Filling Causal Responsibility Gaps in Spatial Interactions using Feasible Action-Space Reduction by Groups - (Full)

Ashwin George*[1,2][0009−0007−7655−0737], Vassil Guenov*[1][0009−0001−8287−9199], Arkady Zgonnikov[1,2][0000−0002−6593−6948], David A. Abbink[1,2][0000−0001−7778−0090], and Luciano Cavalcante Siebert[1,2][0000−0002−7531−3154]

[1] Delft University of Technology, The Netherlands
[2] Centre for Meaningful Human Control A.George@tudelft.nl

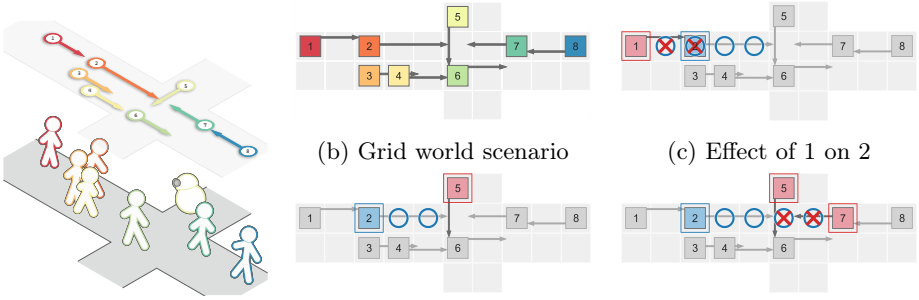
Abstract. Heralding the advent of autonomous vehicles and mobile robots that interact with humans, responsibility in spatial interaction is burgeoning as a research topic. Even though metrics of responsibility tailored to spatial interactions have been proposed, they are mostly focused on the responsibility of individual agents. Metrics of causal responsibility focusing on individuals fail in cases of causal overdeterminism — when many actors simultaneously cause an outcome. To fill the gaps in causal responsibility left by individual-focused metrics, we formulate a metric for the causal responsibility of groups. To identify assertive agents that are causally responsible for the trajectory of an affected agent, we further formalise the types of assertive influences and propose a tiering algorithm for systematically identifying assertive agents. Finally, we use scenario-based simulations to illustrate the benefits of considering groups and how the emergence of group effects vary with interaction dynamics and the proximity of agents.

Keywords: Responsibility · Causal Responsibility · Multi-Agent Systems · Spatial Interactions · Ethics of AI · Emergence

1 Introduction

Navigation of mobile robots and self-driving cars among humans, especially in multi-agent settings, faces many challenges owing to the complexity of multi-agent interactions [31,40]. In these safety-critical interactions, to make agents act responsibly and to attribute responsibility in case of accidents, we need models of responsibility [35,6]. Research related to spatial interactions has focused on 'responsible' or 'responsibility-aware' navigation based on how agents yield to other agents [36,8,15,42,43]. While these methods are not based on the formal definition of "responsibility", a more formal model of causal responsibility in

* These authors contributed equally to this work.



(a) Robot crossing (b) Grid world scenario (c) Effect of 1 on 2

(d) Effect of 5 on 2 (e) Effect of {5,7} on 2

Fig. 1: **Feasible action space reduction [17]:** For the robot crossing (a) represented in the grid world (b), the feasible action space reduction (FeAR) imposed by actors on affected agents are computed based on the feasible actions of the affected when actors follow their *Move de Riguer* (MdR) (represented by ○) and how many of these are rendered infeasible by the actual actions of actors (represented by ✖). For affected agent 2, (c) shows how agent 1 reduces the feasible action space by two, (d) shows how 5 on its own has no influence, and (e) shows how the group $\{5, 7\}$ reduces the feasible action space by two.

spatial interactions was proposed based on how individual agents restrict the feasible action space of other agents [17,18]. However, all of these approaches only focus on the responsibility of individual agents.

Formal models of group responsibility have dealt with the distributing the responsibility of collective outcomes to individual agents in more abstract scenarios with fewer actions, which do not scale to complex spatial interaction [4,1,47,49,30]. Understanding group responsibility is important because the actions of individuals might have superadditive influence on the outcome when acting together [13], which complicates the tracing of responsibility to individuals. This is even more problematic in cases of causal overdetermination, where many agents simultaneously cause an outcome and no individual agent can be deemed causally responsible in isolation [4].

Consider the scenario in Fig. 1 of a robot crossing some pedestrians, which was analysed with the Feasible Action-Space Reduction (FeAR) metric [17]. According to this metric, agents that restrict the feasible action space of another agent are causally responsible for the trajectory of the latter. If we represent this interaction using a grid world (as in [17]), where the actions of agents correspond to their speed (Fig. 1b), we can see that the action of pedestrian 1 is restricting the feasible actions of agent 2, compared to the case when agent 1 would have stayed Fig. 1c. Even though this insight helps us, individual FeAR fails in some cases. For example, if you look at how agent 2 is affected by 5 and 7. Both agents 5 and 7 simultaneously restrict the same actions, and even if one of them were to stay, the feasible action space of agent 2 would not change (Fig. 1d). But when we consider them as a group, we can see how they are collectively reducing the feasible action space of agent 2 (Fig. 1e). Since individual FeAR fails

in cases of causal overdetermination, relying on individual FeAR to ascribe responsibility can lead to responsibility gaps [13]. Responsibility gaps occur when a group of agents have collective responsibility, but no individual agent can be held responsible [13].

To prevent such responsibility gaps, we reformulate the Feasible Action-Space Reduction (FeAR) metric for quantifying the causal responsibility of groups on the trajectory of an affected agent (Section 2.3). For quantifying the contributions of individuals to group outcomes (as in [47,48]), we formally categorise different types of assertive influences (Section 2.4) and use these categorisations to rank the assertive influence of different agents into tiers (Section 2.5). These tiers were further used to explore the emergence of group effects in different scenarios (Section 3.2).

The main contributions of this paper are: (a) a formulation of the feasible action-space reduction (FeAR) metric to quantify the causal responsibility of groups on the trajectory of other agents in spatial interactions, (b) a formal categorisation of the types of assertive influences, (c) a tiering algorithm for ranking the assertive influence of agents on an affected agent, and (d) scenario-based simulations showing how the emergence of group effects are dependent on the dynamics of the interaction and proximity of the affected agent to other agents. Rest of the work is organised as definitions (Section 2), case studies (Section 3), results (Section 4), discussion (Section 5), and conclusion (Section 6).

2 Definitions

After some preliminary notations (Section 2.1) and the definition of the Feasible Action-Space Reduction (FeAR) metric for individual actors [17] (Section 2.2), we present our formulation of FeAR for groups (Section 2.3). Then, we formally categorise the types of assertive influence (Section 2.4) and propose a tiering algorithm for ranking the assertive influences of agents (Section 2.5).

2.1 Preliminaries

As proposed by [17], we model spatial interactions among a set of k agents $\mathcal{K}$ using a grid world where s represents the state of the grid world encompassing information about spatial constraints and the locations of agents. Each agent $i \in \mathcal{K}$, chooses an action a_i which corresponds to the speed with which they move for the time window in consideration. Each agent i has an action space of 17 actions, $\mathcal{A}_i = \{$S0, U1, U2, U3, U4, D1, D2, D3, D4, L1, L2, L3, L4, R1, R2, R3, R4$\}$ — where the first character indicates the direction (Stay, Up, Down, Left or Right) and the second character indicates the speed in steps moved per time window. $A = (a_i)_{i \in \mathcal{K}}$ represents the joint action of all the agents.

We compare the actions of agents a_i against their *Move de Rigueur*(MdR) μ_i which represents expectations of how agents would act in a given state s. Ideally, if all agents follow the joint MdR, $\mu = (\mu_i)_{i \in \mathcal{K}}$, there would not be any crashes. In this paper, we consider that the MdR is staying (S0) for all the

agents in all scenarios. The intervention of replacing the actual action with MdR is represented as follows:

$$[A_G \leftarrow \mu_G] = (a'_i)_{i\in\mathcal{K}} \quad \text{where, } a'_i \quad = \begin{cases} \mu_i \in \mu, & \text{if } i \in G \\ a_i \in A, & \text{if } i \notin G \end{cases} \tag{1}$$

and, $[A_i \leftarrow \mu_i] = \left[A_{\{i\}} \leftarrow \mu_{\{i\}}\right]$ represents the intervention of replacing the action of a single agent i with its MdR. Finally, $n_j\big(s, A\big)$ gives the number of feasible moves available to agent j for a given state s and joint action A.

