From inference to influence: applying causal game theory to complex security environments

Vonk, M.C.

Chapter 6

Real-World Applications

The preceding chapters introduced key concepts from causality and game theory, alongside a methodology for optimizing causal interventions in hybrid Bayesian networks.

This chapter applies these theoretical foundations to the context of complex security environments through two specific case studies. First, causal techniques are applied to uncover a causal structure and estimate causal effects related to environmental conflict in Iraq in Section 6.1. Second, causal game-theoretic concepts are utilized to simulate the effectiveness of countermeasures against hybrid threats in Section 6.2. Environmental conflict and hybrid threats are examined as illustrative examples of complex security environments, as they involve interdependent challenges, diverse threat actors, and evolving risks that contribute to unpredictable and dynamic security conditions.

In doing so, this chapter addresses RQ3: *How can the proposed (strategic) causal concepts be applied to complex security environments?* The content closely reflects two peer-reviewed journal articles [250, 154], to which the reader is directed for further detail.

6.1 Environmental Conflict

Despite significant advances in conflict research methods [188, 238, 132, 44], the outbreak of armed conflict remains difficult to predict, largely due to unresolved questions about its underlying causes. While environmental security research has explored specific causal pathways linking environmental factors to conflict [66, 103, 106, 20, 199,

209], most studies rely on observational data and confirm only limited mechanisms. However, the randomized controlled trials discussed in Chapter 3 are neither practical nor ethical in the context of armed conflict. Therefore, more comprehensive causal explanations remain elusive, leaving a critical methodological gap that hinders both scholarly understanding and effective policy interventions.

In addressing this gap, this section applies the concepts introduced in Chapter 3 to infer causality from non-experimental observations of armed conflict. Using causal discovery and inference methods, commonly hypothesized causal pathways are tested. The considered cross-section consists of 294 non-experimental observations, one for each subdistrict in Iraq (Arabic: nawāḥī) as the unit of analysis. The outcome variable is the count of conflict events. Each observation is additionally described in terms of explanatory variables, including demographics, vital resources, environment, and weather. These observations were sampled from several geocoded maps [188, 208, 164, 159, 50, 114, 1]. From these observations, an empirical causal mechanism of linkages between environment and conflict was retrieved. Represented as a causal graph, the mechanism shows causal pathways from environmental variables to armed conflict outcomes. The mechanism is characterized by causal effects of these variables on the count of conflict events, accounting for causal spillover wherever these effects could be identified and estimated.

Sections 6.1.1 and 6.1.2 present the hypotheses and detail the data and methodological approach. Section 6.1.3 reports the empirical findings, while Section 6.1.4 concludes with a discussion of the implications and potential directions for future research.

6.1.1 Hypotheses

This section derives the causal hypotheses discussed below from relevant findings in the literature. Adapted from Sakaguchi, Varughese, and Auld [209], Figure 6.1 outlines a hypothesized mechanism through which environmental factors may contribute to conflict, as extensively discussed in environmental security literature. The hypothetical linkages are rooted in environmental causes. The figure respectively distinguishes between direct and indirect linkages (i.e., paths A and B, respectively). The latter are indirect because environmental scarcity hypothetically plays a mediating role between environmental causes and armed conflict outcomes.

Longer-term weather patterns have been argued to cause armed conflict directly [112, 111, 225, 81, 21]. The direct link between long-term weather patterns and armed

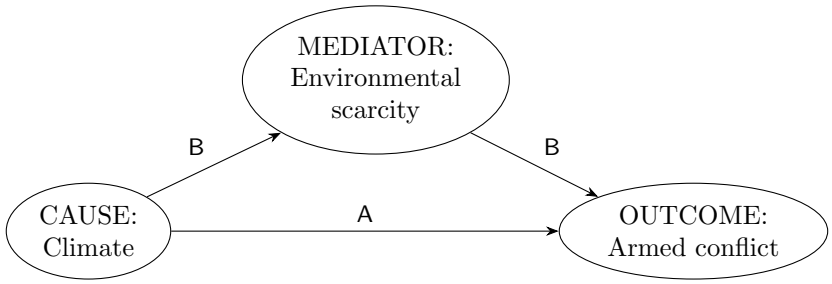

Figure 6.1: Hypothesized causal pathways originating from linkages between environment and conflict.

