Mechanism Design for Defense Coordination in Security Games

Jiarui Gan
University of Oxford
Oxford, United Kingdom
jiarui.gan@cs.ox.ac.uk

Edith Elkind
University of Oxford
Oxford, United Kingdom
elkind@cs.ox.ac.uk

Sarit Kraus
Bar-Ilan University
Ramat Gan, Israel
sarit@cs.biu.ac.il

Michael Wooldridge
University of Oxford
Oxford, United Kingdom
mjw@cs.ox.ac.uk

ABSTRACT

Recent work studied Stackelberg security games with multiple defenders, in which heterogeneous defenders allocate security resources to protect a set of targets against a strategic attacker. Equilibrium analysis was conducted to characterize outcomes of these games when defenders act independently. Our starting point is the observation that the use of resources in equilibria may be inefficient due to lack of coordination. We explore the possibility of reducing this inefficiency by coordinating the defenders—specifically, by pooling the defenders' resources and allocating them jointly. The defenders' heterogeneous preferences then give rise to a collective decision-making problem, which calls for a mechanism to generate joint allocation strategies. We seek a mechanism that encourages coordination, produces efficiency gains, and incentivizes the defenders to report their true preferences and to execute the recommended strategies. Our results show that, unfortunately, even these basic properties clash with each other and no mechanism can achieve them simultaneously, which reveals the intrinsic difficulty of achieving meaningful defense coordination in security games. On the positive side, we put forward mechanisms that fulfill some of these properties and we identify special cases of our setting where more of these properties are compatible.

KEYWORDS

Stackelberg Security games; coordination; mechanism design

ACM Reference Format:
Jiarui Gan, Edith Elkind, Sarit Kraus, and Michael Wooldridge. 2020. Mechanism Design for Defense Coordination in Security Games. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 9 pages.

1 INTRODUCTION

A group of defenders are patrolling the sea to combat their common enemy Captain Jack, a pirate who operates constantly in the patrolled area. Each defender acts independently, ordering their vessels to patrol at certain points of interest, a.k.a. *targets*, where Captain Jack might be looting. The choice of targets for day-to-day patrol is randomized. As a sophisticated pirate, Captain Jack plans carefully: he surveys the randomized patrolling patterns and chooses the best target to attack, taking into consideration both the frequency with which each target is patrolled and the profits from attacking this target. Eventually, the defenders realize that independent movement may be rather inefficient. For example, when a target t is patrolled by vessels of two independent defenders, each with probability 0.5, the probability that t is patrolled by some vessel is $1-(1-0.5)^2 = 0.75$, and with probability $0.5 \times 0.5 = 0.25$ two vessels appear at t simultaneously. Normally, one vessel is already capable of deterring an attack in confrontation with Captain Jack, so the latter scenario is a waste of resources. To increase efficiency, the defenders can reduce the unnecessary double protection by synchronising the defenders' movements, so that at any time exactly one vessel patrols at t. Then t would be covered with probability 1, while each defender would still contribute the same effort as before. This example suggests that it can be very beneficial for a group of defenders who face a common adversary to collaborate.

Unfortunately, coordination may be much more challenging in real life because defenders may have different valuations to the targets. Thus, each defender would prefer the group to adopt a different joint strategy, so as to better protect targets that are more important to them; the problem of choosing a joint strategy becomes very difficult. *In this paper, we show that, indeed, collaboration among defenders is often not possible.*

Specifically, we consider a setting where defenders pool their resources and allocate them jointly according to a strategy that is decided by a coordination mechanism. The defenders' utilities from this strategy are determined by a Stackelberg game model, in which an adversary is expected to best-respond to the resource allocation, attacking the target that maximizes his utility. We aim at designing a coordination mechanism that takes into account the defenders' individual preferences, one that has the following natural properties.

- **Efficiency**. We want the mechanism to exploit all possibilities of efficiency improvement. Hence, we require the allocation strategies it generates to be *Pareto efficient*, so that no other allocation can improve the social welfare while not making any defender worse off.
- **Utility guarantee**. We want the mechanism to ensure that every defender benefits from coordination, so that all defenders are incentivized to participate. Hence, we require the mechanism to guarantee every defender at least their utility in the uncoordinated situation, as captured by a Nash-like equilibrium concept proposed by Gan et al. [10].

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

- **Strategyproofness**. Since collective decision-making relies on information about the defenders' preferences (valuations to the targets), which are held private, we want our mechanism to incentivize every defender to report their true preferences.

In some scenarios the mechanism can only recommend strategies for the defenders, but cannot force the defenders to actually use the recommended strategies. The following additional property is then relevant:

- **Individual rationality**. We want the recommended strategies to form an equilibrium in which no defender has an incentive to modify their strategy.

Beyond hunting for Captain Jack, our model also abstracts scenarios where heterogenous security agencies combat common enemies in the same area, such as patrols along country borders or over international waters carried out by different countries, or co-located operations of police force and private security personnel.

1.1 Our Contribution

Our results shed light on the difficulties of achieving defense coordination in this seemingly cooperative setting. Our first key result is that it is impossible for any mechanism to achieve the above-mentioned basic properties simultaneously. This theoretical barrier then drives us to explore mechanisms that fulfill a subset of these properties or special problem settings that allow more of them to coexist. We show that when strategyproofness is not a concern, there is a mechanism that achieves Pareto efficiency and utility guarantee with respect to the worst-case utility in the uncoordinated situation. Further, when each defender's preference concerning the attacker's response does not depend on how the targets are protected (a special case introduced in [10]), knowing a status-quo equilibrium enables us to design a mechanism that achieves the first three of our properties. We also formulate an individual rationality property, and define a broader class of mechanisms that recommend strategies for the defenders to take. We show that even when strategyproofness is set aside, this additional property is incompatible with efficiency and utility guarantee. Our results indicate that voluntary coordination among defenders may be insufficient to achieve efficient resource utilization in security domains.

1.2 Related Work

Stackelberg security games (SSG) provide a model to study security resource allocation problems. Automated tools that use SSGs have been deployed in many real-world scenarios, such as protection of high-profile infrastructure, transportation systems, and public and natural resources (see, e.g., [2, 22] for an overview). Most SSG models involve only one defender, and then the goal is to maximize the defender's utility [8, 18, 19]. Some recent works studied models with multiple defenders. The problems where the defenders have the same utility function [4, 12] can be studied within the single-defender framework with additional allocation constraints. When defenders are heterogeneous, the outcome of independent defender actions can be predicted by means of equilibrium analysis [10, 13–16, 21]. In our paper, we follow the model of Gan et al. [10], which is arguably the most basic multi-defender model extending the standard SSG; other multi-defender SSG models either involve additional assumptions (e.g., defenders protect disjoint subsets of targets [14, 15]) or are intended for more specific applications (e.g., defense against spear-phishing attacks or traffic light control [13, 16]).

Very recently, Castiglioni et al. [7] studied the problem of computing correlated strategies for a group of leaders to commit to in generic Stackelberg games. Somewhat unusually, in their model a leader has the choice to deviate from the correlated strategy so as to act as a follower in the game; such deviations would not be feasible in many security scenarios that we would like to model. Further, while the idea of correlating the leaders' strategies is conceptually similar to our approach, Castiglioni et al. [7] did not consider strategyproofness and utility guarantee, so their results are very different from ours. Under their solution, players may be incentivized to misreport their preferences to the system and, hence, make other players worse off than even in the uncoordinated situation. In addition, their solution may, in a sense, force players to coordinate by threatening them with a worse outcome if they choose not to; we feel that this approach is not compatible with our agenda.

Our work is related to collective decision-making for public resource allocation, see, e.g., [1, 3, 5, 9]. We distribute security resources which form a divisible probability mass in the mixed strategy setting. Other than the specific utility structure that is realized through the SSG model, one feature of our problem that distinguishes it from the prior work is that resources are contributed by the players, who need additional incentives to agree to cooperate in the first place. A similar phenomenon can be observed in the donor coordination problem studied by Brandl et al. [6], in which philanthropists with different preferences pool their resources and jointly distribute them to charity projects. However, their model is different, and they do not use equilibria of the uncoordinated situation as benchmarks for designing mechanisms.

2 THE MODEL

There are n *defenders* $1, \dots, n$, who want to protect a set $T = \{t_1, \dots, t_m\}$ of m *targets* against an *attacker*. We write $[n] = \{1, \dots, n\}$. Each defender $i \in [n]$ has $k_i \in \mathbb{N}_{\geq 0}$ *security resources* that can be allocated to the targets; let $K = \sum_{i \in [n]} k_i$. A target is said to be protected, or covered, if *at least one* resource is allocated to it; and unprotected, or uncovered, otherwise.