2.2 Individual FeAR

Based on the idea that agents that restrict the feasible action space of other agents are causally responsible for the trajectory of the affected agent, the Feasible Action-Space Reduction (FeAR) metric was defined for individual actors as follows:

Definition 1 (FeAR [17]).
The Feasible Action-Space Reduction (FeAR) *imposed by actor i on affected agent j is defined as:*

$$\mathrm{FeAR}_{i,j}(s, A) = \begin{cases} Z\left(\frac{n_j\left(s,[A_i\leftarrow\mu_i]\right) - n_j\left(s,A\right)}{n_j\left(s,[A_i\leftarrow\mu_i]\right)+\epsilon}\right), & \text{if } i \neq j, \\ Z\left(\frac{n_i\left(s,A\right)}{n_i\left(s,[A_{\neg i}\leftarrow\mu_{\neg i}]\right)+\epsilon}\right), & \text{if } i = j, \end{cases} \tag{2}$$

where $\neg i = \mathcal{K} \setminus \{i\}$, $Z(x) = \min\left(-1, \max(x, 1)\right)$, and $0 < \epsilon \ll 1$.

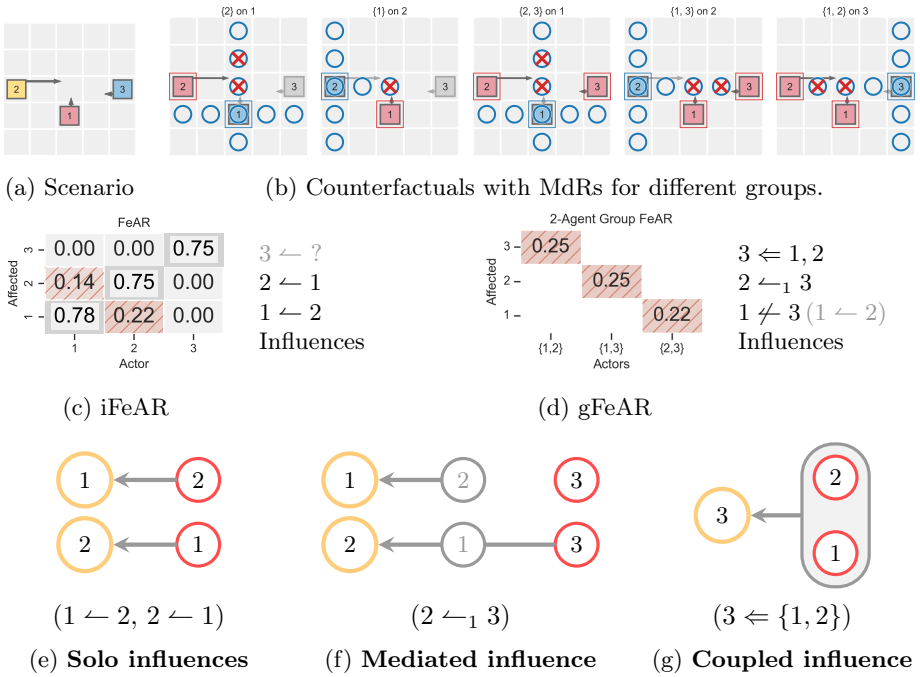
The function $Z(x)$ clips the values of FeAR to $[-1, 1]$ to aid interpretability of FeAR values and ϵ in the denominator ensures that FeAR is defined when $n_j\big(s, [A_i \leftarrow \mu_i]\big) = 0$. Positive values of $\mathrm{FeAR}_{i,j}$ indicate that actor i is being assertive towards the affected agent j and is hence causally responsible for j's trajectory. For example, in the scenario shown in Fig. 2, agent 1 decreases the feasible action space of agent 2 by one (Fig. 2b) and hence $\mathrm{FeAR}_{1,2} > 0$ (Fig. 2c).

2.3 Group FeAR

To better capture causal responsibility in cases of causal overdeterminism, we propose the FeAR metric for groups (gFeAR) as follows:

Definition 2 (Group FeAR). *The Feasible Action-Space Reduction* (FeAR) *imposed by a non-empty group $G \subseteq \neg j$ on an affected agent j , is defined as:*

$$\mathrm{FeAR}_{G,j}(s, A) = \frac{n_j\big(s, [A_G \leftarrow \mu_G]\big) - n_j\big(s, A\big)}{n_j\big(s, [A_G \leftarrow \mu_G]\big) + \epsilon}. \tag{3}$$



(a) Scenario (b) Counterfactuals with MdRs for different groups.

(c) iFeAR (d) gFeAR

(e) **Solo influences** (f) **Mediated influence** (g) **Coupled influence**

Fig. 2: **Types of assertive influence based on group FeAR:** For the illustrative scenario (a) where three agents are moving towards each other, we show how counterfactuals based on the MdR of actors (b) are used to compute iFeAR values (c) for individual actors and gFeAR values for group actors (d). While iFeAR can only identify *solo* influences ($1 \leftharpoonup 2$ and $2 \leftharpoonup 1$) (e), analysing gFeAR can reveal additional *mediated* ($2 \leftharpoonup_1 3$) (f) and *coupled* ($3 \Leftarrow \{1, 2\}$) (g) influences. Even though $\text{FeAR}_{3,3} < 1$ reflects the reduction in feasible action space of 3, iFeAR cannot identify the assertive actors; which are revealed by gFeAR ($3 \Leftarrow \{1, 2\}$). Also note that agent 3 has no influence on agent 1 as $\text{FeAR}_{2,1} = \text{FeAR}_{\{2,3\},1}$.

Looking at Fig. 2 again, the group $\{1, 3\}$ acting together reduces the feasible action space of 2 by 2 Fig. 2b. This is captured by the definition of FeAR for groups and $\text{FeAR}_{\{1,3\},2} > \text{FeAR}_{1,2} > 0$ shows that the group of actors $\{1, 3\}$ has more influence on agent 2 than just the individual actor 1. Furthermore, when considering the effect on agent 3, $\text{FeAR}_{1,3} = \text{FeAR}_{2,3} = 0$, fails to capture any influence of other agents. But $\text{FeAR}_{\{1,2\},3} > 0$ shows that collectively $\{1, 2\}$ is behaving assertively towards agent 3 and is hence collectively causally responsible for the trajectory of agent 3. Thus, gFeAR can quantify casual responsibility in cases of causal overdetermination.

As seen above, agents can have different influences on an affected agent when considered to be acting individually or as part of a group. The following section categorises these types of influences.

2.4 Types of assertive influence

Based on whether agents have assertive influence on their own or as part of groups, we have identified 3 fundamental types of assertive influence: *solo influence*, *mediated influence* and *coupled influence*, and a fourth derived type of influence *mediated coupled influence*. These four types of influence defined below span the whole spectrum of assertive influences and helps us compare the assertiveness of different agents.

Definition 3 (Solo influence $j \leftharpoonup i$).
Agent i has solo influence on agent j if $\mathrm{FeAR}_{i,j} > 0$.

Definition 4 (Mediated influence $j \leftharpoonup_G i$). *Agent i has a mediated influence on agent j if* $\mathrm{FeAR}_{i,j} = 0$, *and* $\exists G \subset \neg j \setminus \{i\}$, $G \neq \emptyset$ *such that* $\mathrm{FeAR}_{G\cup\{i\},j} > \mathrm{FeAR}_{G,j} > 0$.

Definition 5 (Coupled influence $j \Leftarrow G$).
All agents in group $G \subset \neg j \setminus \{i\}$, $G \neq \emptyset$ have coupled influence on agent j if $\mathrm{FeAR}_{i,j} = 0 \quad \forall i \in G$ *and* $\mathrm{FeAR}_{G,j} > 0$.