conflict can be seen in how populations respond to environmental changes. For instance, environmental disruptions affecting livelihoods can prompt community mobilization or lead armed groups to intervene, anticipating adverse outcomes. These actions may lead to direct causal effects of environmental factors to armed conflict. Such effects can originate from soil moisture, temperature, or simply the fact that different physical surroundings absorb or release accumulated heat differently (i.e., also referred to as latent heat or energy) [111, 209]. For example, droughts in East Africa have increased cattle raiding and inter-communal violence among pastoralist groups [72]. Therefore, it is hypothesized that changes in soil moisture, temperature, and latent energy directly cause changes in armed conflict activity (H1).

- H1 a): An increase in latent energy in the form of heat directly causes an increase in armed conflict activity.
- H1 b): An increase in skin temperature directly causes an increase in armed conflict activity.
- H1 c): An increase in soil moisture directly causes a decrease in armed conflict activity.

Further, environmental processes have also been argued to cause armed conflict indirectly [16, 209, 131]. Causal mediation of environmental effects on armed conflict concerns primarily scarcity of vital resources, also referred to as environmental scarcity [103, 104, 105, 129]. Environmental scarcity has been argued to mediate the environmental effects on armed conflict [209, 105]. For instance, land scarcity in Chiapas, Mexico has been hypothesized to mediate environmental pressures into insurgency and civil violence [105]. While causal mediation is elaborated in more detail below,

it can already be hypothesized that the effects of environmental processes indirectly cause armed conflict.

- H2 a): An increase in latent energy indirectly causes an increase in armed conflict activity.
- H2 b): An increase in skin temperature indirectly causes an increase in armed conflict activity.
- H2 c): An increase in soil moisture indirectly causes a decrease in armed conflict activity.

To refine the understanding of indirect causal mechanisms, causal mediation analysis can incorporate specific conditions through which environmental effects manifest. Agricultural and pastoral systems have been shown to shape societal responses to long-term climatic variability, including migration [259]. Land degradation, desertification, and water scarcity are identified as mediators of environmental impacts on violent conflict [98, 72, 116]. Wheat production, a key agricultural output in Iraq [45], has been found to mediate temperature effects on violence [145].

Demographic factors, such as population size, growth, density, and migration, have also been linked to conflict [241, 187, 192]. Resource scarcity is argued to affect denser populations more acutely, increasing the likelihood of conflict under environmental stress [16, 2]. These mediating pathways are captured in the following hypotheses.

- H3 a): Given the indirect paths from the environmental processes to armed conflict activity, wheat production causally mediates the indirect effects of envi-

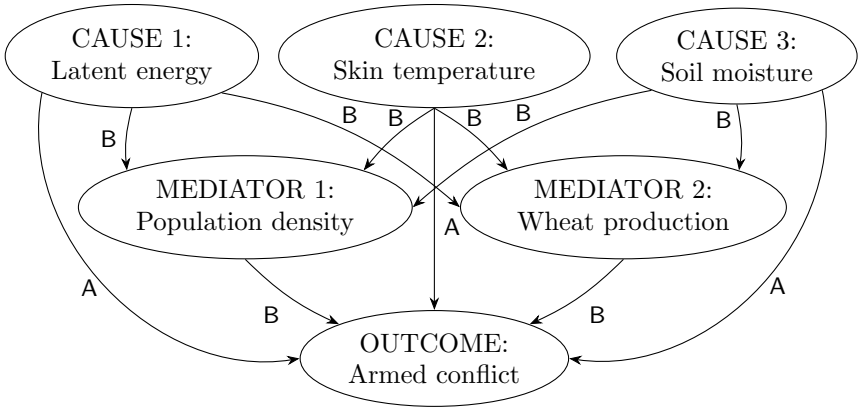

Figure 6.2: The hypothetical causal structure of all hypotheses combined.

ronmental processes on armed conflict activity, by causing an additional decrease in armed conflict activity.

