

Diet: Interpreting Ethical Tradeoffs in Disrupted Sociotechnical Systems

Divya Sundaresan¹, Gordon Briggs², Munindar P. Singh¹

¹North Carolina State University

²US Naval Research Laboratory

dsundar3@ncsu.edu, gordon.m.briggs.civ@us.navy.mil, mpsingh@ncsu.edu

Abstract

Sociotechnical systems comprise human stakeholders and autonomous agents, where stakeholders may prioritize values (such as freedom, security, and welfare) differently. Social norms operationalize these values by specifying expectations about acceptable behavior. Disruptions in sociotechnical systems create situations in which these stakeholder expectations cannot always be simultaneously satisfied, forcing autonomous agents to make ethically consequential decisions and sometimes deviate from otherwise desirable behavior.

We investigate this problem in a multiagent ride-sharing domain in which autonomous vehicles respond to stochastic disruptions such as vehicle breakdowns. Our contributions are threefold: (1) we formulate disruption response as a decentralized multiagent reinforcement learning problem grounded in a transit setting; (2) we analyze reward functions aligned with utilitarian and Rawlsian ethical theories, alongside a profit-maximizing baseline, and demonstrate how ethical objectives shape disruption-response behavior and rider well-being; and (3) we introduce an analysis framework that interprets learned policies by extracting their implicit priorities over ethically relevant system criteria informed by social norms. We find that ethical reward design improves aggregate well-being, induces distinct disruption-response behaviors under different ethical reward formulations, and may concentrate disruption-handling load on ethically motivated agents in heterogeneous populations.

Code — <https://github.com/divtosz/diet>

1 Introduction

A sociotechnical system (STS) comprises stakeholders (people, organizations, and society at large), AI agents, and resources, and is controlled socially through norms (Singh 2014). Ethical questions arise when an STS must reconcile competing stakeholder expectations in allocating limited resources. We illustrate these ideas in the domain of civic transit, via a ride-sharing service in which self-driving vehicles provide on-demand transportation. Such a service must continually allocate a fleet of vehicles to incoming trip requests while managing tradeoffs among efficiency, fairness, and service quality. Each vehicle is controlled by an agent. Riders experience the consequences of the agents' decisions

through waiting times, detours, and the overall reliability of the service. Transit thus provides a concrete and socially consequential setting in which ethical considerations naturally arise from everyday operational decisions.

We define a *disruption* in an STS as an exogenous event that alters the feasibility, timing, or cost of planned actions, thereby forcing the STS to reallocate resources among stakeholders. In our setting, disruptions include vehicle breakdowns, road closures, congestion, or other causes of delays that invalidate previously acceptable plans. Disruptions carry ethical import because they create situations in which not all stakeholder expectations can be satisfied, making tradeoffs unavoidable. As a result, resources (here, vehicles) must be reallocated, which may require adjusting how service is allocated across riders.

Norms operationalize stakeholder values: they encode expectations of fairness, equitable treatment, and other socially meaningful considerations (Bench-Capon and Modgil 2017; Ajmeri et al. 2020; Serramia et al. 2023). For example, a norm may specify that a vehicle should give precedence to riders with time-critical trips or avoid actions that would unduly delay riders currently being served. A responsible agent must both respect applicable norms and deviate from them when blind adherence may lead to diminished outcomes (Singh and Singh 2023; Briggs et al. 2022).

Prior work on ethical decision making in multiagent systems explores how conflicting norms and ethical objectives can be resolved through explicit norm reasoning architectures (Ajmeri et al. 2017, 2020), logical formalisms for ethical governance and norm compliance (Neufeld et al. 2022; Singh and Singh 2023), policy-selection approaches that balance competing objectives (Noothigattu et al. 2019), and reinforcement learning (RL) methods grounded in ethical reward formulations or moral theories (Wu and Lin 2018; Rodriguez-Soto et al. 2022; Tennant, Hailes, and Musolesi 2023; Woodgate, Marshall, and Ajmeri 2025). When decision policies are learned, the ethical tradeoffs they encode may be implicit in the learned behavior. We present an analysis framework for interpreting the priorities over ethical system criteria implicit in those policies, complementing prior efforts to interpret learned behavior (Shea-Blymyer and Abbas 2022; Mao et al. 2023).

We refer to our approach as Diet, a mnemonic name reflecting our focus on interpreting ethical tradeoffs in dis-

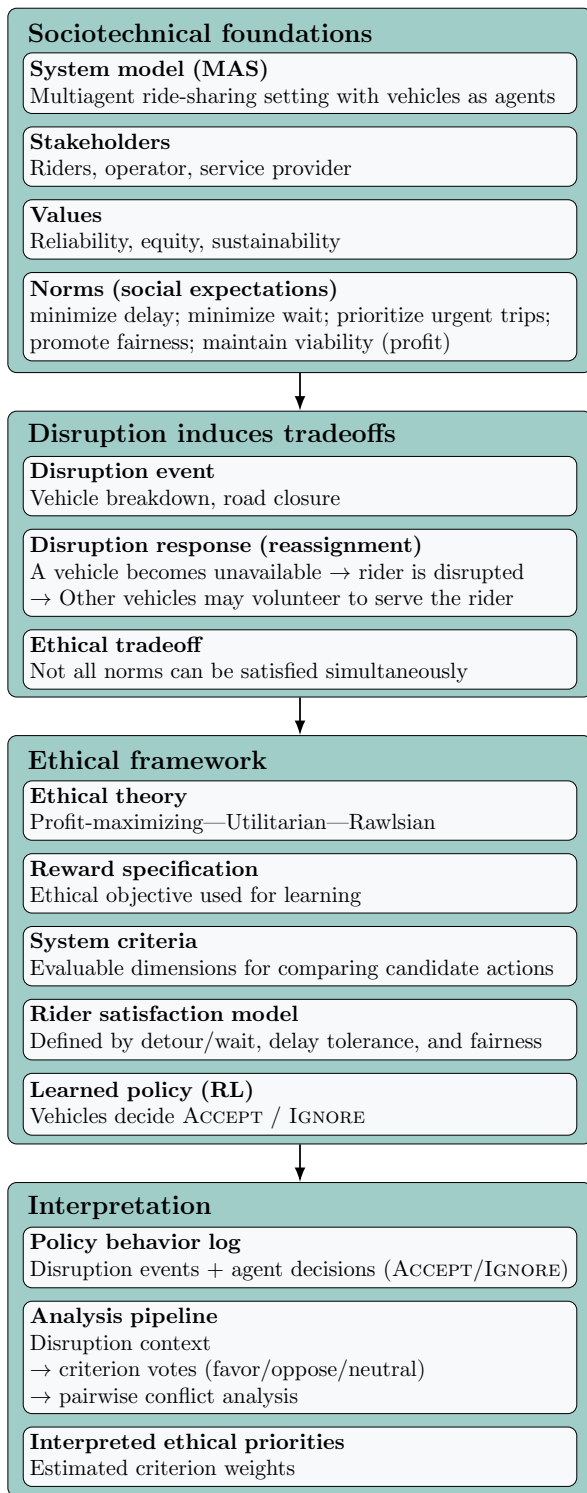


Figure 1: Overview of Diet. We model a multiagent ride-sharing system situated within a sociotechnical context of stakeholders, values, and norms. Disruptions induce ethical tradeoffs in which not all norms can be simultaneously satisfied. Ethical theories shape the reward specifications used for learning, and policies are interpreted by extracting implicit priorities over system criteria from observed behavior.

rupted sociotechnical systems. Figure 1 shows an overview of Diet. We identify the following research questions:

RQ^{Behavior}: How do different ethical theories, operationalized through reward functions, shape the behaviors agents learn when disruptions force ethical tradeoffs that affect multiple stakeholders?