There can also be cases where a group of agents have coupled influence that is mediated by another group of agents. Therefore a more general definition of mediated coupled influence is as follows:

Definition 6 (Mediated coupled influence $j \Leftarrow_{G'} G$).
Group G has a coupled influence on agent j which is mediated by another group G' if $\mathrm{FeAR}_{i,j} = \mathrm{FeAR}_{G',j} \quad \forall i \in G$ *and* $\mathrm{FeAR}_{G'\cup G,j} > \mathrm{FeAR}_{G',j}$

In the example shown in Fig. 2, agents 1 and 2 have solo influences on each other ($1 \leftharpoonup 2$, $2 \leftharpoonup 1$), agent 3 has an influence on agent 2 which is mediated by agent 1 ($2 \leftharpoonup_1 3$), and agents 1 and 2 have a coupled influence on agent 3 ($3 \Leftarrow \{1, 2\}$). Figs. 2e to 2g show how these influences are pictorially represented.

Fig. 3 shows a more complicated scenario with more intricate influences. For example, the group $\{5, 7\}$ has a mediated coupled influence on agent 1 $1 \Leftarrow_{\{2,3,4,6\}} \{5, 7\}$ (Fig. 3d). For systematically unravelling these intricate dependencies of influence by identifying minimal groups that are causally responsible, we propose a tiering algorithm for ranking the assertive influences on each affected agent.

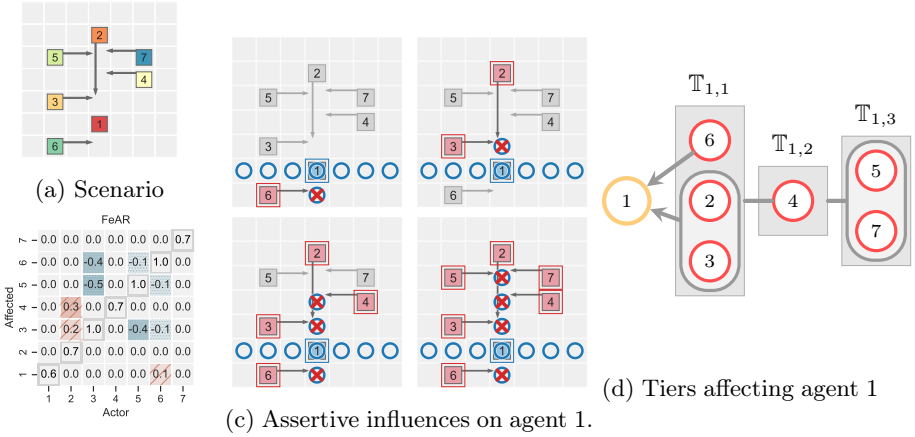
2.5 Ranking agents based on influence

In cases of agents having mediated influence, their influence is conditional to the actions of the mediating agent. Thus, based on the intuition that the assertive influence of agents with mediated influence should be ranked lower than the assertive influence of those agents that mediate these mediated influences, we propose a tiering algorithm for ranking the influence of different agents into tiers (Algorithm 1) [3].

[3] Detailed implementation of the algorithm can be found at github.com/DAI-Lab-HERALD/FeAR.

Algorithm 1 Tiering the assertive influence of agents

$\phi_j \leftarrow \{i : \text{FeAR}_{i,j} < 0\}$ — Find courteous agents ϕ_j.
$\kappa_j \leftarrow \neg j \setminus \phi_j$ — Remove ϕ_j from candidate agents κ_j.
$\mathcal{R}_0 \leftarrow \emptyset$ — $\mathcal{R}_n$ — Assertive agents until $\mathbb{T}_{j,n}$.
for $n = 1, 2, \cdots$: — For each tier,
  $\mathbb{T}_{j,n} \leftarrow \emptyset$
  for $k = 1, \ldots, |\kappa_j|$: — starting with k=1,
    for all $G \subseteq \kappa_j, |G| = k$: — loop through groups with k agents,
      if $\text{FeAR}_{\mathcal{R}_{n-1} \cup G, j} > \text{FeAR}_{\mathcal{R}_{n-1}, j}$: — identify assertive groups,
        $\mathbb{T}_{j,n} \leftarrow \mathbb{T}_{j,n} \cup \{G\}$, — add them to tiers, and
        $\kappa_j \leftarrow \kappa_j \setminus G$ — remove them from candidate agents.
  if $\mathbb{T}_{j,n} = \emptyset$: **break**
  $\mathcal{R}_n \leftarrow \mathcal{R}_{n-1} \cup \bigcup_{G \in \mathbb{T}_{j,n}} G$



(a) Scenario

(b) iFeAR

(c) Assertive influences on agent 1.

(d) Tiers affecting agent 1

Fig. 3: **Ranking the assertiveness of agents into tiers $\mathbb{T}_{j,n}$:** In this illustrative scenario (a), when considering agent 1 as the affected, iFeAR only show agent 6 as being the assertive (b). However, counterfactuals with groups (c) reveal more assertive influences on agent 1 which are systematically ranked into tiers $\mathbb{T}_{1,n}$ (d).

To represent how higher tiers have more influence on the affected agent j, we use the $\succ$ operator : $\mathbb{T}_{j,n} \succ \mathbb{T}_{j,n+1}$.

Consider how different actors influence agent 1 for the example in Fig. 3a. Fig. 3c shows the assertive influence of groups of actors $\{6\}$, $\{2,3\}$, $\{2,3,4,6\}$ and $\{2,3,4,5,6,7\}$. Based on these assertive influences, the algorithm systematically identifies the solo influence of agent 6 ($1 \leftharpoonup 6$), the coupled influence of group $\{2,3\}$ ($1 \Leftarrow \{2,3\}$), the mediated influence of agent 4 ($1 \leftharpoonup_{\{2,3,6\}} 4$) and the mediated coupled influence of group $\{5,7\}$ ($1 \Leftarrow_{\{2,3,4,6\}} \{5,7\}$), and sorts these assertive influences into three tiers $6 \sim 2 \sim 3 \succ 4 \succ 5 \sim 7$ as shown in Fig. 3d. Thus, the tiers provide richer information about assertive influences than the assertive influences found from the positive values of individual FeAR (Fig. 3b).

3 Scenario-based simulations

We defined metrics to quantify how group FeAR can uncover more information about an interaction than individual FeAR (Section 3.1). Using these metrics, we explored group effects in simulations of different scenarios (Section 3.2).

3.1 Metrics

Since the goal of group FeAR and the tiering algorithm is to indentify assertive agents and to rank their assertiveness, we consider two metrics for comparing FeAR and group FeAR, 1) based on the number of assertive agents and 2) based on the alignment of rankings of assertiveness using Kendall's τ. To better understand the relationship of these metrics to the proximity of agents, we plot these against the median Manhattan distance to the affected agent.

Difference in the number of assertive agents: For an affected agent j, the number of assertive agents identified with individual FeAR is the number of actors i with $\text{FeAR}_{i,j} > 0$, and the number of assertive agents identified group FeAR is the number of actors in tiers $\mathbb{T}_{j,n}$. Since, the number of assertive agents vary with scenarios, we use the difference in the number of assertive agents identified using group FeAR and individual FeAR to provide a consistent metric across scenarios:

$$\Delta_j^{\text{Assertive}} = n_{j,\text{gFeAR}}^{\text{Assertive}} - n_{j,\text{iFeAR}}^{\text{Assertive}} = |\cup_n \mathbb{T}_{j,n}| - |\{i : \text{FeAR}_{i,j} > 0\}| \,. \quad (4)$$

Kendall's tau for rankings: For each affected agent, we can rank the assertiveness of other agents in three ways: 1) *iFeAR* ranks: by sorting the positive values of individual FeAR, 2) *gFeAR-Tier* ranks: based on the tiers generated by the tiering algorithm based on group FeAR, and 3) *gFeAR-Shapley* ranks: based on ranking positive Shapely values [44] generated from all the group FeAR values for that affected agent.

