exhibit a bimodal distribution, with 13 of 40 checkpoints receiving zero reward. Consequently, self-matching frequently produces multiple candidate checkpoints with the same return. As shown in Table 27, matching ties occur in 28 of 45 conditions under self-matching, compared with 8 of 45 conditions under cross-matching. We therefore used cross-matched Visibility in the subsequent analyses because it provides a more stable selection of the reference checkpoint.

Main effects and interaction. We next evaluated how Visibility and Head are associated with arbitration control gain and whether their relationships with the gain depend on each other. Because both Visibility and Head are standardized, each coefficient represents the change in expected arbitration control gain associated with a one-standard-deviation increase in the corresponding predictor when the other predictor is at its mean. As shown in Table 28, the coefficient for Visibility is $+0.56$ with a 95% confidence interval of $[-0.07,\ 1.18]$. Thus, when Head is at its mean, a one-standard-deviation increase in Visibility is associated with an estimated increase of 0.56 in arbitration control gain. However, the confidence interval includes zero, so this association is not clearly distinguishable from zero. In contrast, the coefficient for Head is $+17.61$ with a 95% confidence interval of $[16.98,\ 18.23]$. When Visibility is at its mean, a one-standard-deviation increase in Head is therefore associated with an estimated increase of 17.61 in arbitration control gain. This indicates that RDSA's arbitration control advantage tends to increase as the performance difference between the PR-Specialist and SR-Specialist becomes larger. The Visibility $\times$ Head interaction is also positive, with a coefficient of $+2.04$ and a 95% confidence interval of $[1.45,\ 2.63]$. This positive interaction indicates that the relationship between Visibility and arbitration control gain becomes more positive as Head increases. Equivalently, the positive relationship between Head and arbitration control gain becomes stronger as Visibility increases. Thus, larger arbitration control gains tend to occur when impairment related behavioral differences are more pronounced and the performance difference between the two specialists is larger. We evaluate this interaction more directly through simple-slope analyses below.

Evaluation of the interaction between Visibility and Head. To interpret the positive Visibility $\times$ Head interaction in more detail, we estimated simple slopes from the fitted LMM. A simple slope represents the change in expected arbitration control gain associated with a one-standard-deviation increase in one predictor at a specified level of the other predictor (Preacher et al., 2006). We first evaluate the relationship between Visibility and arbitration control gain at three levels of Head: one standard deviation below the mean, the mean, and one standard deviation above the mean. We then evaluate the relationship between Head and arbitration control gain at the corresponding three levels of Visibility. As shown in Table 29, the relationship between Visibility and arbitration control gain changes substantially with Head. When Head is one standard deviation below its mean, the Visibility slope is -1.48 with a 95% confidence interval of $[-2.48,\ -0.49]$. Thus, when the performance difference between the two specialists is small, higher Visibility is associated with lower arbitration control gain. At mean Head, the Visibility slope is $+0.56$, and its confidence interval includes zero, indicating no clear association between Visibility and arbitration control gain at this level. When Head is one standard deviation above its mean, the Visibility slope increases to $+2.60$ with a 95% confidence interval of $[1.89,\ 3.30]$. Thus, when the performance difference between the two specialists is large, greater behavioral differences are associated with larger arbitration control gains. Table 30 provides the complementary evaluation. The Head slope is positive at all three Visibility levels, increasing from $+15.56$ at one standard deviation below mean Visibility to $+17.61$ at mean Visibility and $+19.65$ at one standard deviation above mean Visibility. Thus, larger performance differences between the PR-Specialist and SR-Specialist are associated with larger arbitration control gains across all three Visibility levels, and this relationship becomes stronger as Visibility increases. Taken together, the relationship between Visibility and arbitration control gain depends on Head. Larger arbitration control gains are particularly associated with conditions in which impairment-related behavioral differences are more pronounced and the performance difference between the two specialists is large. These results indicate that the advantage of arbitration control is strongly associated with how much the choice between the two specialists matters for performance, while its relationship with behavioral difference (Visibility) depends on the magnitude of this performance difference (Head).