In the pure strategy setting, an attack on a protected target $t \in T$ will be unsuccessful and results in the attacker receiving a *penalty* value $p^{\mathrm{a}}(t)$, and each defender i receiving a *reward* value $r_i^{\mathrm{d}}(t)$. An attack on an unprotected target will be successful, resulting in a reward $r^{\mathrm{a}}(t)$ for the attacker, and a penalty $p_i^{\mathrm{d}}(t)$ for each defender i. It is assumed that $r_i^{\mathrm{d}}(t) > p_i^{\mathrm{d}}(t)$ and $r^{\mathrm{a}}(t) > p^{\mathrm{a}}(t)$ for each $i \in [n]$ and $t \in T$, so each defender prefers an attack to be unsuccessful, and the attacker prefers the opposite. Thus, although defenders are heterogeneous, they all prefer every target to be safe. The payoff parameters define the *type* of a player. Let $\theta_i = (r_i^{\mathrm{d}}, p_i^{\mathrm{d}})$ denote the type of defender i, and $\alpha = (r^{\mathrm{a}}, p^{\mathrm{a}})$ the type of the attacker. The tuple $(\boldsymbol{\theta}, \mathbf{k})$, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$ and $\mathbf{k} = (k_1, \dots, k_n)$ is called a *defender profile*. Let $\Theta = \left\{(r, p) \in \mathbb{R}^{2m} : r(t) > p(t) \text{ for all } t \in T\right\}$ denote the set of valid player types that are consistent with our assumption above.

Coordination Mechanism. The defenders want to coordinate their resource allocation. They delegate their resources to a coordination authority, which then chooses an allocation strategy to implement according to a *coordination mechanism* π. The mechanism takes as input a defender profile $\boldsymbol{\theta}$ and an attacker type α, and outputs an allocation of K resources. By the payoff structure of the game, the rational ways of resource allocation correspond to subsets of T that contain at most K targets; we denote by $\mathcal{T}_K$ the collection of these subsets. Given a pure strategy $s \in \mathcal{T}_K$, we write $\mathrm{cov}_t(s) = 1$ if s allocates some resource to target t and $\mathrm{cov}_t(s) = 0$ otherwise. More generally, the defenders can employ a *mixed* strategy, randomizing the resource allocation according to a distribution $\mathbf{x} \in \Delta(\mathcal{T}_K)$ over pure strategies. Given a mixed strategy $\mathbf{x}$, each target t is protected with probability $\mathrm{cov}_t(\mathbf{x})$, which we call the *coverage*; we have

$$\mathrm{cov}_t(\mathbf{x}) = \sum_{s \in \mathcal{T}_K} x_s \cdot \mathrm{cov}_t(s), \tag{1}$$

and call $\mathrm{cov}(\mathbf{x}) = \big(\mathrm{cov}_{t_1}(\mathbf{x}), \ldots, \mathrm{cov}_{t_m}(\mathbf{x})\big)$ the *coverage (vector)*.

Now suppose the attacker attacks target t and the defenders employ a strategy $\mathbf{x}$ that results in coverage $\mathbf{c} = \mathrm{cov}(\mathbf{x})$. Then the expected utilities $U_i^{\mathrm{d}}(\mathbf{c}, t)$ of each defender $i \in [n]$ and the expected utility $U^{\mathrm{a}}(\mathbf{c}, t)$ of the attacker are given below:

$$U_i^{\mathrm{d}}(\mathbf{c}, t) = c_t \cdot r_i^{\mathrm{d}}(t) + (1 - c_t) \cdot p_i^{\mathrm{d}}(t);$$
$$U^{\mathrm{a}}(\mathbf{c}, t) = (1 - c_t) \cdot r^{\mathrm{a}}(t) + c_t \cdot p^{\mathrm{a}}(t).$$

By our assumption about the payoffs, $U_i^{\mathrm{d}}(\mathbf{c}, t)$ is strictly increasing with respect to c_t, and $U^{\mathrm{a}}(\mathbf{c}, t)$ strictly decreasing. As in a standard SSG, the attacker observes the mixed resource allocation strategy $\mathbf{c}$, and responds by attacking a target that maximizes his utility against $\mathbf{c}$. Let

$$\mathrm{BR}(\mathbf{c}) = \arg\max_{t \in T} U^{\mathrm{a}}(\mathbf{c}, t)$$

denote the set of attacker best responses. The exact best response the attacker will choose depends on the tie-breaking rule which we will define next.

Hereafter, we will often refer to a mixed strategy $\mathbf{x}$ and its induced coverage $\mathbf{c}$ interchangeably. When no additional allocation constraint is imposed, the set of feasible coverage vectors is $C_K = \{\mathbf{c} \in \mathbb{R}^m : 0 \le c_t \le 1, \sum_{t \in T} c_t = K\}$, i.e., $\mathbf{c} \in C_K$ if and only if $\mathrm{cov}(\mathbf{x}) = \mathbf{c}$ for some $\mathbf{x} \in \Delta(\mathcal{T}_K)$. We will write $f(\mathbf{x}) = f(\mathrm{cov}(\mathbf{x}))$ for any function f that is supposed to take as input a coverage vector, e.g., $U_i^{\mathrm{d}}(\mathbf{x}, t) = U_i^{\mathrm{d}}(\mathrm{cov}(\mathbf{x}), t)$ and $\mathrm{BR}(\mathbf{x}) = \mathrm{BR}(\mathrm{cov}(\mathbf{x}))$.

Tie-breaking Rule. In the single-defender model, the standard solution concept is the *strong Stackelberg equilibrium (SSE)*, which assumes that the attacker breaks ties by choosing an action in $\mathrm{BR}(\mathbf{c})$ to favor the defender. In our setting, we assume that the mechanism explicitly specifies an attacker response in $\mathrm{BR}(\mathbf{c})$. This assumption has the same justification as the optimistic tie-breaking rule in the single-defender setting: a mechanism can induce the attacker's *strict* preference for any target in $\mathrm{BR}(\mathbf{c})$ by reducing protection of that target by an infinitesimal amount, *irrespective of the actual tie-breaking behavior of the attacker*. Thus, we require a coordination mechanism to output a tuple $(\mathbf{x}, t)$, called an *outcome*. An outcome is feasible if and only if $\mathbf{x} \in \Delta(\mathcal{T}_K)$ and $t \in \mathrm{BR}(\mathrm{cov}(\mathbf{x}))$. Our goal in this paper is to find a mechanism $\pi : (\boldsymbol{\theta}, \mathbf{k}, \alpha) \mapsto (\mathbf{x}, t)$ that satisfies certain properties which we will discuss next.

2.1 Desired Properties

First, we want the mechanism to achieve maximal efficiency improvement. Thus, the outcomes it produces should be *Pareto efficient*, so that no defender's utility can be further improved without making any other defender worse off.

Property 1 (Pareto Efficiency (PE)). An outcome $(\mathbf{x}, t)$ *Pareto dominates* another outcome $(\mathbf{x}', t')$ if $U_i^{\mathrm{d}}(\mathbf{x}, t) \ge U_i^{\mathrm{d}}(\mathbf{x}', t')$ for all $i \in [n]$ and $U_i^{\mathrm{d}}(\mathbf{x}, t) > U_i^{\mathrm{d}}(\mathbf{x}', t')$ for some $i \in [n]$. A mechanism π is *Pareto efficient* if for any $\mathbf{k} \in \mathbb{N}^n$, $\boldsymbol{\theta} \in \Theta^n$, and $\alpha \in \Theta$ no feasible outcome Pareto dominates $\pi(\boldsymbol{\theta}, \mathbf{k}, \alpha)$.

For coordination to be meaningful, it is desired that every defender benefits from it. We take the defenders' equilibrium utilities in the uncoordinated situation as our benchmark. In particular, Gan et al. [10] introduced an equilibrium concept called the *Nash Stackelberg equilibrium (NSE)* for multi-defender SSGs. It combines the ideas of the Nash equilibrium and the Stackelberg equilibrium: in an NSE no defender has any incentive to change their strategy, assuming that the other defenders stick to their strategies, while the attacker responds optimally to the modified strategy profile. It is shown that, though an exact NSE may not exist, an ϵ-NSE (where no defender has a deviation that improves her utility by more than ϵ) exists for every $\epsilon > 0$, and so does the limit point of ϵ-NSEs when ϵ approaches 0; therefore, it is proposed to use the limit points as the solution concept for multi-defender SSGs. Following this approach, we adopt this solution concept to describe the outcome of the uncoordinated situation. For simplicity, we will refer to this solution concept as the NSE (Gan et al. refer to it as 0^+-NSE).

In the uncoordinated setting, let $\mathbf{x}^i \in \Delta(\mathcal{T}_{k_i})$ be the strategy of each defender $i \in [n]$, and let $X = (\mathbf{x}^1, \ldots, \mathbf{x}^n)$ be the strategy profile. Each defender strategy $\mathbf{x}^i$ results in coverage $\mathrm{cov}_t(\mathbf{x}^i) = \sum_{s \in \mathcal{T}_{k_i}} x_s \cdot \mathrm{cov}_t(s) \in C_{k_i}$ for each target t as defined in (1), and jointly the strategy profile results in the following overall coverage for each t (with slight abuse of notation):

$$\mathrm{cov}_t(X) = 1 - \prod_{i \in [n]} (1 - \mathrm{cov}_t(\mathbf{x}^i)). \tag{2}$$

Namely, a target is covered as long as some defender is protecting it. We define the NSE and our next property—*utility preservation*.