We compare *iFeAR* and *gFeAR-Tier* ranks against baseline of *gFeAR-Shapley* derived from Shapely values which are the state of the art when computing individual contributions to groups [11,44,48]. Shapley values are calculated based on the marginal changes in group FeAR value when an actor is added to a group of actors [44]. The main difference between *gFeAR-Shapley* ranks and *gFeAR-Tier* ranks is that shapely values consider all possible ways for assembling groups, while tiers are constructed in a more systematic manner starting from solo and coupled influences and then moving onto mediated influences.

For each affected agent in each case, we compare two rankings of assertiveness using Kendall's tau [28] which returns $\tau = +1$ if all pairs of actors have the same relative ranks in the both rankings. When creating the arrays of ranks for comparisons, non-assertive agents are given a rank of $k + 1$ where k is the total number of agents in that scenario.

The rankings of assertive influences, as in the case of the number of assertive agents, are dependent on the scenarios. To better understand the relationship between specific scenarios and these metrics we plot them against a metric for the proximity of agents.

Median Manhattan Distance: All the scenarios for the randomised simulations (Section 3.2) had the same number of agents, and we used the median Manhattan distance between agents to indicate the proximity of agents in each case. Cases where agents are closer together would have lower median Manhattan distances. By plotting the differences in number of assertive agents identified and Kendall's τ comparing ranks for different median Manhattan distances, we can see how group effects are related to proximity of agents.

3.2 Scenarios

We start by analysing a particular scenario (S1) in detail to illustrate how the metrics capture group effects. Later, we use these metrics to explore group effects in pseudo-randomised simulations of three different scenarios (S2, S3, S4).
S1:Robot crossing pedestrians: For the detailed scenario, we consider the description in the introduction where a robot (agent 5) is crossing the path of some pedestrians Fig. 1 and crashing into pedestrians 2, 3, 4 and 7.
S2-4:Randomised simulations: We consider three scenarios *S2:Aggressive*, *S3:Directed* and *S4:Random* for the pseudo-randomised simulations (see Fig. 5a). For *Aggressive*, agents spawn in random locations and the policy makes agents take random actions to aggressively cross the intersection. For *Directed*, agents spawn in fixed locations to the left of the intersection and then the policy makes them randomly choose actions to gently cross to the right and slow down after crossing the intersection. For *Random*, agents spawn in the same location as *Directed*, but then take completely random actions in all directions. 50 simulations of each scenario were run with 5 iterations per simulation. Thus, in total there were 250 separate cases for each scenario with eight agents.

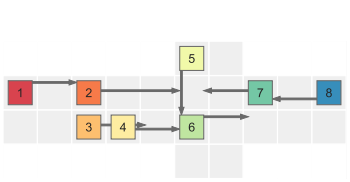
4 Results

Groups effects in four scenarios were analysed with respect to the difference in the number of assertive agents identified and Kendall's τ for comparing rankings of assertiveness.

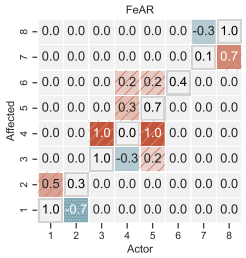
4.1 Number of assertive agents from individual and group FeAR

The tiers for scenario S1 shown in Table 1 show how the robot (agent 5) is affected by the solo influence of pedestrian 4 ($5 \leftharpoonup 4$) and coupled influence of pedestrians 2 and 7 ($5 \Leftarrow \{2,7\}$), and how it affects all the pedestrians either through solo ($3 \leftharpoonup 5$, $4 \leftharpoonup 5$, $6 \leftharpoonup 5$) or coupled influence ($1 \Leftarrow \{5,7\}$, $2 \Leftarrow \{5,7\}$, $7 \Leftarrow \{2,5\}$, $8 \Leftarrow \{2,5\}$). Besides these influences involving agent 5, the algorithm also identifies mediated influences ($4 \leftharpoonup_{\mathbb{T}_{4,1}} 2$, $7 \leftharpoonup_{\mathbb{T}_{7,1}} 6$) and mediated coupled influence ($6 \Leftarrow_{\mathbb{T}_{6,1}} \{2,7\}$) which are mediated by tiers $\mathbb{T}_{j,1}$ of affected agents j.

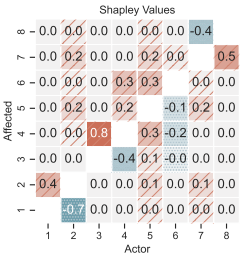
Thus, by identifying coupled and mediated influences, group FeAR identifies more agents which are being assertive towards an affected agent, than



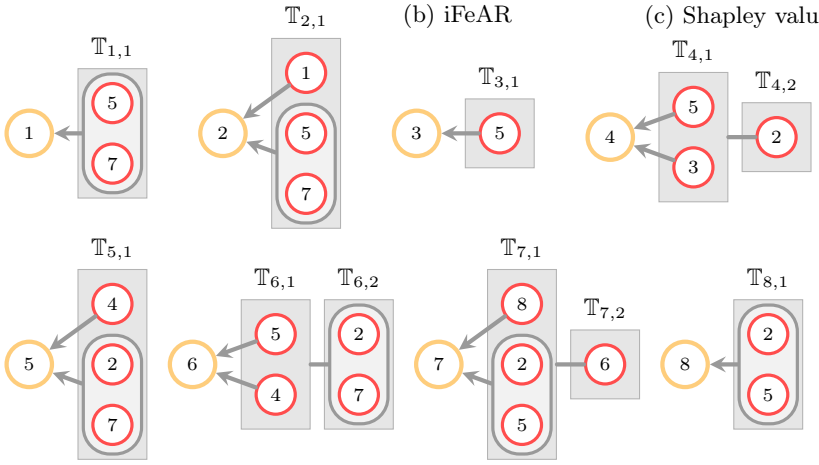
(a) Robot (5) crossing pedestrians (1, 2, 3, 4, 6, 7 and 8)



(b) iFeAR



(c) Shapley values



(d) Tiers affecting different agents from group FeAR.

Fig. 4: **S1: Uncovering group effects with group FeAR:** For the robot crossing scenario from Fig. 1, represented in the grid world as in (a), compared to just using iFeAR (b), we are able to uncover more assertive influences using gFeAR, either through the Shapley values (c) or tiers (d). For example, while iFeAR only shows the assertive influence of the robot 5 on pedestrians 3, 4 and 6, both Shapley values and tiers show that 5 is assertive towards all the pedestrians.

would be possible just by using individual FeAR (as shown by the values of $\Delta_j^{\text{Assertive}} = n_{j,\text{gFeAR}}^{\text{Assertive}} - n_{j,\text{iFeAR}}^{\text{Assertive}}$ in Table 1). Since iFeAR can only identify solo influences, it can never identify more assertive agents than the tiers from gFeAR; so $n_{j,\text{gFeAR}}^{\text{Assertive}} \geq n_{j,\text{iFeAR}}^{\text{Assertive}}$ and $\Delta_j^{\text{Assertive}} \geq 0$.

In S1, the value of $\Delta_j^{\text{Assertive}}$ has a maximum of three for affected agent 7. For agent 7, in addition to the solo influence of agent 8 ($8 \leftharpoonup 7$), the tiers reveal the coupled influence of agents 2 and 5 ($\{2,5\} \Leftarrow 7$), and the mediated influence of agent 6 ($6 \leftharpoonup_{\{2,5,8\}} 7$).

Furthermore, agents 6 and 7 have the maximum number of assertive agents ($n_{j,\text{gFeAR}}^{\text{Assertive}} = 4$) followed by agents 2, 4 and 5 ($n_{j,\text{gFeAR}}^{\text{Assertive}} = 3$). These agents are in the centre of the interaction and their proximity to other agents might be a contributing factor for the group effects on them.

To further explore the effect of proximity on group effects in the randomised simulations, we plot $\Delta_j^{\text{Assertive}}$ versus median Manhattan distance of other agents

Table 1: **Ranks of assertive influence in S1:** Ranking of assertive influence calculated using FeAR, tiers from group FeAR and Shapley values of group FeAR. '■' show affected agents which are not considered in the rankings.