Table 27: Frequency of matching ties under the two return matching criteria.

Matching criterion	Conditions with matching ties
Self-matched	28 / 45
Cross-matched	8 / 45

Table 28: **Fixed-effect estimates for Visibility, Head, and their interaction using cross-matched Visibility.** Visibility and Head are standardized using z-scores. Each main-effect estimate represents the change in expected arbitration control gain per 1 SD increase in the corresponding predictor when the other predictor is at its mean. The interaction estimate represents how this relationship changes with the other predictor. Brackets indicate 95% confidence intervals.

Term	Fixed effect estimate	95% CI
Visibility	$+0.56$	$[-0.07, +1.18]$
Head	$+17.61$	$[+16.98, +18.23]$
Visibility $\times$ Head	$+2.04$	$[+1.45, +2.63]$

Table 29: **Simple slopes of Visibility at three Head levels using cross-matched Visibility.** Head is evaluated at one standard deviation below the mean, the mean, and one standard deviation above the mean. Each slope represents the change in expected arbitration control gain per 1 SD increase in Visibility at the corresponding Head level. Brackets indicate 95% confidence intervals.

Head level	Visibility slope	95% CI
Mean -1 SD	-1.48	$[-2.48, -0.49]$
Mean	$+0.56$	$[-0.07, +1.18]$
Mean $+1$ SD	$+2.60$	$[+1.89, +3.30]$

Table 30: **Simple slopes of Head at three Visibility levels using cross-matched Visibility.** Visibility is evaluated at one standard deviation below the mean, the mean, and one standard deviation above the mean. Each slope represents the change in expected arbitration control gain per 1 SD increase in Head at the corresponding Visibility level. Brackets indicate 95% confidence intervals.

Visibility level	Head slope	95% CI
Mean -1 SD	$+15.56$	$[+14.85, +16.28]$
Mean	$+17.61$	$[+16.98, +18.23]$
Mean $+1$ SD	$+19.65$	$[+18.66, +20.63]$

Summary We additionally evaluated RDSA with an impaired behavioral cloning (BC) human proxy commonly used in ZSC evaluation. RDSA achieved the highest mean episode return across all evaluated BC partner conditions, with significant improvements over the ZSC baselines in most paired comparisons. These results further support the robustness of RDSA across different partner models and impairment settings.

Additional analyses examined how arbitration control responds to partner behavior and impairment severity. Mean p_t decreased as the behavioral difference score increased, and the same pattern was observed as impairment severity increased. These results show that arbitration control changes the relative influence of the PR- and SR-Specialists in response to both observed partner behavior and the severity of decision process impairment. Together with the performance comparisons against the individual specialists and static mixtures, these results support the contribution of arbitration control to RDSA's performance.

We further used linear mixed-effects model (LMM) to analyze arbitration control gain across layouts, impairment types, and severity. At mean within-type severity, the model-estimated marginal means were positive across all evaluated layouts and impairment types, with 95% confidence intervals entirely above zero. The magnitude of these gains decreased with increasing severity, indicating

that the benefit of arbitration control depends on the severity of partner impairment. We then examined how gain relates to behavioral difference (Visibility) and the performance difference between the PR- and SR-Specialists (Head). Performance difference showed a strong positive relationship with gain. The positive interaction indicates that the relationship between behavioral difference and arbitration control gain becomes more positive as performance difference increases. These results indicate that arbitration control provides larger gains when the choice between the two specialists has a greater effect on performance, while the relationship between behavioral difference and gain depends on the magnitude of this performance difference.

D BASELINES

D.1 BASELINES

Below, we describe the implementation details of each baseline, including training settings shared across methods, baseline-specific configurations, and deviations from the original implementations.

FCP and MEP. FCP constructs a partner population from checkpoints of self-play policies at different stages of training (Strouse et al., 2021), whereas MEP trains a diverse population using a population entropy objective before training a cooperator against it (Zhao et al., 2023). We implemented both methods using the HSP codebase and followed the experimental configurations used in the HSP paper. For each method, we trained 12 policies in the first stage and retained their initial, intermediate, and final checkpoints, resulting in a population of 36 policies. Accordingly, the population sizes differ from those used in the original FCP and MEP implementations. Following the HSP configuration, we used uniform population sampling for MEP rather than prioritized sampling. The population entropy coefficient was set to 0.01, following the original MEP study. The common first- and second-stage training settings are reported in Tables 31a and 31b, respectively.