Significance. Disruptions force agents to make decisions with ethically salient consequences (e.g., reallocating service to assist one rider may delay another). Understanding how reward formulations shape learned behavior under disruption is essential for designing STSs that balance effectiveness with ethical considerations.

Novelty. Prior work studies learning ethical behavior through reward design under routine operating conditions whereas we focus on disruption. By comparing profit-maximizing, utilitarian, and Rawlsian rewards, we show how ethical objectives produce systematic differences in agent behavior and stakeholder well-being.

RQ^{Interpretation}: How can we reveal how agents resolve ethical tradeoffs by extracting interpretable priorities from trained models?

Significance. In learning-based approaches to ethical decision making, ethical considerations are often encoded implicitly through reward functions, making agent decisions difficult to understand or justify. This lack of interpretability poses challenges for auditing, governance, and trust in deployed sociotechnical systems.

Novelty. We introduce an analysis pipeline that extracts priorities over ethical system criteria from learned policies. Rather than evaluating policies solely by outcomes, our approach reveals how competing considerations are weighted and traded off in decision making.

RQ^{Heterogeneity}: How does the composition of heterogeneous agent populations trained under different reward formulations affect the distribution of disruption-handling load and overall stakeholder well-being?

Significance. Real-world sociotechnical systems rarely consist of agents optimized for identical system objectives. Understanding how agents trained under different reward formulations interact is therefore essential for anticipating system-level effects of ethical policy adoption.

Novelty. We empirically study heterogeneous agent populations and show how differences in objectives produce asymmetric distributions of disruption-handling load while still improving overall stakeholder well-being.

Scope. We scope our investigation into ethical decision making in a multiagent transit system operating under disruption. Agents make local disruption-response decisions, and our analysis focuses on how ethical reward formulations shape these learned policies and how they prioritize ethical system criteria. We do not model explicit norm enforcement, sanctioning, accountability, or emergent coordination among agents, enabling a controlled comparison of individual decision policies under disruption. Population-level effects observed in heterogeneous settings arise from the interaction of multiple agents following such locally learned policies under shared constraints.

2 Framework

We model a disruption-driven ride-sharing environment as a decentralized RL problem.

2.1 Environment Overview

We model a grid-world ride-sharing environment with n_v autonomous vehicles, each controlled by an independent decision-making agent. Trip requests arrive dynamically, and vehicles move one Manhattan step per time unit toward their assigned destinations. A disruption (here, a vehicle breaking down) occurs stochastically: with probability p_{break} at each time step, a vehicle becomes unavailable for several ticks, forcing reassignment of its passenger to another vehicle. Eligible vehicles (those that are active and not at maximum capacity) may volunteer to serve the disrupted request. At each disruption, each agent observes the features summarized in Table 1 and takes one of two actions: $a \in \{\text{ACCEPT}, \text{IGNORE}\}$. Ultimately, a single vehicle is assigned to any disrupted request. During training, all agents share a common policy, trained using experience collected in parallel across the system.

For our experiments, we use a 20×20 grid with $n_v = 10$ vehicles. At each discrete *world step*, vehicles move and serve riders, and a disruption may occur with probability $p_{\text{break}} = 0.05$. Disrupted vehicles remain inactive for 2 to 4 subsequent world steps, sampled uniformly from $\{2, 3, 4\}$.

Agents act only at disruption events. Each *RL decision step* therefore corresponds to a disruption, at which point an agent decides whether to accept or ignore a reassignment request. Between disruptions, the environment advances without agent intervention. Episodes run for 1,000 world steps, during which multiple RL decision steps occur depending on the realized disruptions.

Planner. The planner computes the routing consequences of candidate reassignment actions, isolating ethical disruption-response decisions from low-level route optimization. The planner locally simulates the insertion of the disrupted rider into an eligible vehicle’s current route without modifying other vehicles’ routes. For a vehicle currently serving rider C , the disrupted rider D may be served either before or after completing C ’s trip; no additional route reordering is performed.

We define burden as the maximum of the detour imposed on the current rider and the disrupted rider’s wait until pickup. For each eligible vehicle, the planner selects the insertion minimizing this burden. Agents then choose whether to ACCEPT or IGNORE. If one or more agents volunteer, the disrupted rider is assigned to the volunteering vehicle with the lowest resulting burden, with ties broken randomly.

2.2 Values

Stakeholders’ *values* are broad motivational goals that justify what the system ought to prioritize (Montes and Sierra 2022). In our transit setting, we focus on *institutional values* relevant to service provision under disruption. Specifically, we consider: (1) **reliability**—providing predictable and timely service; (2) **equity**—treating riders fairly over time and avoiding systematic disadvantage across riders; and

Feature	Description
Busy or inactive	Whether the vehicle is already serving two trips or is itself disrupted.
Burden	Combined detour/wait impact if the agent accepts the disrupted rider (small/medium/large).
Historical fairness	Prior service quality experienced by the current and disrupted riders ($\{-1, 0, +1\}$).
Trip urgencies	Urgency levels of the current and disrupted riders (nonurgent/urgent).
Profit or loss category	Discretized operational profit impact of accepting the disrupted rider.
Alternative availability	Whether alternative lower-burden vehicles are few or many.

Table 1: Observation features available to each agent during disruption-response decisions.

(3) **sustainability**—maintaining long-run operational and financial viability of the service. We operationalize these values via social norms that can be applied to concrete disruption decisions.

2.3 System Criteria

We formalize ethically and operationally relevant aspects of disruption-response decisions as *system criteria*. System criteria are informed by social norms and provide dimensions for evaluating candidate actions. A criterion evaluates how a candidate action affects an ethically or operationally relevant aspect of system performance. A criterion may favor an action, oppose it, or neither. When criteria favor conflicting actions, the agent must trade off among them when selecting an action. In our transit setting, we consider five criteria:

Delay (reduction): minimize additional delay imposed on current riders;

Wait (reduction): minimize waiting time experienced by disrupted riders;

Fairness: prioritize riders who have historically received worse service;

Urgency: prioritize riders with time-critical trips;

(Operational) Viability: maintain long-term operational sustainability, modeled via profit and loss.

2.4 Rider Satisfaction Model

Each disruption decision affects two riders: the *current rider* C who may be delayed by a detour if the disrupted rider is picked up by that vehicle, and the *disrupted rider* D who waits to be picked up.