Definition 2.1 (NSE). *A defender strategy profile* $X = (\mathbf{x}^1, \ldots, \mathbf{x}^n)$ *and an attacker best response* $t \in \mathrm{BR}(X)$ *form an NSE if for every defender* $i \in [n]$ *and every strategy deviation* $\mathbf{x}' \in \Delta(\mathcal{T}_{k_i})$, *it holds that* $\min_{t' \in \mathrm{BR}(\langle \mathbf{x}^{-i}, \mathbf{x}' \rangle)} U_i^{\mathrm{d}}(\langle \mathbf{x}^{-i}, \mathbf{x}' \rangle, t') \le U_i^{\mathrm{d}}(X, t)$.

Property 2 (Utility Preservation (UP)). A mechanism π is *strongly (resp. weakly) utility preserving* if for any $\mathbf{k} \in \mathbb{N}^n$, $\boldsymbol{\theta} \in \Theta^n$, and $\alpha \in \Theta$ the utility $U_i^{\mathrm{d}}(\pi(\boldsymbol{\theta}, \mathbf{k}, \alpha))$ of each defender $i \in [n]$ is at least their utility in the best (resp. worst) NSE.

We note that Definition 2.1 is more concise than the original definition of Gan et al. [10], but they are equivalent: an NSE (X, t) defined above is exactly the limit point of a series of ϵ-NSE with $\epsilon \to 0$, that are induced by letting the defenders reduce their protection to target t by a small amount $\delta \to 0$ (which makes t the only best response of the attacker). The pessimistic tie-breaking assumption

the defenders make when evaluating the benefits of a potential deviation (i.e., $\min_{t' \in \mathrm{BR}(\langle \mathbf{x}^{-i}, \mathbf{x}' \rangle)} U_i^{\mathrm{d}}(\langle \mathbf{x}^{-i}, \mathbf{x}' \rangle, t')$) is appropriate for the uncoordinated situation in which there is no consensus among the defenders about how an attacker response can actually be induced. It does not conflict our previously defined tie-breaking rule *for the mechanism*, under which outcomes of the mechanism can always be induced irrespective of the actual tie-breaking behavior of the attacker. We refer the reader to the paper by Gan et al. [10] for a detailed discussion about the tie-breaking assumptions.

Finally, since the defenders' types (payoffs) are private information, we would like our mechanism to incentivize every defender to report their true type. The *strategyproofness* property requires that no defender can improve her utility by non-truthful reporting.

Property 3 (Strategyproofness (SP)). A mechanism π is *not strategyproof* if for some $\mathbf{k} \in \mathbb{N}^n$, $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta^n$, $\alpha \in \Theta$, and $i \in [n]$, we have $\theta'_j = \theta_j$ for all $j \in [n] \setminus \{i\}$ and $U_i^{\mathrm{d}}(\pi(\boldsymbol{\theta}, \mathbf{k}, \alpha)) < U_i^{\mathrm{d}}(\pi(\boldsymbol{\theta}', \mathbf{k}, \alpha))$, where each U_i^{d} is the utility function of defender type θ_i. Otherwise, π is strategyproof.

3 IMPOSSIBILITY RESULTS

Unfortunately, the basic properties defined so far clash with each other. In this section, we show several impossibility results. We begin by introducing *non-overlapping payoffs*, a special class of defender types that will be useful for our proofs.

Definition 3.1 (Non-overlapping payoffs). *A defender type $(r_i^{\mathrm{d}}, p_i^{\mathrm{d}})$ is said to have* non-overlapping payoffs *if for any $t, t' \in T$, either $r_i^{\mathrm{d}}(t) < p_i^{\mathrm{d}}(t')$ or $r_i^{\mathrm{d}}(t') < p_i^{\mathrm{d}}(t)$.*

When a defender has non-overlapping payoffs, her preference concerning the attacker's response does not depend on the coverage vector. Indeed, under our assumption that $r_i^{\mathrm{d}}(t) > p_i^{\mathrm{d}}(t)$ for all t, the utility $U_i^{\mathrm{d}}(\mathbf{c}, t)$ should always lie in $[p_i^{\mathrm{d}}(t), r_i^{\mathrm{d}}(t)]$ for any $\mathbf{c}$. Thus, if $r_i^{\mathrm{d}}(t) < p_i^{\mathrm{d}}(t')$, we have $U_i^{\mathrm{d}}(\mathbf{c}, t) \le r_i^{\mathrm{d}}(t) < p_i^{\mathrm{d}}(t') \le U_i^{\mathrm{d}}(\mathbf{c}', t')$, irrespective of $\mathbf{c}$ and $\mathbf{c}'$, so the defender always prefers the attacker to go for t'; we write $t \prec_i t'$ in this case. Likewise, if $r_i^{\mathrm{d}}(t') < p_i^{\mathrm{d}}(t)$, we have $U_i^{\mathrm{d}}(\mathbf{c}', t') < U_i^{\mathrm{d}}(\mathbf{c}, t)$ and we write $t' \prec_i t$. The defender's utility is then captured by the preference order $\prec_i$ over T.

3.1 Strong UP Cannot be Guaranteed

We first show that strong UP cannot be satisfied on its own.

Theorem 3.2. *No mechanism is strongly UP.*

Proof. It suffices to show a game with no strongly UP outcome. Consider a game with two defenders and two targets $T = \{a, b\}$. Each defender has one resource, and both of them have non-overlapping payoffs captured by preference orders $a \prec_1 b$ and $b \prec_2 a$. The targets are identical to the attacker, so the attacker's best response is always the target(s) with the lowest coverage.

Suppose that defender 1 always protects a and defender 2 always protects b. The coverage contributed by the defenders is $\mathbf{c}^1 = (c_a^1, c_b^1) = (1, 0)$ and $\mathbf{c}^2 = (c_a^2, c_b^2) = (0, 1)$, respectively. Overall, both targets are fully protected and are the attacker's best responses. Let the joint strategy be $C = (\mathbf{c}^1, \mathbf{c}^2)$. It is easy to see that (C, b) is an NSE. Indeed, for defender 1 this is already the best outcome. Defender 2 prefers the attacker to choose a, but she cannot induce this attacker action: she can neither increase the coverage of b to make it less attractive to the attacker nor decrease the coverage of a to make it more attractive. Similarly, (C, a) is also an NSE, and it is the best for defender 2.

Now suppose some outcome $(\mathbf{z}, t)$ is strongly UP. If $t = a$, we have $U_1^{\mathrm{d}}(\mathbf{z}, t) < U_1^{\mathrm{d}}(\mathbf{z}, b) \le U_1^{\mathrm{d}}(C, b)$, so defender 1 would be worse off than in her most preferred NSE. By symmetry, if $t = b$, defender 2 would be worse off than in her most preferred NSE. Thus, no outcome of this game is strongly UP. □

3.2 SP and Weak UP are Incompatible

As strong UP turns out to be too demanding, we focus on weak UP. Trivially, every NSE (X, t) can be turned into an outcome $(\mathrm{cov}(X), t)$ (by simulating independent defender movement specified by X) that is a weakly UP by definition, so a weakly UP outcome always exists. Nevertheless, the weak UP property clashes with SP.

Theorem 3.3. *If $|T| \ge 3$, there exists no weakly UP mechanism that satisfies strategyproofness, even when the defenders' payoffs are restricted to non-overlapping payoffs.*

To prove Theorem 3.3, we show that for $|T| \ge 3$ a strategyproof coordination mechanism has to be a *dictatorship*, i.e., it has to always output the favorite outcome of a distinguished defender (the dictator). While a dictatorship is a valid coordination mechanism, its output will not be acceptable to a defender whose preference conflicts with that of the dictator; indeed, under this mechanism her utility may be lower than in her worst NSE. Our argument proceeds by transforming a coordination mechanism into a social choice function, and invoking the famous Gibbard–Satterthwaite theorem [11, 20], which states that for more than two alternatives any onto strategyproof social choice function is a dictatorship.