(a) Ranking's of assertiveness and $\Delta_j^{\text{Assertive}}$.

Affected: 1	$\Delta_j^{\text{Assertive}} = 2$
iFeAR ranks	[■ . . . . . .]
gFeAR (Tier) ranks	[■ . . . 1 . 1 .]
gFeAR (Shapley) ranks	[■ . . . 1 . 1 .]

Affected: 2	$\Delta_j^{\text{Assertive}} = 2$
iFeAR ranks	[1 ■ . . . . . .]
gFeAR (Tier) ranks	[1 ■ . . . 1 . 1 .]
gFeAR (Shapley) ranks	[1 ■ . . . 2 . 2 .]

Affected: 3	$\Delta_j^{\text{Assertive}} = 0$
iFeAR ranks	[. . ■ . 1 . . .]
gFeAR (Tier) ranks	[. . ■ . 1 . . .]
gFeAR (Shapley) ranks	[. . ■ . 1 . . .]

Affected: 4	$\Delta_j^{\text{Assertive}} = 1$
iFeAR ranks	[. . 1 ■ 1 . . .]
gFeAR (Tier) ranks	[. 3 1 ■ 1 . . .]
gFeAR (Shapley) ranks	[. 3 1 ■ 2 . . .]

Affected: 5	$\Delta_j^{\text{Assertive}} = 2$
iFeAR ranks	[. . . 1 ■ . . .]
gFeAR (Tier) ranks	[. 1 . 1 ■ . 1 .]
gFeAR (Shapley) ranks	[. 1 . 3 ■ . 1 .]

Affected: 6	$\Delta_j^{\text{Assertive}} = 2$
iFeAR ranks	[. . . 1 1 ■ . .]
gFeAR (Tier) ranks	[. 3 . 1 1 ■ 3 .]
gFeAR (Shapley) ranks	[. 3 . 1 1 ■ 3 .]

Affected: 7	$\Delta_j^{\text{Assertive}} = 3$
iFeAR ranks	[. . . . . ■ 1]
gFeAR (Tier) ranks	[. 1 . . 1 4 ■ 1]
gFeAR (Shapley) ranks	[. 2 . . 2 4 ■ 1]

Affected: 8	$\Delta_j^{\text{Assertive}} = 2$
iFeAR ranks	[. . . . . . . ■]
gFeAR (Tier) ranks	[. 1 . . 1 . . ■]
gFeAR (Shapley) ranks	[. 1 . . 1 . . ■]

(b) Kendall's τ comparing rankings. τ close to 1 indicate similar rankings.

Affected	1	2	3	4	5	6	7	8
τ(iFeAR, gFeAR-(Tier))	-	0.47	1.00	0.85	0.47	0.79	0.42	-
τ(iFeAR, gFeAR-(Shapley))	-	0.65	1.00	0.82	0.22	0.79	0.59	-
τ(gFeAR-(Tier), gFeAR-(Shapley))	1.00	0.93	1.00	0.97	0.93	1.00	0.94	1.00

to the affected agent (Fig. 5b). Low values of the median Manhattan distance would imply that other agents were more proximal to the affected agent. It should be noted that the median Manhattan distance to affected agent is a meaningful metric only because all the scenarios considered here have the same map and number of agents. As the median Manhattan distance increases beyond a threshold, the maximum $n_{j,\text{gFeAR}}^{\text{Assertive}}$ line show how the number of assertive agents start to drop. Counts of instances of $\Delta_j^{\text{Assertive}}$ for different scenarios also show mirror this trend of decreasing group effects as the affected agents gets farther from other agents. To better understand the trends in group effects we also plot the fraction of non-zero $\Delta_j^{\text{Assertive}}$ and mean $\Delta_j^{\text{Assertive}}$ for each scenario in Fig. 5b. Both of these show how group effects decrease as the proximity of the affected agent to others decrease.

Another key finding is how group effects vary across different simulation scenarios. The fraction of non-zero $\Delta_j^{\text{Assertive}}$ and mean $\Delta_j^{\text{Assertive}}$ for each mean Manhattan distance show how group effects are generally the largest in case

of the *Aggressive* scenario where all the agents are aggressively crossing the intersection. Also, for larger distances from the affected agent, group effects are still present for *Assertive*, while they die off for *Directed* and *Random*. Of the three scenarios, *Random* has the lowest values of $\Delta_j^{\text{Assertive}}$ while *Directed* has intermediate values. In summary, the values of $\Delta_j^{\text{Assertive}}$ show that group effects are strongest in *Assertive* and weakest in *Random*.

4.2 Comparing rankings of assertiveness

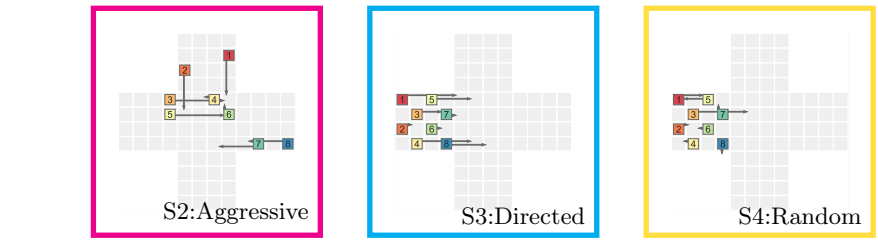
So far, we have explored the difference in the number of assertive agents identified using individual FeAR and group FeAR. Besides just identifying assertive agents, our algorithm ranks assertive influences into tiers. Here, we will compare these *gFeAR-Tier* ranks against *iFeAR* ranks generated from individual FeAR values and *gFeAR-Shapley* ranks generated from Shapley values based in group FeAR.

The rankings for the robot crossing scenario (CS1) are summarised in Table 1 in terms of ranks. So for affected agent 7, the *gFeAR-Tier* ranks of $(1, 1, 1, 4)$ for agents 2, 5, 8 and 6 represent the ranking $2 \sim 5 \sim 8 \succ 6$ and the *gFeAR-Shapley* ranks of $(1, 2, 2, 4)$ for agents 8, 2, 5 and 6, represent the ranking $8 \succ 2 \sim 5 \succ 6$.

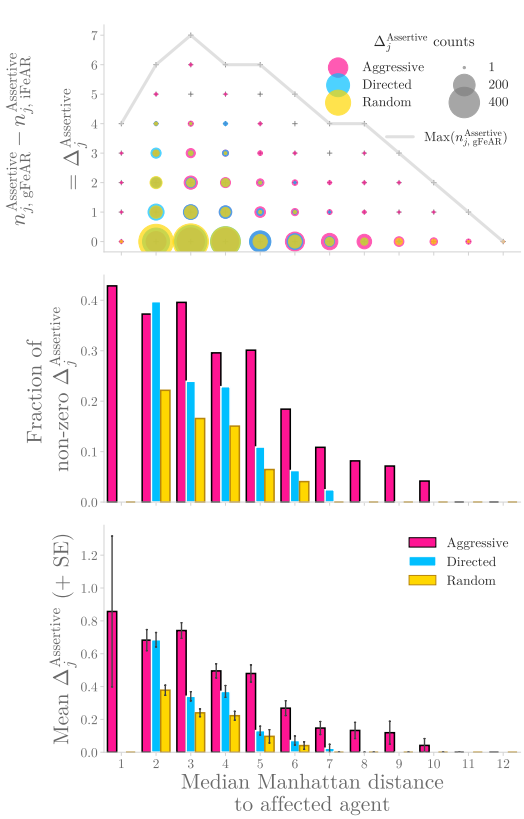
Apart from fewer assertive agents identified using iFeAR, most of the other rankings are in agreement with each other. One difference between *gFeAR-Tier* and *gFeAR-Shapley* ranks concerns the ranking of coupled influences. Coupled influences are consistently ranked along solo influences in *iFeAR* ranks, while the *gFeAR-Shapley* ranks for coupled influences is higher than the solo influence for affected agent 5, and lower than solo influences for agents 2 and 7. Another difference between *gFeAR-Shapley* and *gFeAR-Tier* ranks is in the ranks of solo influence — agents with solo influence always have rank 1 in *gFeAR-Tier*, whereas *gFeAR-Shapley* ranks for agents with solo influence can be different. For example, in the *gFeAR-Shapley* ranks for affected agent 4, the ranks of agents 3 and 5 are 1 and 2 respectively, whereas both of them have *gFeAR-Tier* of 1.