Given detour d for C and wait w for D , with historical fairness $H_C, H_D \in \{-1, 0, +1\}$ and tolerances for delay (based on trip urgency) τ_C, τ_D , we model delay-based satisfaction as a monotonically decreasing function of delay, consistent with classic models (Small 1982). We adopt an exponential form to capture diminishing sensitivity to increasing delay, reflecting prospect theory (Kahneman and

Tversky 1979). For rider $i \in \{C, D\}$, we compute delay-based satisfaction S_i^{delay} as follows, where η adjusts the influence of historical fairness on delay tolerance.

$$S_i^{\text{delay}} = \exp\left(-\frac{x_i}{\tau_i(1 + \eta H_i)}\right), \quad x_C = d, \quad x_D = w.$$

We augment the current rider C 's satisfaction with a small *altruistic warm glow* g to reflect the finding that modest prosocial acts can generate intrinsic “warm-glow” satisfaction (Andreoni 1990), especially for prosocial personality types (Sundaresan et al. 2025). At the same time, large sacrifices diminish well-being (Hobfoll 1989). Accordingly, we model g as an inverse-U function of detour that peaks for moderate detours and tapers as detours become excessive.

$$g(d; \tau_C) = \frac{d}{\tau_C} \exp\left(1 - \frac{d}{\tau_C}\right),$$

Figure 2 shows how the delay-based satisfaction and altruistic warm glow vary with delay.

The bonus g is applied only when the disrupted rider has greater need than the current rider (e.g., $u_D > u_C$), captured by a binary indicator $\mathbb{I}_{\text{help}} \in \{0, 1\}$, where $\mathbb{I}_{\text{help}} = 1$ if this condition holds and 0 otherwise. λ is a scaling coefficient.

The resulting satisfaction values are:

$$S_C = \min\left(S_C^{\text{delay}} + \lambda \mathbb{I}_{\text{help}} g(d; \tau_C), 1\right),$$

$$S_D = S_D^{\text{delay}}.$$

2.5 Experimental Setup

We evaluate the learned profit-maximizing, utilitarian, and Rawlsian policies with respect to (1) rider well-being and (2) system performance, measured by induced detour for the current rider, waiting time for the disrupted rider, and monetary profit, as well as (3) the implicit criterion priorities expressed by each policy. For each reward specification, we trained three independently seeded PPO policies (seeds 0–2) for 200 iterations using a shared two-layer fully connected network with ReLU activations (64 units per layer). Well-being and system performance results are averaged over the three training seeds, with 50 evaluation episodes per seed (150 total episodes per policy). For policy interpretation and heterogeneous population analyses, we evaluate a single representative trained policy (seed = 1) per reward specification.

3 Agent Behavior (RQ_{Behavior})

3.1 Ethical Reward Functions

We compare three reward functions: one profit-maximizing baseline and two based on established ethical theories. Each defines an evaluation function over action outcomes.

Profit-maximizing baseline. Rewards equal the normalized profit or loss resulting from a given agent accepting the disrupted request. For each disruption event, a profit value is sampled uniformly from $[-1, 1]$ for each candidate vehicle:

$$r_t^{\text{profit}} \in [-1, 1].$$

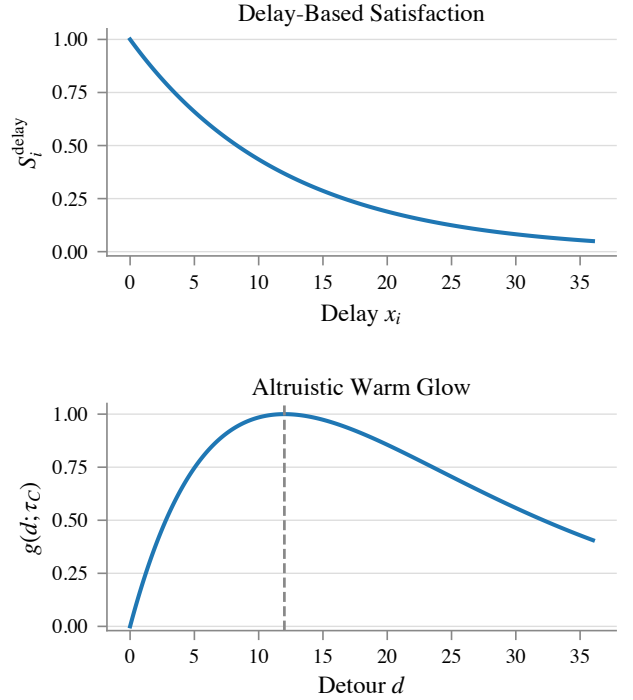


Figure 2: Components of rider satisfaction. Delay-based satisfaction S_i^{delay} decreases with experienced delay, whereas warm glow $g(d; \tau_C)$ peaks at $d = \tau_C$ and then declines.

Utilitarian. Following the utilitarian principle of maximizing aggregate welfare (Mill 1863), the ethical value of an action is the total satisfaction of all affected riders in the resulting outcome. We define the utilitarian evaluation function as:

$$J_{\text{util}}(a) = S_C^a + S_D^a.$$

Rawlsian. Following Rawls’ theory of justice (Rawls 1999), the ethical value of an action is given by the satisfaction of the worst-off individual in the resulting outcome. The corresponding evaluation function is:

$$J_{\text{rawls}}(a) = \min(S_C^a, S_D^a).$$

For both ethical reward definitions, we compute a counterfactual difference (Wolpert and Tumer 2001) by comparing the chosen action a with the alternative action $\bar{a}$ that the agent could have taken.

$$\Delta J^{\text{theory}} = J^{\text{theory}}(a) - J^{\text{theory}}(\bar{a}).$$

This difference is scaled with a responsibility factor R_t , which reflects how well suited the deciding vehicle is to serve the disrupted request relative to other vehicles. The resulting value is converted into an RL reward using a threshold parameter $\delta > 0$ to distinguish clearly positive, clearly negative, and near-neutral effects. The resulting bounded reward is defined as:

$$r_t^{\text{theory}} = \begin{cases} +1, & R_t \Delta J \geq \delta, \\ -1, & R_t \Delta J \leq -\delta, \\ \frac{R_t \Delta J}{\delta}, & \text{otherwise.} \end{cases}$$

Metric	Definition
Geometric mean well-being	$WB_{GM} = \sqrt{S_C S_D}$
Minimum well-being	$WB_{\min} = \min(S_C, S_D)$
Total well-being	$S_C + S_D$
Satisfaction gap	$\Delta S = S_C - S_D $

Table 2: Evaluation metrics derived from rider satisfaction.

3.2 Evaluation Metrics

We evaluate disruption-response decisions using the metrics summarized in Table 2. Here, S_C and S_D denote individual rider satisfaction, well-being denotes aggregate measures derived from them, and the satisfaction gap captures inequality between riders.

3.3 Well-Being and System Performance

To address RQ_{Behavior} , we examine how the three reward functions influence policy behavior under disruption. Table 3 summarizes well-being and operational outcomes across three training seeds with 50 evaluation episodes per seed (150 total episodes per policy). Table 4 reports the corresponding effect sizes. Relative to the profit-maximizing baseline, ethical policies improve total well-being ($d \approx 2.35$), minimum well-being ($d \approx 2.68$), and geometric mean well-being ($d \approx 2.76$), while reducing the satisfaction gap between riders ($d \approx -1.7$).

Operationally, ethical policies reduce detours and waiting times, but also reduce profit: the profit-maximizing policy achieves the highest profit, whereas the ethical policies incur a slight loss on average ($d \approx -8$). Differences between utilitarian and Rawlsian policies are minimal, suggesting that aggregate metrics alone may be insufficient to distinguish between policies derived from different ethical theories.