Let g be a function that maps a linear order $\prec$ over a candidate set T to a defender type $(r^{\mathrm{d}}, p^{\mathrm{d}})$, such that for each $t \in T$,

$$r^{\mathrm{d}}(t) = j_t + \frac{1}{2} \quad \text{and} \quad p^{\mathrm{d}}(t) = j_t,$$

where $j_t = |\{\tau \in T : \tau \prec t\}|$ is the index of t in $\prec$. Hence, a defender i of type $g(\prec)$ has non-overlapping payoffs captured by $\prec$: for any coverage $\mathbf{c}$ and $\mathbf{c}'$ it holds that $U_i^{\mathrm{d}}(\mathbf{c}, t) \le j_t + \frac{1}{2} < j_{t'} \le U_i^{\mathrm{d}}(\mathbf{c}', t')$ if $t \prec t'$. Overloading this notation, given a preference profile $(\prec_1, \ldots, \prec_n)$, we write $g(\prec_1, \ldots, \prec_n) = (g(\prec_1), \ldots, g(\prec_n))$.

Given a mechanism π, we construct a social choice function f_π: for any preference profile $(\prec_1, \ldots, \prec_n)$, we let $f_\pi(\prec_1, \ldots, \prec_n) = t$ if $\pi(g(\prec_1, \ldots, \prec_n), \mathbf{k}, \alpha^*) = (\mathbf{x}, t)$ for some $\mathbf{x}$, where

- $\mathbf{k}$ is a defender resource profile such that $k_1 = k_2 = 1$ and $k_i = 0$ for all $i = 3, \ldots, n$ (hence, $K = 2$).
- $\alpha^* = (r^{\mathrm{a}}, p^{\mathrm{a}})$ is an attacker type such that $r^{\mathrm{a}}(t) = 0$ and $p^{\mathrm{a}}(t) = -1$ for all $t \in T$ (so all the targets are identical to α^* and $U^{\mathrm{a}}(\mathbf{c}, t) = -c_t$).

For the Gibbard–Satterthwaite theorem to be applicable, we need to show that f_π is *onto* T, i.e., for every $t \in T$ there exists an input $(\prec_1, \ldots, \prec_n)$ such that $f(\prec_1, \ldots, \prec_n) = t$.

Lemma 3.4. *If π is weakly UP then f_π is onto T.*

Proof. We will show that f_π satisfies a weak form of unanimity, i.e., for any $\tau \in T$ and any linear order $\prec$ such that $t \prec \tau$ for all $t \in T \setminus \{\tau\}$ it holds that $f_\pi(\prec, \ldots, \prec) = \tau$.

To this end, we consider a game where the defender profile is given by $(g(\prec,\dots,\prec),\mathbf{k})$ and the attacker type is α^*. The core of the proof is then to argue the following: *if any strategy profile (X,t) forms an NSE in this game, then $t=\tau$.* As long as this claim holds, the assumption that π is weakly UP will imply that $\pi(g(\prec,\dots,\prec),\mathbf{k},\alpha^*)=(\mathbf{x},\tau)$ for some $\mathbf{x}$. Indeed, since $\tau'\prec\tau$ by our assumption, any outcome $(\mathbf{x}',\tau')$, $\tau'\neq\tau$ would give $U_i^{\mathrm{d}}(\mathbf{x}',\tau')<U_i^{\mathrm{d}}(Y,\tau)$ even for the worst NSE (Y,τ) for defender i; a mechanism that outputs $(\mathbf{x}',\tau')$ cannot be weakly UP. Thus, we have $f_\pi(\prec,\dots,\prec)=\tau$ by definition, as desired.

We prove the above core claim to complete the proof. Suppose that (X,τ') is an NSE but $\tau'\neq\tau$. We first show that,

$$U^{\mathrm{a}}(X,t)=U^{\mathrm{a}}(X,\tau')\quad\text{for all } t\in T.$$

Indeed, we have $U^{\mathrm{a}}(X,t)\le U^{\mathrm{a}}(X,\tau')$ for all $t\in T$ since $\tau'\in\mathrm{BR}(X)$ by the definition of the NSE. To see that $U^{\mathrm{a}}(X,t)\ge U^{\mathrm{a}}(X,\tau')$ for all t, suppose for the sake of contradiction that $U^{\mathrm{a}}(X,h)<U^{\mathrm{a}}(X,\tau')$ for some $h\in T$. Let $\mathbf{c}^i=\mathrm{cov}(\mathbf{x}^i)$ be the contribution of each defender i to the overall coverage. Now that $\mathrm{cov}_h(X)=-U^{\mathrm{a}}(X,h)>-U^{\mathrm{a}}(X,\tau')=\mathrm{cov}_{\tau'}(X)\ge 0$, there exists $i\in[n]$, such that $c_h^i>0$. We modify $\mathbf{c}^i$ slightly; let $\tilde{\mathbf{c}}^i$ be such that $\tilde{c}_t^i=\min\{1,c_t^i+\delta_t\}$ for all $t\neq h$, and $\tilde{c}_h^i=c_h^i-\delta_h$. When the changes δ_t's are sufficiently small and $\sum_{t\neq h}\delta_t\le\delta_h$, we have $\tilde{\mathbf{c}}^i\in C_{k_i}$, so $\tilde{\mathbf{c}}^i$ corresponds to some feasible strategy $\tilde{\mathbf{x}}^i$ of defender i. Suppose defender i now deviates to playing $\tilde{\mathbf{x}}^i$; let $\tilde{X}=\langle\mathbf{x}^{-i},\tilde{\mathbf{x}}^i\rangle$. We can further let $\delta_{\tau'}$ be sufficiently smaller than all other δ_t's, so that after the change we have $U^{\mathrm{a}}(\tilde{X},t)<U^{\mathrm{a}}(\tilde{X},\tau')$ for all $t\in T\setminus\{\tau'\}$ (recall that we had $U^{\mathrm{a}}(X,t)\le U^{\mathrm{a}}(X,\tau')$ and $U^{\mathrm{a}}(X,t)$ changes continuously with $\mathbf{c}^i$). Thus, τ' becomes the only best response of the attacker while its coverage increases; $\mathrm{BR}(\tilde{X})=\{\tau'\}$ and $\mathrm{cov}_{\tau'}(\tilde{X})>\mathrm{cov}_{\tau'}(X)$. For defender i, it follows that $\min_{t\in\mathrm{BR}(\tilde{X})}U_i^{\mathrm{d}}(\tilde{X},t)=U_i^{\mathrm{d}}(\tilde{X},\tau')>U_i^{\mathrm{d}}(X,\tau')$, contradicting the assumption that (X,τ') is an NSE.

It follows that in this game $\mathrm{cov}_t(X)=-U^{\mathrm{a}}(X,t)=-U^{\mathrm{a}}(X,\tau')=\mathrm{cov}_{\tau'}(X)$ for all t. We should also have $\mathrm{cov}_{\tau'}(X)>0$, since otherwise no target is protected at all and defender 1 can allocate some resource to improve her utility. Moreover, $\mathrm{cov}_{\tau'}(X)<1$ since otherwise $\sum_{t\in T}\mathrm{cov}_t(X)=|T|\ge 3$, which is a contradiction because

$$\sum_{t\in T}\mathrm{cov}_t(X)=\sum_{t\in T}\left(1-\prod_{i\in[n]}\left(1-\mathrm{cov}_t(\mathbf{x}^i)\right)\right)\le\sum_{t\in T,\,i\in[n]}\mathrm{cov}_t(\mathbf{x}^i)\le K=2. \tag{3}$$

Therefore, when X is played, all targets are equally appealing to the attacker. Some defender i with $\mathrm{cov}_\tau(\mathbf{x}^i)>0$ can then reduce it to zero (by removing the resource to be allocated to τ in every pure strategy). Since $\mathrm{cov}_\tau(X)=\mathrm{cov}_{\tau'}(X)<1$, this will reduce $\mathrm{cov}_\tau(X)$ (see (2)) and result in τ to be the attacker's only best response. Since $\tau\prec_i\tau'$, defender i is better off making this deviation, which contradicts the assumption that (X,τ') forms an NSE. □

Lemma 3.5. *If π is strategyproof then f_π is strategyproof.*

Proof. Suppose that f_π is not strategyproof. Then there exist linear orders $\prec_1,\dots,\prec_n$ and $\prec_i'$ over T and alternatives $\tau,\tau'\in T$ such that $f(\prec_1,\dots,\prec_i,\dots,\prec_n)=\tau$, $f(\prec_1,\dots,\prec_i',\dots,\prec_n)=\tau'$, and $\tau\prec_i\tau'$, i.e., voter i benefits from reporting $\prec_i'$ instead of $\prec_i$.

Let $\boldsymbol{\theta}=g(\prec_1,\dots,\prec_i,\dots,\prec_n)$ and $\boldsymbol{\theta}'=g(\prec_1,\dots,\prec_i',\dots,\prec_n)$. Hence, $\theta_j=\theta_j'$ for all $j\in[n]\setminus\{i\}$. By the definition of f_π, we have $\pi(\boldsymbol{\theta},\mathbf{k},\alpha^*)=(\mathbf{x},\tau)$ and $\pi(\boldsymbol{\theta}',\mathbf{k},\alpha^*)=(\mathbf{x}',\tau')$ for some $\mathbf{x}$ and $\mathbf{x}'$. Consider a defender i of type θ_i. Since $\tau\prec_i\tau'$, we have

$$U_i^{\mathrm{d}}(\pi(\boldsymbol{\theta},\mathbf{k},\alpha^*))=U_i^{\mathrm{d}}(\mathbf{x},\tau)<U_i^{\mathrm{d}}(\mathbf{x}',\tau')=U_i^{\mathrm{d}}(\pi(\boldsymbol{\theta}',\mathbf{k},\alpha^*)).$$

Thus, π is not strategyproof. □

We are now ready to prove Theorem 3.3.