The difference among these rankings were quantified using Kendall's τ which had values in the range $[-1, 1]$ and complete agreement among rankings had $\tau = 1$. Based on the difference of $\tau(\textit{iFeAR, gFeAR-Tier})$ from 1, the difference between *iFeAR* and *gFeAR-Tier* ranks was largest for agent 7 and smallest for agent 3 ($7 > 2 \sim 5 > 6 > 4 > 3$). It must be noted that even though agents 1 and 8 experience group effects due to coupled influences, the lack of *iFeAR* makes computing $\tau(\textit{iFeAR, gFeAR-Tier})$ impossible. Even with this caveat, the values of $\tau(\textit{iFeAR, gFeAR-Tier})$ give a good indication for the overall presence of group effects in the scenario. Thus, the values of $\tau(\textit{iFeAR, gFeAR-Tier})$ show how group effects are stronger on central agents that are more proximal to other agents. Similarly, the values of $\tau(\textit{iFeAR, gFeAR-Shapley})$ which represent the difference between the *iFeAR* and *gFeAR-Shapley*, also highlight the prevalence of groups effects on central agents — with $1 - \tau$ values for different affected agents ranked as $5 > 7 > 2 > 6 > 4 > 3$.

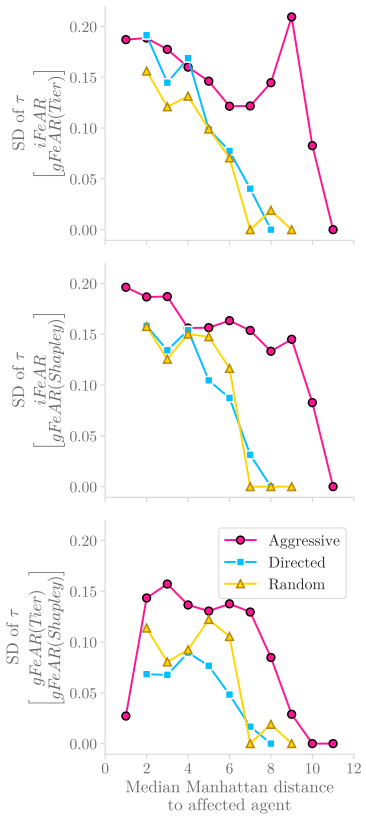
In addition to comparing iFeAR and gFeAR using $\tau(\textit{iFeAR, gFeAR-Tier})$, we can use $\tau(\textit{gFeAR-Tier, gFeAR-Shapley})$ to compare the two rankings generated from



(a) Scenarios



(b) Difference in the number of assertive agents identified using gFeAR and iFeAR $\Delta_j^{\text{Assertive}}$.



(c) Standard deviation of Kendall's τ for comparing ranks.

Fig. 5: **Emergence of group effects in randomised simulations:** Different simulation scenarios are shown in (a). Difference in the number of assertive agents $\Delta_j^{\text{Assertive}}$ identified using individual FeAR (iFeAR) and (tiers of) group FeAR (gFeAR) for different proximity to the affected agent are shown in (b). Rankings made using FeAR, tiers from gFeAR or Shapley values of gFeAR we compared using Kendall's τ. Variation of Kendall's τ with respect to the scenarios and median Manhattan distance between agents is shown in (c).

gFeAR. All the values of $\tau(_{gFeAR\text{-}Tier},\ _{gFeAR\text{-}Shapley}) < 0.9$ indicate strong agreement between *gFeAR-Tier* and *gFeAR-Shapley* ranks. Again the values of $1 - \tau$ for different affected agents ($2 \sim 5 > 7 > 4 > 1 \sim 3 \sim 6 \sim 8$) indicate greater differences in central agents.

Thus, since the group effects on an affected agent might be related to its centrality and proximity to other agents, quantified proximity to other agents using the median Manhattan distance and explored how taus were related to this distance in different scenarios. We are interested in the spread of τ values and hence plot the standard deviation (SD) of τ in different scenarios for different median Manhattan distance (Fig. 5c).

For both $\tau(_{iFeAR},\ _{gFeAR\text{-}Tier})$ and $\tau(_{iFeAR},\ _{gFeAR\text{-}Shapley})$ comparing iFeAR and gFeAR, we can see that the SD in τ drops as the median Manhattan distance to the affected agent increases. Furthermore, we can see that the SD in τ is mcuh greater for *Aggressive* than for *Directed* or *Random*.

Regarding the values of $\tau(_{gFeAR\text{-}Tier},\ _{gFeAR\text{-}Shapley})$, the spread of the SDs indicate differences between *gFeAR-Tier* and *gFeAR-Shapley*, but these differences are smaller than their difference with *iFeAR* ranks.

5 Discussion

5.1 Filling causal responsibility gaps with Group FeAR

One of the main challenges with AI agents and collective actions is the possibility of responsibility voids [32,5,39,21,46]. Responsibility voids occur when no one can be held responsible for an outcome that resulted from collective action [5,12]. When a group of agents can be held responsible, but no individual can be held responsible, this is called a responsibility gap [13]. The discourse on responsibility voids and responsibility gaps revolve around moral responsibility, which depends on conditions like intention, knowledge, and wrong-doing on top of causally contributing to the outcome. In this paper, we skip the all other conditions and focus on causal responsibility and thereby on causal responsibility gaps.

We define that causal responsibility gaps occur when a group of agents have collective causal responsibility, but none of the individuals on their own can be ascribed causal responsibility. Such causal responsibility gaps can occur if we solely rely on the FeAR values for individual agents [17,19]. According to the definition of iFeAR as in Eq. (2), the causal responsibility of an agent for it's own trajectory is determined as the complement of the FeAR imposed on it by all other agents ($\text{FeAR}_{j,j} = 1 - \text{FeAR}_{\neg j,j}$). Thus, cases of coupled influence can lead to $\text{FeAR}_{i,j} = 0\, \forall i \in \mathcal{K}$, where no individual agent is causally responsible, but $\neg j$ as a collective has assertive influence on j. If we were to blindly hold all agents in $\neg j$ responsible, this would lead to responsibility gluts "where too many agents are held responsible" [13]. To prevent both causal responsibility gaps and gluts, our tiering algorithm for sorting assertive influences, systematically probes the assertiveness of individual agents and groups to identify minimal groups with assertive influence.

Minimality of the sufficiency set is an important criteria when inferring causality, i.e., all elements much be necessary to cause the outcome [45,25,4]. Instead of focussing on a particular outcome, we are interested in a group's degree of causal influence on the trajectory of an affected agent, which can increase with the number of agents within the group. Using systematic and incremental interventions grounded on the actual joint action of agents, the tiering algorithm is able to unravel the structure of causal influence on affected agents.

Shapley values have been widely used to compute the contributions of individual agents to collective rewards or costs [44]. To compute Shapely values of group FeAR for k agents on an affected agent, we need to consider 2^k counterfactual scenarios. By eliminating courteous agents (with $\text{FeAR}_{i,j} < 0$) and grouping agents in higher tiers, the tiering algorithm potentially saves on computation cost. After identifying all assertive agents using the tiering algorithm, Shapley values could be useful in distributing penalties to individuals.