3.4 Robustness to Changes in Disruption Dynamics

To assess sensitivity to environment parameterization, we evaluated one representative trained policy from each reward specification (seed = 1) under moderate shifts in disruption dynamics, without retraining. We varied the probability of disruption occurrence while holding repair duration fixed, and varied repair duration while holding disruption probability fixed. Tables 5 and 6 report the resulting well-being and operational metrics.

Across all tested settings, the qualitative ordering of policies remains unchanged. Utilitarian and Rawlsian policies consistently outperform the profit-maximizing baseline on total, minimum, and geometric mean well-being, while the profit-maximizing policy achieves the highest profit. Ethical policies also continue to reduce detour and waiting time. Although absolute values shift as disruptions become more frequent or longer-lasting, no tested setting reverses the main experiment conclusions, indicating robustness to moderate variation in disruption dynamics.

4 Policy Interpretation ($RQ_{\text{Interpretation}}$)

4.1 Extracting Criterion Priorities from Policies

To characterize how policies resolve ethical tradeoffs, we introduce a criterion analysis pipeline that recovers implicit ethical priority structure from observed behavior. We identify states in which the various criteria yield opposing action recommendations, estimate pairwise win probabilities between criteria, and aggregate these pairwise conflict outcomes to infer relative criterion priorities.

Signature abstraction. For each disruption, we construct its *signature*, a reduced six-dimensional feature representation that contains the state features required to determine how each criterion evaluates the candidate actions. Here, d_C and w_D denote the current rider’s detour and the disrupted rider’s wait time; u_C and u_D their urgencies; H_C and H_D their historical fairness values; $\Delta\pi$ the profit difference between accepting and ignoring the request; and candidate suitability whether the vehicle is among the better-suited candidates for serving the request.

$$\text{sig} = \left(\mathbf{1}[d_C > 0], \mathbf{1}[w_D > 0], \text{sign}(u_D - u_C), \right. \\ \left. \text{sign}(H_C - H_D), \text{sign}(\Delta\pi), \text{candidate suitability} \right)$$

Criterion votes. Given a signature, we encode the action preference of each criterion using a symbolic vote vector:

$$v = (v_{\text{delay}}, v_{\text{wait}}, v_{\text{urg}}, v_{\text{fair}}, v_{\text{prof}}) \in \{-1, 0, +1\}^5,$$

where each entry indicates whether the criterion favors ACCEPT (+1), IGNORE (−1), or is neutral (0) in that situation.

Pairwise conflict matrices. We focus on situations in which two criteria issue opposing action recommendations under the same signature. For each ordered pair of criteria (c_i, c_j) , we estimate the empirical probability that the policy follows criterion c_i ’s recommendation over criterion c_j ’s when they are in conflict. Aggregating across all such signatures yields a pairwise conflict matrix, where each entry captures how often one criterion prevails over another. These matrices provide an interpretable view of how learned policies resolve ethical tradeoffs.

Conflict-derived criterion weights. To compute each criterion’s overall priority, we fit a Bradley-Terry model (Bradley and Terry 1952) to the empirical win probabilities $\{P(c_i \succ c_j)\}$. This model estimates a latent strength parameter for each criterion such that criteria with larger strengths are more likely to prevail in pairwise conflicts. Formally,

$$P(c_i \succ c_j) \approx \sigma(\theta_i - \theta_j),$$

where θ_i and θ_j denote the latent strengths of criteria c_i and c_j , respectively, and $\sigma(\cdot)$ denotes the logistic function. We normalize the resulting strengths using a softmax transformation to obtain conflict-derived criterion weights.

4.2 Learned Criterion Priorities

We evaluate one policy (seed = 1) for each reward specification, which exhibited representative performance across the three trained seeds.

Policy	S_C	S_D	S_C+S_D	$WB_{\min}$	WB_{GM}	ΔS	Detour _C	Wait _D	Profit
Profit-max	0.79	0.38	1.17	0.32	0.49	0.53	3.45	9.55	0.50
Utilitarian	0.81	0.47	1.28	0.41	0.57	0.46	2.66	7.14	0.01
Rawlsian	0.81	0.47	1.28	0.41	0.58	0.46	2.54	7.05	-0.02

Table 3: Well-being and operational outcomes across policies, averaged over three independent training seeds (50 evaluation episodes per seed, 150 total episodes per policy). Results are stable across seeds, with low variance across all metrics (standard deviations ≤ 0.01 for well-being metrics and ≤ 0.07 for operational metrics). For ΔS , Detour_C, and Wait_D, lower is better; for all remaining metrics, higher is better. Utilitarian and Rawlsian policies improve total, minimum, and geometric mean well-being relative to the profit-maximizing baseline, while also reducing the satisfaction gap, detour, and waiting time. These gains come at the cost of reduced profit relative to the profit-maximizing policy.

Metric	Utilitarian vs Profit	Rawlsian vs Profit
S_C	0.60	0.63
S_D	2.32	2.25
S_C+S_D	2.36	2.34
$WB_{\min}$	2.76	2.62
WB_{GM}	2.78	2.73
ΔS	-1.73	-1.68
Detour _C	-1.31	-1.60
Wait _D	-3.52	-3.58
Profit	-7.46	-8.55

Table 4: Episode-level effect sizes (Cohen’s d) relative to the profit-maximizing baseline, computed from episode-level metrics pooled across three training seeds (50 evaluation episodes per seed, 150 total episodes per policy). Positive values indicate higher metric values under the ethical policy. For ΔS , Detour_C, and Wait_D, lower values indicate improvement. Utilitarian and Rawlsian policies increase disrupted-rider welfare and minimum well-being ($d \approx 2.5$), reduce waiting time ($d \approx -3.5$), and moderately increase current-rider welfare ($d \approx 0.6$). Both substantially reduce operator profit ($d \approx -8$).

For each criterion conflict, we record which criterion’s preferred action matches the policy’s chosen action. From these pairwise comparisons, we estimate empirical dominance probabilities, where each entry of the conflict matrix (i, j) denotes the proportion of conflicts in which row criterion i prevails over column criterion j . Figures 3–5 visualize the resulting criterion dominance patterns for each policy.

Clear differences emerge in how each policy resolves criterion conflicts:

- **Profit-maximizing:** Profit dominates all other criteria, indicating that disruption-response decisions are driven entirely by profitability considerations.
- **Utilitarian:** Delay is the dominant criterion, whereas wait is least influential, indicating a preference for minimizing detour to riders who are currently being served.
- **Rawlsian:** Urgency and fairness dominate, whereas delay receives comparatively low priority, reflecting an emphasis on protecting disadvantaged or time-critical riders even at the cost of additional delay to riders who are currently being served.

We next quantify these dominance patterns using the

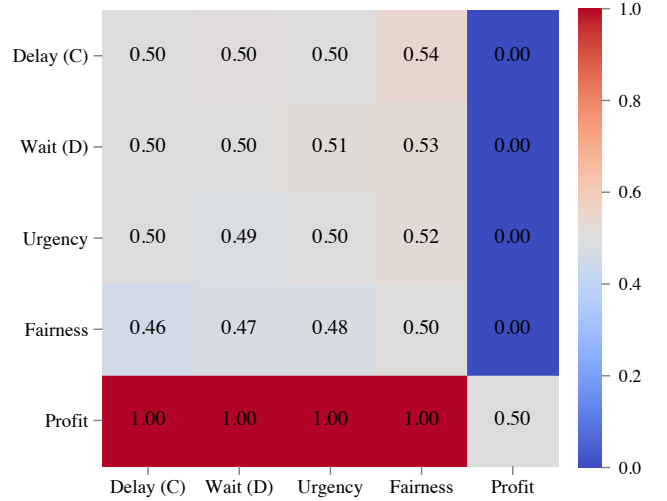


Figure 3: Criterion conflict matrix for the profit-maximizing policy. Entry (i, j) denotes the fraction of conflict cases in which row criterion i prevails over column criterion j (1.0 = 100% dominance). Whenever profit conflicts with any other criterion, the policy selects the profit-favoring action.