HSP. HSP represents partner preferences using hidden utility functions defined over event-based features and trains a cooperator against a diverse partner population (Yu et al., 2023). Following the original two-stage procedure, we trained 36 biased policies using independently sampled hidden utilities. To reduce behavioral redundancy, we followed HSP's greedy selection procedure and retained 18 policies that maximized differences in their expected event counts. We combined these policies with 18 MEP policies obtained from the initial, intermediate, and final checkpoints of six MEP policies, resulting in a population of 36 policies. Because HSP does not provide configurations for Risky Multipath and Risky Dualpath, we applied the Many Orders configuration, which covers a broad range of event-based preferences, to both layouts. The common first- and second-stage training settings are reported in Tables 31a and 31b, respectively.

GAMMA. GAMMA learns a variational autoencoder (VAE) from joint coordination trajectories and uses the learned generative model to produce partner behavior during cooperator training (Liang et al., 2024). We followed the implementation and training settings provided in the original GAMMA paper and codebase. The VAE training data were generated using FCP policies that had already been trained for the FCP baseline. Following the original GAMMA setting, we selected eight FCP policies and retained their initial, intermediate, and final checkpoints, resulting in a data-generating population of 24 policies. The first eight policies in the fixed training order were used without performance-based selection. Joint trajectories were then collected by sampling pairs from this population and used to train the VAE. The architecture and training configuration of the VAE are reported in Table 31c.

E3T. For a fair comparison, we included E3T because it models partner behavior during interaction (Yan et al., 2023). E3T is an end-to-end method that constructs its training partner by mixing the ego policy with a uniform random action distribution. The mixture coefficient ϵ controls the contribution of this distribution. We fixed $\epsilon = 0.3$ across all layouts. E3T also includes a partner modeling module that predicts partner actions from recent interaction history. We otherwise followed the E3T paper and codebase for its network architecture and PPO configuration. To improve training stability, we additionally applied ValueNorm to E3T. We also adjusted the reward-shaping horizon so that its coefficient was annealed over the full 100M-step training period.

Table 31: **Hyperparameters used in the first stage, second stage, and VAE models.**

Hyperparameter	Value
CNN kernels	$[3\times3,\ 3\times3,\ 3\times3]$
CNN channels	$[32,\ 64,\ 32]$
Hidden layer size	64
Activation function	ReLU
Entropy coefficient	0.01
Environment steps	10M
Learning rate	5×10^{-4}
Episode length	200
Discount factor γ	0.99
Weight decay	0
Parallel environment threads	100
Reward shaping horizon	100M
Optimizer	Adam ($\epsilon = 10^{-5}$)
PPO epochs	15
Batch	40,000

(a) Hyperparameters used in the first stage.

Hyperparameter	Value
Architecture	1 GRU layer
Hidden layer size	64
Activation function	ReLU
Entropy coefficient	0.01
Environment steps	100M
Learning rate	5×10^{-4}
Discount factor γ	0.99
Weight decay	0
Parallel environment threads	300
Reward shaping horizon	100M
Optimizer	Adam ($\epsilon = 10^{-5}$)
PPO epochs	15
Policy pool size	36

(b) Common hyperparameters used in the second stage.

Hyperparameter	Value
CNN kernels	$[3\times3,\ 3\times3,\ 3\times3]$
CNN channels	$[32,\ 64,\ 32]$
Hidden layer size	256
Recurrent layer size	256
Activation function	ReLU
Weight decay	0.0001
Parallel environments	200
Episode length	200
Epochs	500
Chunk length	100
Learning rate	0.0005
KL penalty coefficient β	$0 \rightarrow 1$
Latent variable dimension	16

(c) Hyperparameters for VAE models.