In this paper, we compute causal responsibility based on feasible action-space reduction (FeAR) which simply looks at how actions of agents reduce the feasible action-space of others by causing collisions. There are other models of responsibility that are base on game-theoretic formulations [13,20], logical formulations of ability and obligations [30,47,49], and probabilistic models of causation [7,24,14,2] — some even consider epistemic states like knowledge and intentions to properly ascribe moral responsibility. Compared to these models of responsibility, FeAR provides a simplistic and model-agnostic metric for causal responsibility, that can be applied to a particular time window of spatial interaction. These simplifications make it an ideal candidate for parsing causal responsibility in large datasets of spatial interactions. Short-listed scenarios with (problematic) causal responsibility ascriptions can be further subjected to more rigorous scrutiny with regard to epistemic, probabilistic and motivational characteristics.

5.2 Metric for emergence in spatial interactions

In our simulations there was little room for emergence as agents did not learn or interact with each other. But by prescribing top-down policies we were hoping to generate "emergence-like" behaviour. In doing so, we have stumbled on some metrics that might be useful in flagging potential emergent behaviours in spatial interactions.

Emergence can be defined as "the appearance of patterns, properties and behaviours within a system that are not evident in individual components [23]." Based on the relationship between the macroscopic property and microscopic factors that cause it, emergence has been classified into *nominal*, *weak* and *strong* emergence [3,41,50]. Traditionally, weak emergence has been quantified using information-theoretic metrics pertaining to the causal relationship between macroscopic properties and microscopic factors [9,41,26,38,23,37], which rely on probabilistic models of how systems evolve over time. Group FeAR, on the other hand, provides a measure of causal responsibility of groups of agents without needing probabilistic or prediction models.

More assertiveness from other agents mean that the trajectory of the affected agent is less dependent on it's own actions and more dependent on the collective actions of others — which necessitates coordination among agents. Unlike the case of weak emergence where properties emerge at macroscopic scales [27,26], the assertiveness of groups of agents act on the trajectories of an affected agent on the same scale which obviates the need to identify the scale at which emergence occurs. The standard deviation of Kendall's τ between the rankings of assertiveness from iFeAR and gFeAR are indicative of the superadditivity of assertive influences in a scenario. The results from the randomised simulations (Section 4.2) showing more group effects for *Aggressive* than *Random*, agree with the intuition that complexity is maximum in the space between complete order and complete chaos [22,10,29,33,16]. Thus, the standard deviation in Kendall's τ comparing the ranks of assertiveness provides a model-agnostic metric for detecting the emergence of complexity in spatial interactions.

5.3 Applications

Group FeAR along with the tiering algorithm can be used to ascribe backward-looking causal responsibility in spatial interactions. As grid-world simulations with discrete actions and non-adaptive agents limit the external validity of our results, future research incorporating continuous formulations of FeAR [19] and agents with explicit normative reflections [34] might be better at explaining empirically observable behaviours in real-life spatial interactions.

As group effects necessitate more coordination between agents, in a decentralised setting, group effects should be minimised. The presence of group effects might warrant interventions in the form of infrastructural changes (barricades, roundabouts, lanes), active monitoring and policing (crowd control, traffic signals), or through training and testing humans (driver's license).

Furthermore, if we know that a group of agents can communicate or coordinate amongst themselves, then group FeAR can use to generate collective actions that maximise courteousness to other agents by minimising group FeAR.

6 Conclusion

We presented a reformulation of the feasible action-space reduction (FeAR) metric to quantify the causal responsibility of groups on the trajectory of an affected agent. Based on marginal changes in FeAR, we identified four types of assertive influences - 'solo', 'mediated', 'coupled' and 'mediated coupled'. Base on these assertive influences, we proposed a tiering algorithm for ranking the assertiveness of agents. Furthermore, through scenario-based simulations, we demonstrated how group FeAR along with the tiering algorithm can be used to identify the emergence of group effects in multi-agent spatial interactions.

Acknowledgements: This work is supported by the TU Delft AI Labs programme.

References

1. Alechina, N., Halpern, J.Y., Logan, B.: Causality, Responsibility and Blame in Team Plans. In: Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems. pp. 1091–1099 (2017)
2. Alechina, N., Halpern, J.Y., Logan, B.: Causality, Responsibility and Blame in Team Plans (May 2020). https://doi.org/10.48550/arXiv.2005.10297
3. Bedau, M.A.: Weak Emergence. Noûs **31**(s11), 375–399 (Jan 1997). https://doi.org/10.1111/0029-4624.31.s11.17
4. Braham, M., Van Hees, M.: Degrees of Causation. Erkenntnis **71**(3), 323–344 (Nov 2009). https://doi.org/10.1007/s10670-009-9184-8
5. Braham, M., Van Hees, M.: Voids or Fragmentation: Moral Responsibility For Collective Outcomes. The Economic Journal **128**(612), F95–F113 (Jul 2018). https://doi.org/10.1111/ecoj.12507
6. Calvert, S.C., Johnsen, S.O., George, A.: Designing automated vehicle and traffic systems towards meaningful human control. In: Research Handbook on Meaningful Human Control of Artificial Intelligence Systems. Edward Elgar Publishing, United Kingdom (2023 (Accepted)). https://doi.org/10.48550/arXiv.2303.05091
7. Chockler, H., Halpern, J.Y.: Responsibility and Blame: A Structural-Model Approach. Journal of Artificial Intelligence Research **22**, 93–115 (Oct 2004). https://doi.org/10.1613/jair.1391
8. Cosner, R.K., Chen, Y., Leung, K., Pavone, M.: Learning responsibility allocations for safe human-robot interaction with applications to autonomous driving. In: Proc. IEEE Conf. on Robotics and Automation (2023)
9. Crutchfield, J.P.: The calculi of emergence: Computation, dynamics and induction. Physica D: Nonlinear Phenomena **75**(1-3), 11–54 (Aug 1994). https://doi.org/10.1016/0167-2789(94)90273-9
10. Crutchfield, J.P., Young, K.: Inferring statistical complexity. Physical Review Letters **63**(2), 105–108 (Jul 1989). https://doi.org/10.1103/PhysRevLett.63.105
11. Dobzinski, S., Mehta, A., Roughgarden, T., Sundararajan, M.: Is Shapley cost sharing optimal? Games and Economic Behavior **108**, 130–138 (Mar 2018). https://doi.org/10.1016/j.geb.2017.03.008
12. Duijf, H.: Responsibility Voids and Cooperation. Philosophy of the Social Sciences **48**(4), 434–460 (Jul 2018). https://doi.org/10.1177/0048393118767084
13. Duijf, H.: A Logical Study of Moral Responsibility. Erkenntnis (Sep 2023). https://doi.org/10.1007/s10670-023-00730-2
14. Engl, F.: A Theory of Causal Responsibility Attribution. SSRN Electronic Journal (2018). https://doi.org/10.2139/ssrn.2932769
15. Geisslinger, M., Poszler, F., Lienkamp, M.: An ethical trajectory planning algorithm for autonomous vehicles. Nature Machine Intelligence **5**(2), 137–144 (Feb 2023). https://doi.org/10.1038/s42256-022-00607-z
16. Gell-Mann, M., Lloyd, S.: Information measures, effective complexity, and total information. Complexity **2**(1), 44–52 (Sep 1996). https://doi.org/10.1002/(SICI)1099-0526(199609/10)2:1<44::AID-CPLX10>3.0.CO;2-X
17. George, A., Cavalcante Siebert, L., Abbink, D., Zgonnikov, A.: Feasible Action-Space Reduction as a Metric of Causal Responsibility in Multi-Agent Spatial Interactions. In: Gal, K., Nowé, A., Nalepa, G.J., Fairstein, R., Rădulescu, R. (eds.) Frontiers in Artificial Intelligence and Applications, Frontiers in Artificial Intelligence and Applications, vol. 372, pp. 819–826. IOS Press (Sep 2023). https://doi.org/10.3233/FAIA230349