Bradley-Terry procedure described above. Figure 6 shows the resulting normalized criterion weights, which reinforce the conflict-level patterns above:

- **Profit-maximizing:** Assigns nearly all weight to profit, with negligible weight on all other criteria.
- **Utilitarian:** Places greatest weight on delay, reflecting a preference for minimizing detour to riders who are currently being served.
- **Rawlsian:** Places greatest weight on urgency and fairness, reflecting a stronger emphasis on riders with greater need or disadvantage.

Together, these analyses show that different ethical reward formulations induce distinct criterion priorities, providing interpretable insight into how learned policies resolve ethical tradeoffs under disruption.

5 Heterogeneous Populations ($RQ_{\text{Heterogeneity}}$)

All preceding experiments assume *homogeneous* populations in which every agent follows the same learned pol-

p_{break}	Policy	$S_C + S_D$	$WB_{\min}$	WB_{GM}	ΔS	Detour _C	Wait _D	Profit
0.02	Profit-max	1.18	0.33	0.50	0.53	3.39	9.42	0.52
0.02	Utilitarian	1.29	0.42	0.58	0.46	2.60	7.10	0.06
0.02	Rawlsian	1.29	0.42	0.58	0.46	2.51	7.05	0.02
0.05	Profit-max	1.17	0.32	0.49	0.53	3.46	9.55	0.50
0.05	Utilitarian	1.28	0.41	0.57	0.46	2.67	7.16	0.01
0.05	Rawlsian	1.28	0.41	0.58	0.46	2.58	7.05	-0.02
0.10	Profit-max	1.15	0.31	0.48	0.53	3.77	9.78	0.50
0.10	Utilitarian	1.26	0.40	0.56	0.47	2.88	7.41	0.02
0.10	Rawlsian	1.26	0.40	0.56	0.47	2.85	7.35	-0.01

Table 5: Robustness to variation in disruption probability for one representative trained seed (seed = 1), evaluated without retraining over 50 episodes. We vary the disruption probability $p_{\text{break}} \in \{0.02, 0.05, 0.10\}$ while holding all other environment parameters fixed at their training values. Across all settings, utilitarian and Rawlsian policies continue to outperform the profit-maximizing baseline on well-being and operational delay metrics, whereas the profit-maximizing policy continues to achieve the highest profit.

Repair duration	Policy	$S_C + S_D$	$WB_{\min}$	WB_{GM}	ΔS	Detour _C	Wait _D	Profit
1-2	Profit-max	1.16	0.32	0.49	0.53	3.63	9.44	0.50
1-2	Utilitarian	1.28	0.41	0.57	0.46	2.67	7.07	0.00
1-2	Rawlsian	1.28	0.41	0.58	0.45	2.71	6.92	-0.01
2-4	Profit-max	1.17	0.32	0.49	0.53	3.46	9.55	0.50
2-4	Utilitarian	1.28	0.41	0.57	0.46	2.67	7.16	0.01
2-4	Rawlsian	1.28	0.41	0.58	0.46	2.58	7.05	-0.02
4-6	Profit-max	1.16	0.32	0.48	0.52	3.71	9.44	0.50
4-6	Utilitarian	1.27	0.40	0.57	0.47	2.63	7.40	0.03
4-6	Rawlsian	1.28	0.41	0.57	0.47	2.59	7.26	-0.02

Table 6: Robustness to variation in repair duration for one representative trained seed (seed = 1), evaluated without retraining over 50 episodes. We vary the repair duration over 1 to 2, 2 to 4, 4 to 6 while holding all other environment parameters fixed at their training values. The qualitative ordering of policies remains unchanged across all tested settings.

icy, enabling a controlled comparison of reward specifications reflecting different ethical theories. However, real sociotechnical systems often comprise agents guided by distinct objectives operating within the same environment. For example, different ride-sharing organizations may configure the subsets of autonomous vehicles they control according to different priorities.

To study how population composition influences outcomes, we evaluate *heterogeneous* populations composed of mixtures of agents trained under different reward specifications. We vary the fraction of agents following each policy type while holding fleet size constant, and analyze the resulting disruption-handling load and rider well-being. For these experiments, we use one representative trained policy (seed = 1) for each reward specification and evaluate each heterogeneous configuration over 50 episodes.

Figure 7 shows the fraction of disruption events handled by Rawlsian agents as a function of their population fraction in mixed populations with profit-maximizing agents. When the two policy types coexist, Rawlsian agents accept and resolve a disproportionately large share of disrupted requests relative to their representation in the fleet. This effect is strongest when Rawlsian agents constitute a small minority of the population, in which case their share of handled

disruptions exceeds their population share by approximately 60%. As the fraction of Rawlsian agents increases, this overload effect diminishes, approaching proportional load sharing under fully Rawlsian populations. We observe qualitatively similar trends in mixed populations of utilitarian and profit-maximizing agents, although the concentration of disruption-handling load is less pronounced than in mixed populations of Rawlsian and profit-maximizing agents.

Despite the asymmetric distribution of the load for handling disruptions, increasing the proportion of ethically guided agents improves rider outcomes. Figure 8 shows geometric mean well-being as a function of the Rawlsian agent fraction in mixed populations with profit-maximizing agents. Geometric mean well-being increases monotonically as the fraction of Rawlsian agents in the population grows, with gains observed even under partial Rawlsian adoption. A similar monotonic improvement holds in mixed populations of utilitarian and profit-maximizing agents as the fraction of utilitarian agents increases.

To further examine how ethical objectives shape *which* disruptions agents handle, we analyze heterogeneous populations composed of Rawlsian and utilitarian agents in equal proportion. Figure 9 reports the distribution of disruption handling across policies in a 50-50 Rawlsian-utilitarian

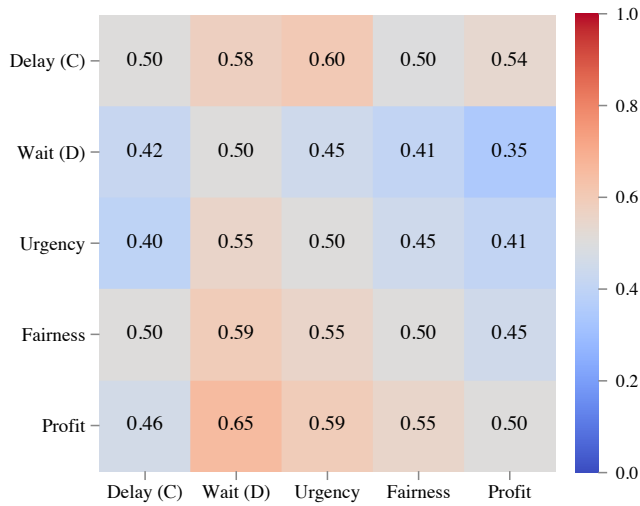


Figure 4: Criterion conflict matrix for the utilitarian policy. Entry (i, j) denotes the fraction of conflict cases in which row criterion i prevails over column criterion j (1.0 = 100% dominance). Delay is most influential in pairwise conflicts, whereas wait is outweighed by all other criteria.