Proof of Theorem 3.3. Let π be an arbitrary strategyproof mechanism. We show that π cannot be weakly UP. By Lemmas 3.4 and 3.5, the social choice function f_π is onto T and strategyproof. Since $|T|\ge 3$, by the Gibbard–Satterthwaite theorem f_π is a dictatorship. Assume that voter λ is the dictator; we can assume without loss of generality that $\lambda\neq 1$ (if $\lambda=1$, we can exchange voters 1 and 2 below). Let τ be the favourite alternative of voter λ, i.e., $\tau\succ_\lambda t$ for all $t\in T\setminus\{\tau\}$; then $f_\pi(\prec_1,\dots,\prec_n)=\tau$.

Now, consider linear orders $\prec_1',\dots,\prec_n'$, in which

$$t_1\prec_\lambda' t_2\prec_\lambda'\cdots\prec_\lambda' t_m,\quad\text{and}\quad t_m\prec_1' t_{m-1}\prec_1'\cdots\prec_1' t_1.$$

Then $f_\pi(\prec_1',\dots,\prec_n')=t_m$ and by the definition of f_π this implies $\pi(g(\prec_1',\dots,\prec_n'),\mathbf{k},\alpha^*)=(\mathbf{x},t_m)$ for some $\mathbf{x}$. Let $\boldsymbol{\theta}=g(\prec_1',\dots,\prec_n')$. Thus, in the game given by defender profile $(\boldsymbol{\theta},\mathbf{k})$ and attacker type α^*, the mechanism offers defender 1 utility

$$U_1^{\mathrm{d}}(\pi(\boldsymbol{\theta},\mathbf{k},\alpha^*))=U_1^{\mathrm{d}}(\mathbf{x},t_m)\le r_1^{\mathrm{d}}(t_m)=\frac{1}{2}.$$

However, we will show next that defender 1 can obtain utility at least 1 even in the worst NSE, so π cannot be weakly UP.

Observe that defender 1 can prevent t_m (her least preferred attacker response) from being adopted by the attacker by always allocating her resource to t_m: the resulting strategy profile $\tilde{X}$ gives $\mathrm{cov}_{t_m}(\tilde{X})=1$ and $\mathrm{cov}_t(\tilde{X})<1$ for at least one target $t\in T$ (for the same reason as (3)). Thus, $U^{\mathrm{a}}(\tilde{X},t_m)>U^{\mathrm{a}}(\tilde{X},t)$, so $t_m\notin\mathrm{BR}(\tilde{X})$. Since $t_m\prec_1' t$ for all $t\in T\setminus\{t_m\}$, we have

$$\min_{t\in\mathrm{BR}(\tilde{X})}U_1^{\mathrm{d}}(\tilde{X},t)\ge U_1^{\mathrm{d}}(\tilde{X},t_{m-1})\ge p_1^{\mathrm{d}}(t_{m-1})=1.$$

Therefore, a strategy profile cannot be an NSE if it does not offer defender 1 utility 1, as otherwise the defender can always allocate her resource to t_m to increase her utility. The worst NSE should offer defender 1 utility 1, which completes the proof. □

4 CONSTRAINED SOCIAL WELFARE MAXIMIZATION MECHANISMS

Since our desirable properties cannot be satisfied simultaneously, we can ask what is a maximal subset of them that can be satisfied at the same time. SP alone is easy to achieve, by appointing one defender (e.g., one who contributes the most resources) as a dictator and maximizing her utility, which also gives us PE if we fix a tie-breaking order. Nevertheless, since some defenders may end up even less happy than in the worst NSE, this defeats the purpose of coordination. In this section, we propose a Constrained Social Welfare Maximization mechanism (CSW-Max) that achieves PE and weak UP (but not SP), and runs in polynomial time. For the special case where all defenders have non-overlapping payoffs we show that when an NSE is provided, a variant of CSW-Max mechanism

can achieve PE, Weak UP, and SP simultaneously; this mechanism, too, runs in polynomial time.

4.1 A PE and Weakly UP Mechanism

CSW-Max works as follows.

1. Compute an NSE (X, τ) using the algorithm by Gan et al. [10] (which runs in polynomial time).
2. Solve the following linear program LP-t for each $t \in T$, where $\mathbf{c}$ are the variables.

$$\text{LP-}t\text{:}\quad \text{maximize} \quad \sum_{i\in[n]} U_i^{\mathrm{d}}(\mathbf{c}, t) \tag{4}$$

$$\text{subject to} \quad U_i^{\mathrm{d}}(\mathbf{c}, t) \geq U_i^{\mathrm{d}}(X, \tau) \qquad \forall i \in [n] \tag{4a}$$

$$U^{\mathrm{a}}(\mathbf{c}, j) \leq U^{\mathrm{a}}(\mathbf{c}, t) \qquad \forall j \in T \tag{4b}$$

$$\sum_{j\in T} c_j \leq K \tag{4c}$$

$$0 \leq c_j \leq 1 \qquad \forall j \in T \tag{4d}$$

3. Pick a $t \in T$ that maximizes the optimal value of LP-t; let it be t^* (breaking ties lexicographically). Output $(\mathbf{c}^*, t^*)$.

In LP-t, Constraint (4a) requires each defender to obtain at least their utility in the NSE (X, τ); (4b) requires that the attacker is incentivized to attack target t; the remaining two constraints ensure that $\mathbf{c} \in C_K$ is a feasible strategy. We will now argue that LP-t is feasible for at least one $t \in T$, so the mechanism will always output some outcome. Moreover, this outcome is PE and weakly UP.

Lemma 4.1. *LP-t is feasible for at least one $t \in T$.*

Proof. In fact, LP-τ is always feasible as $\mathbf{z} = \mathrm{cov}(X)$ satisfies all the constraints. Specifically, (4a) is satisfied trivially when $\mathbf{z} = \mathrm{cov}(X)$. (4b) is satisfied because (X, τ) is an NSE, so by definition $\tau \in \mathrm{BR}(X) = \mathrm{BR}(\mathbf{z})$. Let $X = (\mathbf{x}^1, \ldots, \mathbf{x}^n)$. We have

$$z_j = 1 - \prod_{i=1}^{n}(1 - \mathrm{cov}_j(\mathbf{x}^i)) \leq \sum_{i=1}^{n} \mathrm{cov}_j(\mathbf{x}^i)$$

for each $j \in T$, and summing up the inequalities gives

$$\sum_{j\in T} z_j \leq \sum_{j\in T, i\in[n]} \mathrm{cov}_j(\mathbf{x}^i) \leq \sum_{i\in[n]} k_i = K,$$

so Constraint (4c) is satisfied. Finally, Constraint (4d) is satisfied as $0 \leq \mathrm{cov}_j(X) \leq 1$ for all $j \in T$ by definition. □

Theorem 4.2. CSW-Max *is PE and weakly UP.*

Proof. Let $(\mathbf{c}^*, t^*)$ be the outcome of CSW-max. It is weakly UP since by Constraint (4a) every defender gets at least their utility in the NSE (X, τ).

Suppose for the sake of contradiction that $(\mathbf{c}^*, t^*)$ is not PE, and it is Pareto dominated by another outcome $(\mathbf{z}, h)$. We have $\sum_{i\in[n]} U_i^{\mathrm{d}}(\mathbf{z}, h) > \sum_{i\in[n]} U_i^{\mathrm{d}}(\mathbf{c}^*, t^*)$. Since $(\mathbf{c}^*, t^*)$ maximizes the optimal values of LP-t's that admit feasible solutions, $\mathbf{z}$ is not a feasible solution of LP-h. However, since $(\mathbf{z}, h)$ is a feasible outcome, it satisfies (4b)–(4d), so the only constraint it can violate is (4a); we have $U_j^{\mathrm{d}}(\mathbf{z}, h) < U_j^{\mathrm{d}}(X, \tau)$ for some $j \in [n]$. Further, since $(\mathbf{c}^*, t^*)$ is a feasible solution of LP-t^*, we have $U_i^{\mathrm{d}}(\mathbf{c}^*, t^*) \geq U_i^{\mathrm{d}}(X, \tau)$ for all $i \in [n]$. Thus, $U_j^{\mathrm{d}}(\mathbf{z}, h) < U_j^{\mathrm{d}}(X, \tau) \leq U_j^{\mathrm{d}}(\mathbf{c}^*, t^*)$, which contradicts the assumption that $(\mathbf{z}, h)$ Pareto dominates $(\mathbf{c}^*, t^*)$. □

4.2 NSE-induced SP Mechanism

Sometimes a (truthful) NSE is known before the defenders decide to collaborate, e.g., when the status quo is an NSE (indeed, since an NSE in our model can be computed efficiently, it is plausible to assume that one may arise from interaction between the players). If in this NSE every target is protected with some probability but not probability 1 and every defender has non-overlapping payoffs, we can achieve SP, PE, and weak UP simultaneously. Intuitively, the known NSE spares us the trouble of dealing with non-truthful input information when we compute the NSE in Step 1 of CSW-Max. Given also the defenders' consistent preferences with non-overlapping payoffs, we are able to generate a PE and weakly UP outcome without any additional information.