18. George, A., Siebert, L.C., Abbink, D., Zgonnikov, A.: Feasible Action-Space Reduction as a Metric of Causal Responsibility in Multi-Agent Spatial Interactions (2023 (ECAI 2023 Accepted))
19. George, A., Siebert, L.C., Abbink, D.A., Zgonnikov, A.: Feasible Action Space Reduction for Quantifying Causal Responsibility in Continuous Spatial Interactions (May 2025). https://doi.org/10.48550/arXiv.2505.17739
20. Gladyshev, M., Alechina, N., Dastani, M., Doder, D.: Group Responsibility for Exceeding Risk Threshold. In: Proceedings of the Twentieth International Conference on Principles of Knowledge Representation and Reasoning. pp. 322–332. International Joint Conferences on Artificial Intelligence Organization, Rhodes, Greece (Sep 2023). https://doi.org/10.24963/kr.2023/32
21. Goetze, T.S.: Mind the Gap: Autonomous Systems, the Responsibility Gap, and Moral Entanglement. In: 2022 ACM Conference on Fairness Accountability and Transparency. pp. 390–400. ACM, Seoul Republic of Korea (Jun 2022). https://doi.org/10.1145/3531146.3533106
22. Grassberger, P.: Toward a quantitative theory of self-generated complexity. International Journal of Theoretical Physics **25**(9), 907–938 (Sep 1986). https://doi.org/10.1007/BF00668821
23. Green, D.G.: Emergence in complex networks of simple agents. Journal of Economic Interaction and Coordination **18**(3), 419–462 (Jul 2023). https://doi.org/10.1007/s11403-023-00385-w
24. Halpern, J.Y.: Cause, responsibility and blame: A structural-model approach. Law, Probability and Risk **14**(2), 91–118 (Jun 2015). https://doi.org/10.1093/lpr/mgu020
25. Halpern, J.Y.: A modification of the halpern-pearl definition of causality. In: Proceedings of the 24th International Conference on Artificial Intelligence. pp. 3022–3033. IJCAI'15, AAAI Press, Buenos Aires, Argentina (2015)
26. Hoel, E.: When the Map Is Better Than the Territory. Entropy **19**(5), 188 (Apr 2017). https://doi.org/10.3390/e19050188
27. Hoel, E.P., Albantakis, L., Tononi, G.: Quantifying causal emergence shows that macro can beat micro. Proceedings of the National Academy of Sciences **110**(49), 19790–19795 (Dec 2013). https://doi.org/10.1073/pnas.1314922110
28. Kendall, M.G.: A New Measure of Rank Correlation. Biometrika **30**(1/2), 81 (Jun 1938). https://doi.org/10.2307/2332226
29. Langton, C.G.: Computation at the edge of chaos: Phase transitions and emergent computation. Physica D: Nonlinear Phenomena **42**(1-3), 12–37 (Jun 1990). https://doi.org/10.1016/0167-2789(90)90064-V
30. Lorini, E., Longin, D., Mayor, E.: A logical analysis of responsibility attribution: Emotions, individuals and collectives. Journal of Logic and Computation **24**(6), 1313–1339 (Dec 2014). https://doi.org/10.1093/logcom/ext072
31. Markkula, G., Madigan, R., Nathanael, D., Portouli, E., Lee, Y.M., Dietrich, A., Billington, J., Schieben, A., Merat, N.: Defining interactions: A conceptual framework for understanding interactive behaviour in human and automated road traffic. Theoretical Issues in Ergonomics Science **21**(6), 728–752 (Nov 2020). https://doi.org/10.1080/1463922X.2020.1736686
32. Matthias, A.: The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology **6**(3), 175–183 (2004). https://doi.org/10.1007/s10676-004-3422-1
33. Mitchell, M., Crutchfield, J.P., Hraber, P.T.: Evolving cellular automata to perform computations: Mechanisms and impediments. Physica D: Nonlinear Phenomena **75**(1-3), 361–391 (Aug 1994). https://doi.org/10.1016/0167-2789(94)90293-3

34. Morales, J., Lopez-Sanchez, M., Rodriguez-Aguilar, J.A., Wooldridge, M., Vasconcelos, W.: Automated synthesis of normative systems. In: Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems. pp. 483–490. AAMAS '13, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2013)
35. Papadimitriou, E., Farah, H., van de Kaa, G., Santoni de Sio, F., Hagenzieker, M., van Gelder, P.: Towards common ethical and safe 'behaviour' standards for automated vehicles. Accident Analysis & Prevention **174**, 106724 (Sep 2022). https://doi.org/10.1016/j.aap.2022.106724
36. Remy, I., Fridovich-Keil, D., Leung, K.: Learning responsibility allocations for multi-agent interactions: A differentiable optimization approach with control barrier functions (Oct 2024)
37. Rodríguez-Falcón, S., Stucchi, L.: Quantifying Emergent Behaviors in Agent-Based Models using Mean Information Gain (Oct 2025). https://doi.org/10.48550/arXiv.2510.10381
38. Rosas, F.E., Mediano, P.A.M., Jensen, H.J., Seth, A.K., Barrett, A.B., Carhart-Harris, R.L., Bor, D.: Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. PLOS Computational Biology **16**(12), e1008289 (Dec 2020). https://doi.org/10.1371/journal.pcbi.1008289
39. Santoni De Sio, F., Mecacci, G.: Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. Philosophy & Technology **34**(4), 1057–1084 (Dec 2021). https://doi.org/10.1007/s13347-021-00450-x
40. Schwarting, W., Pierson, A., Alonso-Mora, J., Karaman, S., Rus, D.: Social behavior for autonomous vehicles. Proceedings of the National Academy of Sciences **116**(50), 24972–24978 (Dec 2019). https://doi.org/10.1073/pnas.1820676116
41. Seth, A.K.: Measuring Autonomy and Emergence via Granger Causality. Artificial Life **16**(2), 179–196 (Apr 2010). https://doi.org/10.1162/artl.2010.16.2.16204
42. Shalev-Shwartz, S., Shammah, S., Shashua, A.: On a Formal Model of Safe and Scalable Self-driving Cars (Oct 2018)
43. Shalev-Shwartz, S., Shammah, S., Shashua, A.: Vision Zero: On a Provable Method for Eliminating Roadway Accidents without Compromising Traffic Throughput (Jan 2019)
44. Shapley, L.S.: A Value for n-Person Games. In: Kuhn, H.W., Tucker, A.W. (eds.) Contributions to the Theory of Games (AM-28), Volume II, pp. 307–318. Princeton University Press (Dec 1953). https://doi.org/10.1515/9781400881970-018
45. Triantafyllou, S., Singla, A., Radanovic, G.: Actual Causality and Responsibility Attribution in Decentralized Partially Observable Markov Decision Processes. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. pp. 739–752. ACM, Oxford United Kingdom (Jul 2022). https://doi.org/10.1145/3514094.3534133
46. Veluwenkamp, H.: What responsibility gaps are and what they should be. Ethics and Information Technology **27**(1), 14 (Mar 2025). https://doi.org/10.1007/s10676-025-09823-8
47. Yazdanpanah, V., Dastani, M.: Distant Group Responsibility in Multi-agent Systems. In: Baldoni, M., Chopra, A.K., Son, T.C., Hirayama, K., Torroni, P. (eds.) PRIMA 2016: Princiles and Practice of Multi-Agent Systems, vol. 9862, pp. 261–278. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-44832-9_16
48. Yazdanpanah, V., Dastani, M., Jamroga, W., Alechina, N., Logan, B.: Strategic Responsibility Under Imperfect Information (2019)

49. Yazdanpanah, V., Stein, S., Gerding, E.H., Jennings, N.R.: Applying strategic reasoning for accountability ascription in multiagent teams. In: Espinoza, H., McDermid, J.A., Huang, X., Castillo-Effen, M., Chen, X.C., Hernández-Orallo, J., hÉigeartaigh, S.Ó., Mallah, R., Pedroza, G. (eds.) Proceedings of the Workshop on Artificial Intelligence Safety 2021 Co-Located with the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI 2021), Virtual, August, 2021. CEUR Workshop Proceedings, vol. 2916. CEUR-WS.org (2021)
50. Yuan, B., Zhang, J., Lyu, A., Wu, J., Wang, Z., Yang, M., Liu, K., Mou, M., Cui, P.: Emergence and Causality in Complex Systems: A Survey of Causal Emergence and Related Quantitative Studies. Entropy **26**(2), 108 (Jan 2024). https://doi.org/10.3390/e26020108