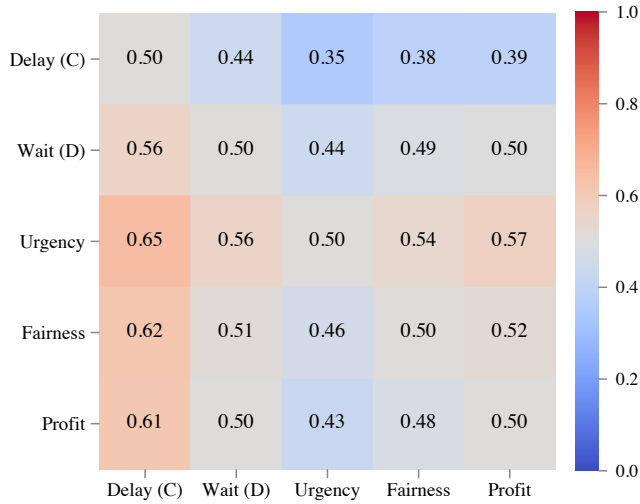


Figure 5: Criterion conflict matrix for the Rawlsian policy. Entry (i, j) denotes the fraction of conflict cases in which row criterion i prevails over column criterion j (1.0 = 100% dominance). Urgency and fairness prevail in most pairwise conflicts, whereas delay is outweighed by all other criteria.

population, conditioned on disruption characteristics. Although both policies contribute substantially, Rawlsian agents consistently handle a larger share of disruptions overall. This imbalance is most pronounced in ethically salient cases: high-burden disruptions, cases with large satisfaction gaps, and urgent requests.

These patterns indicate that Rawlsian agents preferentially accept disruptions involving greater burden, urgency, or rider disadvantage, whereas utilitarian agents more often handle lower burden cases. Figure 10 further groups

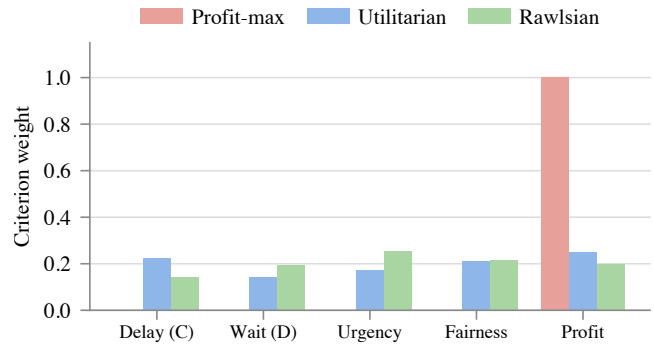


Figure 6: Learned criterion weights per policy.

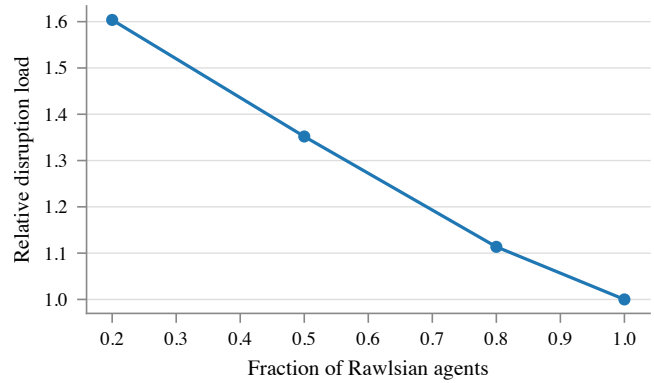


Figure 7: Relative disruption load of Rawlsian agents in mixed populations with profit-maximizing agents, measured as their share of disrupted requests handled divided by their population share. Values above 1 indicate that Rawlsian agents handle a disproportionately large share of disruptions relative to their representation in the population.

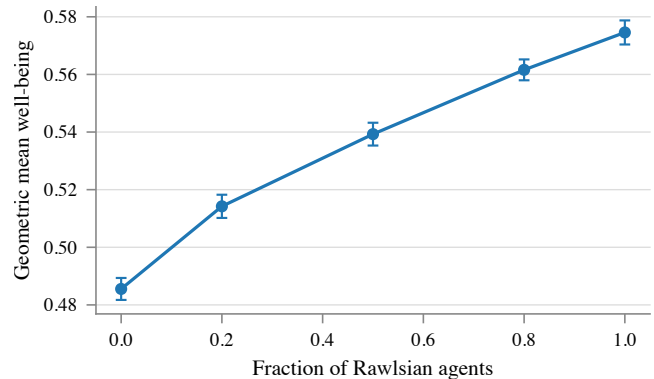


Figure 8: Geometric mean well-being as a function of the fraction of Rawlsian agents in mixed populations with profit-maximizing agents. Well-being increases monotonically as the proportion of Rawlsian agents grows, with gains observed even under partial ethical adoption.

disruptions by burden range, expressed as a fraction of the maximum observed burden. As burden increases, Rawlsian agents handle an increasing share of disruptions, account-

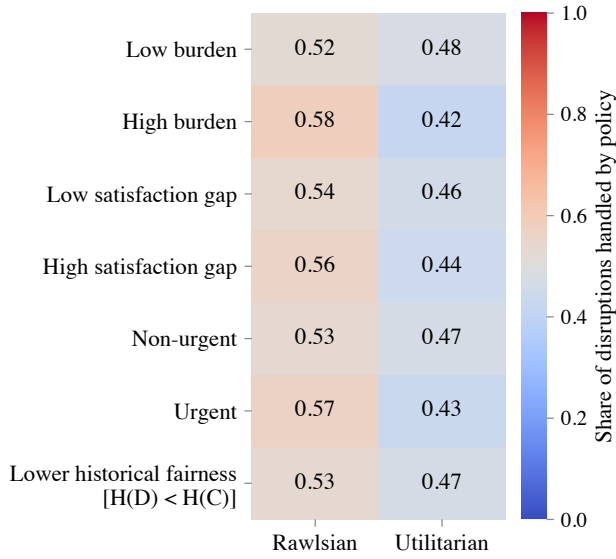


Figure 9: Distribution of disruptions handled by Rawlsian and utilitarian policies in a 50-50 heterogeneous population. Values indicate the fraction of disruptions assigned to each policy conditioned on disruption characteristics. Rawlsian agents consistently handle a larger share of the disruptions, particularly when there is urgency, a high burden, or a high satisfaction gap between the involved riders.