Let (X, τ) be a known NSE such that $0 < \mathrm{cov}_t(X) < 1$ for all $t \in T$. Our mechanism CSW-Max-NSE is a simple variant of CSW-Max:

1. Solve LP-τ defined in (4). Let the optimal solution be $\mathbf{c}^*$.
2. Output $(\mathbf{c}^*, \tau)$.

The argument in Lemma 4.1 shows that LP-τ is always feasible. We will show that CSW-Max-NSE is SP, PE, and weakly UP.

Theorem 4.3. CSW-Max-NSE *is SP, PE, and weakly UP when the defenders have non-overlapping payoffs.*

Proof. Observe that $U_i^{\mathrm{d}}(\mathbf{c}, \tau)$ depends only on c_τ and is increasing with respect to c_τ. Thus, LP-τ is equivalent to the following LP with variables $\mathbf{c}$:

$$\begin{aligned} &\text{maximize} && c_\tau \\ &\text{subject to} && c_\tau \geq \mathrm{cov}_\tau(X) && \forall i \in [n] \\ &&& \text{Constraints (4b)–(4d)} \end{aligned}$$

The formulation does not rely on the defenders' payoffs. Thus, misreporting does not change the output and CSW-Max-NSE is SP.

To show that it is PE, suppose for the sake of contradiction that $(\mathbf{c}^*, \tau)$ is Pareto dominated by another outcome $(\mathbf{z}, h)$.

If $h = \tau$, we have $U_i^{\mathrm{d}}(\mathbf{z}, \tau) \geq U_i^{\mathrm{d}}(\mathbf{c}^*, \tau)$ for all $i \in [n]$, so $\mathbf{z}$ is a feasible solution to LP-τ; and $\sum_{i\in[n]} U_i^{\mathrm{d}}(\mathbf{z}, \tau) > \sum_{i\in[n]} U_i^{\mathrm{d}}(\mathbf{c}^*, \tau)$, so $\mathbf{z}$ is actually a better solution than $\mathbf{c}^*$, contradicting the fact that $\mathbf{c}^*$ is an optimal solution of LP-τ.

Therefore, $h \neq \tau$. Since $\tau \in \mathrm{BR}(X)$, we have $U^{\mathrm{a}}(X, t) \leq U^{\mathrm{a}}(X, \tau)$ for all $t \in T$. We can further argue that $U^{\mathrm{a}}(X, t) = U^{\mathrm{a}}(X, \tau)$, so $\mathrm{BR}(X) = T$. Indeed, if $U^{\mathrm{a}}(X, t') < U^{\mathrm{a}}(X, \tau)$, target t' is overly protected; some defender i can rebalance the coverage, adjusting her contribution $\mathbf{c}^i = \mathrm{cov}(\mathbf{x}^i)$ to $\tilde{\mathbf{c}}^i$ such that $\tilde{c}^i_{t'} = c^i_{t'} - \delta_{t'}$ and $\tilde{c}^i_t = c^i_t + \delta_t$ for all $t \neq t'$. When the changes δ_t's are sufficiently small, $\tilde{\mathbf{c}}^i \in C_{k_i}$, so it is feasible, and by letting δ_τ be sufficiently smaller than all other δ_t's, we can make τ the unique attacker best response after the deviation, in which case the utility of defender i will increase, contradicting the assumption that (X, τ) is an NSE.

Now that $\mathrm{BR}(X) = T$ and the payoffs are non-overlapping, it must be that $h \prec_i \tau$ for those defenders i with $\mathrm{cov}_h(\mathbf{x}^i) > 0$. Indeed, if $h \succ_i \tau$ these defenders would be better off reducing $\mathrm{cov}_h(\mathbf{x}^i)$ (and hence $\mathrm{cov}_h(X)$) to attract the attacker to attack h. This implies

$U_i^{\mathrm{d}}(\mathbf{z}, h) < U_i^{\mathrm{d}}(\mathbf{c}^*, \tau)$—a contradiction to the assumption that $(\mathbf{z}, h)$ Pareto dominates $(\mathbf{c}^*, \tau)$.

Finally, weak UP is guaranteed by Constraint (4a) of LP-τ. □

5 STABILITY AGAINST STRATEGY MODIFICATION

So far, we assumed that the mechanism was in charge of the defenders' resources. In some scenarios the mechanism does not have direct access to the resources; it can only *recommend* allocation strategies to the defenders and cannot enforce that the defenders will indeed follow the recommendations. In this section, we consider an additional property: *individual rationality* (IR). This property requires that no defender has an incentive to deviate from the strategy dictated by the coordination mechanism.

5.1 Recommendation Strategies and IR

In allocation strategies considered so far, resources are treated as anonymous; we do not care about the ownership of resources in the allocation. We will now extend the definition of allocation strategies to incorporate the ownership information, and call the newly defined strategies *recommendation strategies*. In this new model, a pure strategy is a joint strategy $S = (s^1, \ldots, s^n)$ with each $s^i \in \mathcal{T}_{k_i}$ being a pure allocation strategy of defender i. We have $\mathrm{cov}_t(S) = 1$ if and only if $\mathrm{cov}_t(s^1) = 1$ for some $i \in [n]$. A recommendation strategy $\mathbf{x}$ is then a distribution over joint pure strategies, i.e., $\mathbf{x} \in \Delta(\mathcal{T}_{\mathbf{k}})$ where we write $\mathcal{T}_{\mathbf{k}} = \prod_{i\in[n]} \mathcal{T}_{k_i}$. A mechanism generates an outcome $(\mathbf{x}, t)$ conditioned on the input player profile as previously defined, and to implement the outcome, the mechanism samples a joint pure strategy $(s^1, \ldots, s^n)$ from $\mathbf{x}$ and recommends each defender i to take strategy s^i. The recommendation is privately sent to each defender, and only the distribution $\mathbf{x}$ is publicly known to everyone. We define *strategy modification* to formalize how a defender could deviate from the recommended strategy.

Definition 5.1 (Strategy modification). *A strategy modification of a defender $i \in [n]$ is a function $\phi : \mathcal{T}_{k_i} \to \Delta(\mathcal{T}_{k_i})$, under which the defender plays a mixed strategy $\phi(s)$ whenever she is instructed to play a pure strategy $s \in \mathcal{T}_{k_i}$ by the coordination mechanism.*

For a distribution $\mathbf{x} \in \Delta(\mathcal{T}_{\mathbf{k}})$, we abuse the notation and let $\phi(\mathbf{x})$ denote the distribution that arises when defender i adopts ϕ. By $\mathbf{x}' = \phi(\mathbf{x})$, the probability that each joint pure strategy $S = (s^1, \ldots, s^n) \in \mathcal{T}_{\mathbf{k}}$ is chosen is

$$x'_S = \sum_{Q=(q^1,\ldots,q^n)\in\mathcal{T}_{\mathbf{k}}} x_Q \cdot \phi_{s^i}(q^i),$$

where $\phi_{s^i}(q^i)$ is the probability s^i is chosen by the mixed strategy $\phi(q^i)$. We are now ready to define individual rationality.

Property 4 (Individual rationality (IR)). An outcome $(\mathbf{x}, t)$ is *individually rational* if for every defender $i \in [n]$ it holds that $U_i^{\mathrm{d}}(\mathbf{x}, t) \geq \min_{t' \in \mathrm{BR}(\phi(\mathbf{x}))} U_i^{\mathrm{d}}(\phi(\mathbf{x}), t')$ for every strategy modification $\phi : \mathcal{T}_{k_i} \to \Delta(\mathcal{T}_{k_i})$. A coordination mechanism is IR if only outputs IR outcomes.

Put differently, IR requires correlated-equilibrium-like outcomes and every NSE is an IR outcome. Given our results in the previous sections, we would like to find a PE and weakly UP mechanism that is also IR. Unfortunately, such mechanism does not exist.

5.2 PE, Weak UP, and IR are Incompatible

We describe a game in which no outcome is PE, weakly UP, and IR.