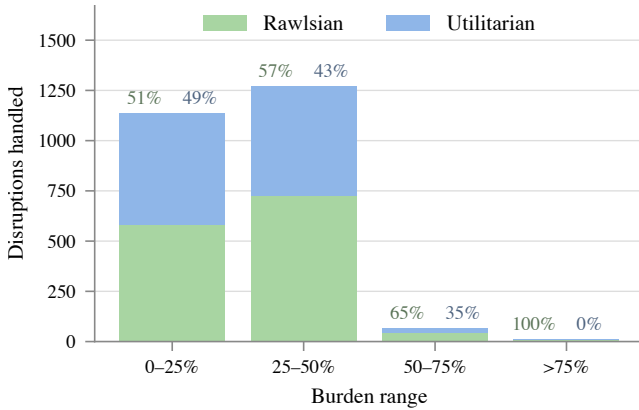


Figure 10: Disruptions handled in each burden range (defined as % of maximum observed burden) in a 50-50 Rawlsian-utilitarian population. As burden increases, Rawlsian agents handle an increasingly large share of disruptions, while utilitarian agents more frequently handle lower- and medium-burden cases. In the highest burden range, Rawlsian agents handle all disruptions.

ing for 100% of disruptions in the highest burden range. The utilitarian agents' share decreases steadily as burden increases. Taken together, Figures 9 and 10 suggest a structured division of labor in which Rawlsian agents shoulder the most ethically demanding disruptions, whereas utilitarian agents handle comparatively lighter cases.

Together, these results show that heterogeneous deployments can yield system-level improvements while concentrating disruption-handling load on ethically motivated agents, potentially causing a narrow subset of ethically-motivated agents to absorb a disproportionate share of disruption-handling costs in the absence of coordination or governance mechanisms.

6 Related Work

Prior work on ethical and normative reasoning in multiagent systems spans several complementary directions. Norm-aware architectures study how agents reason about stakeholder expectations, fairness, and values (Ajmeri et al. 2017, 2020; Bench-Capon and Modgil 2017; Serramia et al. 2023), while other work examines responsible norm deviation, resilience, and decentralized norm emergence (Singh and Singh 2023; Mahala et al. 2023; Vinitsky et al. 2023; Agrawal, Ajmeri, and Singh 2022). Together, these approaches provide sociotechnical and normative foundations for ethical decision making in autonomous systems.

A complementary line of work incorporates ethics into RL through reward shaping or policy-selection mechanisms. Existing approaches include ethics shaping based on similarity to human behavior (Wu and Lin 2018), multiobjective RL with ethical reward components (Rodriguez-Soto et al. 2022), Rawlsian reward formulations (Woodgate, Marshall, and Ajmeri 2025), moral-theory-derived rewards in social dilemmas (Tennant, Hailes, and Musolesi 2023), combinations of different moral theories to prevent extreme outcomes (Ecoffet and Lehman 2021), policy selection between task-optimal and constrained behavior for ethical alignment (Noothigattu et al. 2019), and normative supervision over learned policies (Neufeld et al. 2022). These works demonstrate multiple ways of embedding ethical considerations into learning-based agents.

Another related direction focuses on interpreting learned behavior. Existing approaches infer obligations from learned policies (Shea-Blymyer and Abbas 2022), analyze cost sensitivity in helping behavior (Mao et al. 2023), or develop interpretable representations of learned policies (Verma et al. 2018; Liu et al. 2018). Figure 11 summarizes the themes integrated in Diet and Table 7 situates Diet relative to representative approaches most closely related to our setting and objectives.

7 Discussion

This work suggests important directions for extending learning-based approaches to ethical decision making in sociotechnical systems. Although our framework operationalizes ethical considerations through system criteria motivated by stakeholder norms, we do not model norms explicitly. Future work may therefore investigate richer normative representations for guiding learning of ethical behavior, including norms that capture temporal structure, responsibility, and norm evolution. Extending the framework to hybrid or context-dependent ethical objectives may additionally enable the study of moral pluralism and ethical uncertainty in disrupted sociotechnical settings.

Work	Study Setting	Primary Contribution and Contrast with Diet
Woodgate, Marshall, and Ajmeri (2025)	Resource allocation	Shows that Rawlsian reward shaping improves minimum welfare and reduces inequality. $\Rightarrow$ Diet compares Rawlsian and utilitarian objectives under disruption and interprets the ethical priority structures learned by the resulting policies.
Rodriguez-Soto, Lopez-Sanchez, and Rodriguez-Aguilar (2021); Rodriguez-Soto et al. (2022)	Balancing task performance with ethical behavior	Demonstrates how moral requirements can be encoded as reward components in multiobjective RL to guarantee ethical agent behavior. $\Rightarrow$ Diet models disruption-response decision making where ethical criteria conflict, rather than treating ethics primarily as compliance with a fixed requirement.
Tennant, Hailes, and Musolesi (2023)	Social dilemma games	Shows that moral-theory-derived rewards shape cooperation and exploitation in social dilemmas. $\Rightarrow$ Diet shows that policies trained with different ethical objectives have distinct priorities over ethical system criteria in a disruption-driven STS.
Mao et al. (2023)	Helping scenarios	Introduces behavioral probes as a method to reveal whether agents continue helping as helping becomes costly. $\Rightarrow$ Diet introduces an analysis framework to extract implicit criterion priorities from learned policies through behavioral conflict analysis.
Ecoffet and Lehman (2021)	Abstract RL benchmarks	Shows that training under uncertainty across moral theories can reduce extreme behavior induced by commitment to a single theory. $\Rightarrow$ Diet fixes the ethical objective during training and instead analyzes how different objectives manifest in disruption-response behavior.
Noothigattu et al. (2019)	Sequential decision tasks	Balances task reward with ethical constraints through adaptive policy selection. $\Rightarrow$ Diet learns ethically grounded policies directly and interprets the tradeoffs implicit in their behavior.
Serramia et al. (2023)	Offline norm optimization	Shows how norm sets can be optimized for alignment with collective moral values. $\Rightarrow$ Diet treats values and norms as motivating criteria for the evaluation of learned disruption-response behavior.
Shea-Blymyer and Abbas (2022)	RL policy analysis	Shows that deontic obligations can be inferred from utility-maximizing learned policies. $\Rightarrow$ Diet infers relative priorities among conflicting criteria from learned policy behavior.
Diet	Multiagent transit disruption	Focuses on ethical decision making under disruption rather than routine operation, demonstrating that ethical reward formulations induce distinct disruption-response behaviors and learned criterion priorities, and that heterogeneous populations produce asymmetric disruption-handling load.

Table 7: Representative approaches to ethical decision making in autonomous systems. The table highlights the study setting and primary contribution of each line of work and its contrast with Diet.

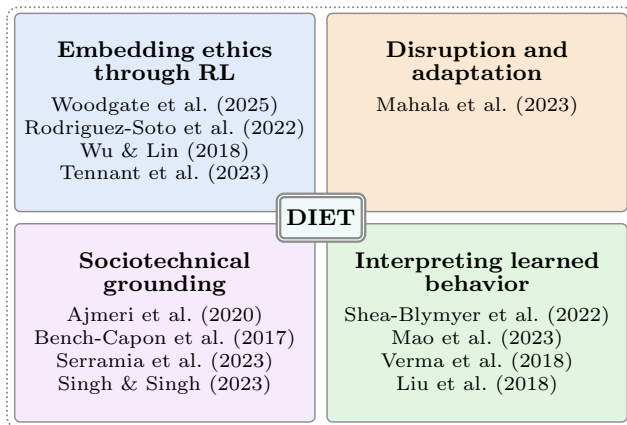


Figure 11: Key themes emphasized across related work and integrated in Diet.

Future work may also investigate more sophisticated methods for interpreting learned policies. For example, causal influence analysis or structured symbolic explanations could help reveal how specific observations, stakeholder characteristics, or environmental conditions shape the disruption-response decisions of trained models.

Our heterogeneous population results further raise questions about how disruption-handling responsibilities should be allocated across agents with different objectives. Future work may investigate coordination or governance mechanisms that prevent ethically motivated agents from disproportionately absorbing ethically demanding disruptions.

Finally, another important direction is to model richer forms of disruption and recovery. Real-world disruptions often evolve over time, involve uncertainty about future system conditions, and require sequences of interdependent decisions rather than isolated responses. Our formulation models disruptions as single-step decisions in a small simulated environment. Extending the framework to larger operational settings and multistep decision dynamics would enable analysis of cumulative action effects and long-term fairness.

Acknowledgments

We thank the anonymous reviewers for their thoughtful feedback and constructive suggestions, which helped improve this paper. Divya Sundaresan was supported in part by a Goodnight Doctoral Fellowship and an NC State College of Engineering Graduate Merit Award.

References

- Agrawal, R.; Ajmeri, N.; and Singh, M. P. 2022. Socially Intelligent Genetic Agents for the Emergence of Explicit Norms. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, 10–16. Vienna: IJCAI.
- Ajmeri, N.; Guo, H.; Murukannaiah, P. K.; and Singh, M. P. 2020. Elessar: Ethics in Norm-Aware Agents. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.
- Ajmeri, N.; Murukannaiah, P. K.; Guo, H.; and Singh, M. P. 2017. Arnor: Modeling Social Intelligence via Norms to Engineer Privacy-Aware Personal Agents. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 230–238.
- Andreoni, J. 1990. Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving. *The Economic Journal*, 100(401): 464–477.
- Bench-Capon, T.; and Modgil, S. 2017. Norms and value based reasoning: justifying compliance and violation. *Artificial Intelligence and Law*, 25: 29–64.
- Bradley, R. A.; and Terry, M. E. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3–4): 324–345.
- Briggs, G.; Williams, T.; Jackson, R. B.; and Scheutz, M. 2022. Why and how robots should say ‘no’. *International Journal of Social Robotics*, 14(2): 323–339.
- Ecoffet, A.; and Lehman, J. 2021. Reinforcement Learning Under Moral Uncertainty. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 1–18.
- Hobfoll, S. E. 1989. Conservation of Resources: A New Attempt at Conceptualizing Stress. *American Psychologist*, 44(3): 513–524.
- Kahneman, D.; and Tversky, A. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2): 263–291.
- Liu, G.; Schulte, O.; Zhu, W.; and Li, Q. 2018. Toward Interpretable Deep Reinforcement Learning with Linear Model U-Trees. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2018, Dublin, Ireland, September 10-14, 2018, Proceedings, Part II*, volume 11052 of *Lecture Notes in Computer Science*, 414–429. Springer.
- Mahala, G.; Kafalı, Ö.; Dam, H. K.; Ghose, A.; and Singh, M. P. 2023. A Normative Approach for Resilient Multiagent Systems. *Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS)*, 37(2): 46:1–46:40.
- Mao, Y.; Reinecke, M. G.; Kunesch, M.; Duéñez-Guzmán, E. A.; Comanescu, R.; Haas, J.; and Leibo, J. Z. 2023. Doing the right thing for the right reason: Evaluating artificial moral cognition by probing cost insensitivity. *CoRR*, abs/2305.18269.
- Mill, J. S. 1863. *Utilitarianism*. London: Longmans, Green and Company.
- Montes, N.; and Sierra, C. 2022. Synthesis and Properties of Optimally Value-Aligned Normative Systems. *Journal of Artificial Intelligence Research (JAIR)*, 74: 1739–1774.
- Neufeld, E. A.; Bartocci, E.; Ciabattoni, A.; and Governatori, G. 2022. Enforcing ethical goals over reinforcement-learning policies. *Ethics and Information Technology*, 24(4): 43.
- Noothigattu, R.; Bouneffouf, D.; Mattei, N.; Chandra, R.; Madan, P.; Varshney, K. R.; Campbell, M.; Singh, M.; and Rossi, F. 2019. Teaching AI Agents Ethical Values Using Reinforcement Learning and Policy Orchestration. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 6377–6381. ijcai.org.
- Rawls, J. 1999. *A Theory of Justice*. Cambridge, Massachusetts: Harvard University Press, 2nd edition. ISBN 9780674000780.
- Rodriguez-Soto, M.; Lopez-Sanchez, M.; and Rodriguez-Aguilar, J. A. 2021. Multi-Objective Reinforcement Learning for Designing Ethical Multi-Agent Environments. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*.
- Rodriguez-Soto, M.; Serramia, M.; Lopez-Sanchez, M.; and Rodriguez-Aguilar, J. A. 2022. Instilling Moral Value Alignment by Means of Multi-Objective Reinforcement Learning. *Ethics and Information Technology*.
- Serramia, M.; Rodriguez-Soto, M.; Lopez-Sanchez, M.; Rodriguez-Aguilar, J. A.; Bistaffa, F.; Boddington, P.; Wooldridge, M.; and Ansotegui, C. 2023. Encoding Ethics to Compute Value-Aligned Norms. *Minds and Machines*, 33: 421–449.
- Shea-Blymyer, C.; and Abbas, H. 2022. Generating Deontic Obligations From Utility-Maximizing Systems. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 653–663. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392471.
- Singh, A. M.; and Singh, M. P. 2023. Norm Deviation in Multiagent Systems: A Foundation for Responsible Autonomy. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*, 289–297. Macau: IJCAI.
- Singh, M. P. 2014. Norms as a Basis for Governing Sociotechnical Systems. *ACM Transactions on Intelligent Systems and Technology*, 5(1): 1–23.
- Small, K. A. 1982. The Scheduling of Consumer Activities: Work Trips. *The American Economic Review*, 72(3): 467–479.
- Sundaresan, D.; Watson, A.; Bardaka, E.; Lee, C. C.; Mayhorn, C. B.; and Singh, M. P. 2025. Prosociality in Microtransit. *Journal of Artificial Intelligence Research (JAIR)*, 82: 77–110.

- Tennant, E.; Hailes, S.; and Musolesi, M. 2023. Modeling Moral Choices in Social Dilemmas with Multi-Agent Reinforcement Learning. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, 317–325.
- Verma, A.; Murali, V.; Singh, R.; Kohli, P.; and Chaudhuri, S. 2018. Programmatically Interpretable Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, 5052–5061. PMLR.
- Vinitsky, E.; Köster, R.; Agapiou, J. P.; Duéñez-Guzmán, E. A.; Vezhnevets, A. S.; and Leibo, J. Z. 2023. A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. *Collective Intelligence*, 2(2).
- Wolpert, D. H.; and Tumer, K. 2001. Optimal payoff functions for members of collectives. *Advances in Complex Systems*, 4(2-3): 265–279.
- Woodgate, J.; Marshall, P.; and Ajmeri, N. 2025. Operationalising Rawlsian Ethics for Fairness in Norm Learning Agents. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI)*.
- Wu, Y.; and Lin, S. 2018. A Low-Cost Ethics Shaping Approach for Designing Reinforcement Learning Agents. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 1687–1694. AAAI Press.