Example 5.2. *There are three defenders and five targets $T = \{a, b, c, d, e\}$. Each defender has one resource. The defenders have non-overlapping payoffs captured by the following preference orders:*

$$a \prec_1 b \prec_1 c \prec_1 d \prec_1 e;$$
$$c \prec_2 b \prec_2 a \prec_2 d \prec_2 e;$$
$$e \prec_3 d \prec_3 c \prec_3 b \prec_3 a.$$

The attacker's payoffs are as follows.

	a	b	c	d	e
r^{a}	1	1	1	0.7	0.7
p^{a}	0	0	0	0	0

Lemma 5.3. *In Example 5.2 the attacker's best response is to attack target b in every NSE.*

Proof. First, since a, c, and e are the least favored attacker responses of some defender, the attacker's response cannot be any of these targets in an NSE: otherwise the defender for whom this target is the lest favored attacker response can deviate to always allocating her resource to this target; the target will have coverage 1 and cannot be an attacker best response in this example.

We will now argue that target d cannot be the attacker's best response either. We show that defender 3 can always avoid targets d and e to be the attacker's best responses by playing the strategy $\mathbf{c}^3 = (0, 0, 0, \frac{1}{2}, \frac{1}{2})$, irrespective of the other defenders' strategies. When $\mathbf{c}^3$ is played, the coverage of both targets d and e will be at least 0.5. The attacker obtains utility at most $0.7 \times 0.5 + 0 \times 0.5 = 0.35$ by attacking any of them. Thus, for either of them to be a best response of the attacker, each of a, b, and c needs to receive coverage at least 0.65 (for the attacker to get utility at least 0.35 by attacking them); this is impossible. Essentially, let $\mathbf{c}^1$ and $\mathbf{c}^2$ be the first two defenders' strategy, so each target $t \in \{a, b, c\}$ has coverage $1-(1-c_t^1)(1-c_t^2)$; we show that the following constraints cannot be satisfied simultaneously to complete the proof.

$$\begin{cases} 1-(1-c_t^1)(1-c_t^2) \geq 0.65, & \text{for all } t \in \{a, b, c\} \\ 0 \leq c_t^i \leq 1, & \text{for all } i = 1, 2, \text{ and } t \in \{a, b, c\} \\ c_a^i + c_b^i + c_c^i \leq k_i = 1, & \text{for all } i = 1, 2 \end{cases}$$

Indeed, suppose that the above constraints are satisfied by some c_t^i's. For simplicity, let $(x_1, x_2, x_3) = (1-c_a^1, 1-c_b^1, 1-c_c^1)$, $(y_1, y_2, y_3) = (1-c_a^2, 1-c_b^2, 1-c_c^2)$. Then the following constraints should be satisfied.

$$\begin{cases} x_j \cdot y_j \leq 0.35, & \text{for all } j = 1, 2, 3 \\ 0 \leq x_j \leq 1, \text{ and } 0 \leq y_j \leq 1, & \text{for all } j = 1, 2, 3 \\ x_1 + x_2 + x_3 \geq 2, \text{ and } y_1 + y_2 + y_3 \geq 2 & \end{cases}$$

Assume without loss of generality that $x_1 \leq x_2 \leq x_3$. In fact, if the inequalities are satisfiable, there is a satisfying solution such that $y_1 \geq y_2 \geq y_3$ because if, e.g., $y_1 \leq y_2$, we can exchange the values of y_1 and y_2 while the inequalities will still be satisfied. Further, assume without loss of generality that $x_1 \cdot y_3 \leq x_3 \cdot y_1$. Let $(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) =$

$(x_1-\delta,\ x_2,\ x_3+\delta)$ and $(\tilde{y}_1,\tilde{y}_2,\tilde{y}_3)=(y_1+\frac{\delta\cdot y_3}{x_3+\delta},\ y_2,\ y_3-\frac{\delta\cdot y_3}{x_3+\delta})$. We have

$$\tilde{x}_1+\tilde{x}_2+\tilde{x}_3=x_1+x_2+x_3\geq 2,$$
$$\tilde{y}_1+\tilde{y}_2+\tilde{y}_3=y_1+y_2+y_3\geq 2,$$
$$\tilde{x}_3\cdot\tilde{y}_3=(x_3+\delta)\cdot(y_3-\tfrac{\delta}{x_3+\delta}\cdot y_3)=x_3\cdot y_3\leq 0.35,$$
$$\tilde{x}_2\cdot\tilde{y}_2=x_2\cdot y_2\leq 0.35,$$
$$\tilde{x}_1\cdot\tilde{y}_1=(x_1-\delta)\cdot(y_1+\tfrac{\delta}{x_3+\delta}\cdot y_3)$$
$$=x_1y_1+\tfrac{\delta(x_1\cdot y_3-y_1\cdot x_3-y_1\cdot\delta-y_3\cdot\delta)}{x_3+\delta}\leq x_1y_1\leq 0.35.$$

However, if we gradually increase δ, $0\leq x_j\leq 1$ and $0\leq y_j\leq 1$ will be violated. Consider the first constraint that becomes tight.

- Suppose that $\tilde{x}_1$ drops to 0 first. We have $\tilde{x}_1=0$ while all the constraints are satisfied. By $\tilde{x}_1+\tilde{x}_2+\tilde{x}_3\geq 2$ and $\tilde{x}_2,\tilde{x}_3\in[0,1]$ we have $\tilde{x}_2=\tilde{x}_3=1$. By $\tilde{x}_j\cdot\tilde{y}_j\leq 0.35$ we have $\tilde{y}_2\leq 0.35$ and $\tilde{y}_3\leq 0.35$. It follows that, $\tilde{y}_1+\tilde{y}_2+\tilde{y}_3\leq 1+0.35+0.35<2$, which is a contradiction.
- Suppose that $\tilde{x}_3$ reaches 1 first. We have $\tilde{x}_3=1$ while all the constraints are satisfied. It follows by $\tilde{x}_3\cdot\tilde{y}_3\leq 0.35$ that $\tilde{y}_3\leq 0.35$ and, by $\tilde{y}_1+\tilde{y}_2+\tilde{y}_3=2$ and $\tilde{y}_1\geq\tilde{y}_2$, that $\tilde{y}_1\geq 0.825$. Applying these constraints repeatedly, we have
$$\tilde{x}_1\leq\tfrac{0.35}{\tilde{y}_1}<0.425\ \Rightarrow\ \tilde{x}_2\geq 2-\tilde{x}_1-\tilde{x}_3\geq 0.575$$
$$\Rightarrow\tilde{y}_2\leq\tfrac{0.35}{\tilde{x}_2}<0.61\ \Rightarrow\ \tilde{y}_1=2-\tilde{y}_2-\tilde{y}_3>1,$$
which is a contradiction.
- When $\tilde{y}_1$ reaches 1 first or $\tilde{y}_3$ drops to 0 first, the same analysis can be applied. Both cases lead to contradictions as above.

Thus, no x_j's and y_j's can satisfy the constraints. □

Lemma 5.4. *Suppose $(\mathbf{x},\tau)$ is a PE outcome in Example 5.2, and $\mathbf{c}=\mathrm{cov}(\mathbf{x})$. Then $c_a=c_b=c_c=\frac{24}{31}$ and $c_d=c_e=\frac{21}{31}$.*

Proof. We first show that $\mathbf{c}$ indeed corresponds to a feasible outcome. Note that in this example, $K=3=\sum_{t\in T}c_t$ by the above valuation of $\mathbf{c}$, so $\mathbf{c}\in C_K$ and hence it can be implemented by a distribution $\mathbf{x}\in\Delta(\mathcal{T}_K)$ (see Section 2). Trivially, any coverage vector that can be implemented by some distribution $\mathbf{x}\in\Delta(\mathcal{T}_K)$ can also be implemented by a distribution in $\mathbf{x}\in\Delta(\mathcal{T}_{\mathbf{k}})$ over the non-anonymous joint pure strategy $\mathcal{T}_{\mathbf{k}}$, so $\mathbf{c}$ is feasible.

Suppose for the sake of contradiction that $(\mathbf{x}',\tau')$ is a PE outcome, but $\mathbf{c}'=\mathrm{cov}(\mathbf{x}')\neq\mathbf{c}$. We have $\sum_{t\in T}c'_t\leq K=\sum_{t\in T}c_t$, which implies that $c'_h<c_h$ for some $h\in T$. Further, observe that $U^{\mathrm{a}}(\mathbf{c},t)=\frac{7}{31}$ for all $t\in T$, so $\mathrm{BR}(\mathbf{c})=T$. Since $U^{\mathrm{a}}(\mathbf{c},t)$ decreases with c_t, we have $\max_{t\in T}U^{\mathrm{a}}(\mathbf{c}',t)\geq U^{\mathrm{a}}(\mathbf{c}',h)>U^{\mathrm{a}}(\mathbf{c},h)=\frac{7}{31}$. By definition $\tau'\in\mathrm{BR}(\mathbf{c}')$, so $U^{\mathrm{a}}(\mathbf{c}',\tau')>\frac{7}{31}=U^{\mathrm{a}}(\mathbf{c},\tau')$. On the other hand, the utility $U_i^{\mathrm{d}}(\mathbf{c},t)$ of each defender i increase with c_t, so $U^{\mathrm{a}}(\mathbf{c}',\tau')>U^{\mathrm{a}}(\mathbf{c},\tau')$ implies $U_i^{\mathrm{d}}(\mathbf{c}',\tau')<U_i^{\mathrm{d}}(\mathbf{c},\tau')$ for all $i\in[n]$, which means that now there is another outcome $(\mathbf{x},\tau')$ that Pareto dominates $(\mathbf{x}',\tau')$, contradicting the assumption that $(\mathbf{x}',\tau')$ is PE. (Here $(\mathbf{x},\tau')$ is a feasible outcome since $\tau'\in\mathrm{BR}(\mathbf{c})=T$.) □

Theorem 5.5. *There exists no mechanism that is PE, weakly UP, and IR, even when the defenders' payoffs are non-overlapping.*

Proof. It suffices to show that Example 5.2 does not admit a PE, weakly UP, and IR outcome.

First, by Lemma 5.3 and the fact that the defenders have non-overlapping utilities, if an outcome $(\mathbf{x},\tau)$ is weakly UP then we must have $\tau=b$. By Lemma 5.4, $\mathrm{BR}(\mathbf{c})=T$. This implies that no defender i will receive a recommendation to protect any target $t\succ_i b$ with positive probability, i.e., for all $S=(s^1,s^2,s^3)\in\mathcal{T}_{\mathbf{k}}$ such that $x_S>0$, $\mathrm{cov}_t(s^i)=0$ if $t\succ_i b$ (otherwise, the defender can remove the resource allocated to this target to induce a better attacker response). Hence, targets d and e will only be protected by defender 3 who has only one resource. We have $\mathrm{cov}_e(S)+\mathrm{cov}_d(S)\leq 1$ in every pure joint strategy in the support set of $\mathbf{x}$; it follows that $\mathrm{cov}_e(\mathbf{x})+\mathrm{cov}_d(\mathbf{x})=\sum_{S\in\mathcal{T}_{\mathbf{k}}:x_S>0}x_S\cdot(\mathrm{cov}_e(S)+\mathrm{cov}_d(S))\leq\sum_{S\in\mathcal{T}_{\mathbf{k}}}x_S=1$. This contradicts Lemma 5.4 where we have $c_d+c_e=\frac{42}{31}>1$. □

We are again forced to select a strict subset of the conflicting properties. Now if we drop PE, essentially we will be looking at mechanisms that generate correlated equilibria. Trivially, every NSE, as a special correlated equilibrium, is weakly UP and IR, and we know that they always exist and can be computed efficiently. However, in this way we do not obtain any of the benefits of coordination, violating our original motivation.

Alternatively, we can aim to find correlated equilibria that satisfy certain optimality criteria among all correlated equilibria, e.g., equilibria that maximize the social welfare, or Pareto optimal equilibria. We leave this question open for future work and only highlight the challenges here. Unlike in normal-form games or many other succinctly representable multiplayer games where correlated equilibria can be computed efficiently [17], our model features a very different utility structure, where a player's utility from playing a mixed strategy is not a linear combination of their pure strategy utilities because of different attacker responses induced by these strategies. This means that we are unable to use existing approaches in order to compute correlated equilibria in our setting.

6 CONCLUSION

This paper demonstrates an intriguing phenomenon in security scenarios involving multiple defenders: even when the defenders face a common enemy, it is often not possible for them to reach a consensus on how to coordinate. Our analysis highlights the underlying causes of this phenomenon: there is a competition among the defenders, where each defender drives the attacker towards a target that she considers less important. Our impossibility results immediately call for more innovative ways to promote coordination. In scenarios where strategyproofness is not necessary, one open question outlined above is how to compute a correlated equilibrium with good social welfare properties. While we view the problem from the perspective of mechanism design, in reality there are other natural means to be considered as well, such as negotiation, contracting, and coalition formation. These will be interesting directions to explore in order to obtain more positive results.

ACKNOWLEDGMENTS

This work was supported by the European Research Council (ERC) under grant number 639945 (ACCORD). Jiarui Gan was supported by the EPSRC International Doctoral Scholars Grant EP/N509711/1. Sarit Kraus was partially supported by Ministry of Science and Technology, Israel and the Japan Science and Technology Agency (JST), Japan.

REFERENCES

[1] Stéphane Airiau, Haris Aziz, Ioannis Caragiannis, Justin Kruger, Jérôme Lang, and Dominik Peters. 2019. Portioning Using Ordinal Preferences: Fairness and Efficiency. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19)*. 11–17.

[2] Bo An and Milind Tambe. 2017. *Stackelberg Security Games (SSG) Basics and Application Overview*. Cambridge University Press, 485–507.

[3] Haris Aziz, Anna Bogomolnaia, and Hervé Moulin. 2019. Fair mixing: the case of dichotomous preferences. In *Proceedings of the 2019 ACM Conference on Economics and Computation*. ACM, 753–781.

[4] Nicola Basilico, Andrea Celli, Giuseppe De Nittis, and Nicola Gatti. 2017. Computing the team-maxmin equilibrium in single-team single-adversary team games. *Intelligenza Artificiale* 11, 1 (2017), 67–79.

[5] Anna Bogomolnaia, Hervé Moulin, and Richard Stong. 2005. Collective choice under dichotomous preferences. *Journal of Economic Theory* 122, 2 (2005), 165–184.

[6] Florian Brandl, Felix Brandt, Dominik Peters, Christian Stricker, and Warut Suksompong. 2020. Funding Public Projects: A Case for the Nash Product Rule. (2020). Working paper.

[7] Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. 2019. Be a Leader or Become a Follower: The Strategy to Commit to with Multiple Leaders. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19)*. 123–129.

[8] Francesco Maria Delle Fave, Albert Xin Jiang, Zhengyu Yin, Chao Zhang, Milind Tambe, Sarit Kraus, and John P Sullivan. 2014. Game-theoretic patrolling with dynamic execution uncertainty and a case study on a real transit system. *Journal of Artificial Intelligence Research* 50 (2014), 321–367.

[9] Conal Duddy. 2015. Fair sharing under dichotomous preferences. *Mathematical Social Sciences* 73 (2015), 1–5.

[10] Jiarui Gan, Edith Elkind, and Michael Wooldridge. 2018. Stackelberg Security Games with Multiple Uncoordinated Defenders. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS'18)*. 703–711.

[11] Allan Gibbard. 1973. Manipulation of voting schemes: a general result. *Econometrica* 41, 4 (1973), 587–601.

[12] Albert Xin Jiang, Ariel D Procaccia, Yundi Qian, Nisarg Shah, and Milind Tambe. 2013. Defender (mis) coordination in security games. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI'13)*. 220–226.

[13] Aron Laszka, Jian Lou, and Yevgeniy Vorobeychik. 2016. Multi-defender strategic filtering against spear-phishing attacks. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'16)*. 537–543.

[14] Jian Lou, Andrew M Smith, and Yevgeniy Vorobeychik. 2017. Multidefender Security Games. *IEEE Intelligent Systems* 32, 1 (2017), 50–60.

[15] Jian Lou and Yevgeniy Vorobeychik. 2015. Equilibrium analysis of multi-defender security games. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI'15)*. 596–602.

[16] Jian Lou and Yevgeniy Vorobeychik. 2016. Decentralization and security in dynamic traffic light control. In *Proceedings of the Symposium and Bootcamp on the Science of Security*. 90–92.

[17] Christos H Papadimitriou and Tim Roughgarden. 2008. Computing correlated equilibria in multi-player games. *Journal of the ACM (JACM)* 55, 3 (2008), 14.

[18] Praveen Paruchuri, Jonathan P Pearce, Janusz Marecki, Milind Tambe, Fernando Ordonez, and Sarit Kraus. 2008. Efficient Algorithms to Solve Bayesian Stackelberg Games for Security Applications.. In *AAAI*. 1559–1562.

[19] James Pita, Manish Jain, Milind Tambe, Fernando Ordóñez, and Sarit Kraus. 2010. Robust solutions to Stackelberg games: Addressing bounded rationality and limited observations in human cognition. *Artificial Intelligence* 174, 15 (2010), 1142–1171.

[20] Mark Allen Satterthwaite. 1975. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of economic theory* 10, 2 (1975), 187–217.

[21] Andrew Smith, Yevgeniy Vorobeychik, and Joshua Letchford. 2014. Multi-Defender security games on networks. *ACM SIGMETRICS Performance Evaluation Review* 41, 4 (2014), 4–7.

[22] Milind Tambe. 2011. *Security and game theory: algorithms, deployed systems, lessons learned*. Cambridge University Press.