- H3 b): Given the indirect paths from the environmental processes to armed conflict activity, population density causally mediates the indirect effects of environmental processes on armed conflict activity, by causing an additional increase in armed conflict activity.

With these hypotheses, it is possible to compose the entire hypothetical causal structure of linkages between environment and conflict, as shown in Figure 6.2. The figure again distinguishes between direct and indirect linkages, corresponding to edges A and B, respectively. As can be seen in the figure, grounded in environmental processes, the scarcity of vital resources exposes the population to existential stress. Both the density of population and the scarcity of agricultural resources aggravate these effects.

6.1.2 Data and Methods

Data

The units of analysis are all Iraq's 294 subdistricts from January 1, 2020, to January 1, 2022, aggregated as a single cross-section. Data included conflict events, such as violence against civilians and battles from ACLED [188], where total conflict events—aggregated across battles, explosions, violence against civilians, protests, riots, and strategic developments—served as the outcome variable. Environmental and demographic explanatory variables were sourced as geo-coded grids from the Humanitarian Data Exchange, ECMWF's ERA5-Land [164], NASA [114, 208, 159], CIESIN at Columbia University [50], and MapSPAM [232]. Skin temperature, representing the interface temperature between the earth's surface and atmosphere, was selected due to its relevance for agriculture and water availability [122]. Soil moisture at 28–100 cm depth, which impacts crop viability and water access, was also included. Latent energy, measuring heat flux linked to evaporation and condensation, was used to capture broader hydrometeorological dynamics [208]. As a proxy for environmental scarcity, wheat production data were retrieved from MapSPAM, given wheat's importance to Iraqi agriculture [232, 145]. Population density, previously linked to conflict vulnerability, was sourced from CIESIN's Gridded Population of the World dataset [50]. Due to data scarcity in Iraq, other societal and political variables commonly used in conflict research were unavailable at an appropriate resolution [261, 3].

Methods

By applying causal methodology to non-experimental observations, causal paths and effects behind the causal mechanism of such linkages can be disentangled and quantified. As the methodological concepts are introduced in Chapter 3, the discussion here is limited to the specification and implementation of the introduced methods.

Since there are a limited number of observations, GES was selected as the causal discovery algorithm as it has been found appropriate in simulation studies involving small sample sizes [155]. The considered loss function was the Bayesian information criterion. The output of GES was the likeliest DAG, given the observations. The nodes of the DAG correspond to the armed conflict activity and explanatory variables. The edges correspond to respective causal relationships between them [155]. Causal estimands of the explanatory variables were identified using the backdoor criterion and adjustment formula. Acknowledging that the observations in question are spatially correlated in the sense that the climatological variables of one municipality may have an influence on the climate-conflict dynamics of another municipality, the estimation procedure invokes SESEM to account for spatial confounding [137].

A straightforward way to make this spatial confoundedness explicit is to use distances between municipalities. Distances are modeled by computing municipal centroids and the shortest paths connecting them. The first step of applying SESEM is fitting an initial non-spatial SEM to the data [136], assuming independence of errors. Then, spatially explicit variance–covariance matrices are computed for a series of lag distances divided into bins corresponding to distance ranges. Each bin contains 500 data pairs to ensure sufficient sample size for meaningful inference. Inference focuses on the lowest 20% of distances, as spillover effects are more likely in nearby municipalities. Finally, SEM models are fitted for each chosen lag distance, and edge coefficients, standard errors, and p-values are computed. In addition, parameters are defined for individual causal paths such that path-specific coefficients, standard errors, and p-values can be computed.

6.1.3 Results

Empirical causal structure

Section 6.1.1 outlined the hypothetical causal structure linking environmental factors to conflict. Figure 6.3 presents the empirically derived causal structure based on the available non-experimental observations.

Albeit somewhat less expressive, the empirical causal structure largely corresponds with the hypothetical one in Figure 6.2. The conflict nodes cluster together. The only node with only incoming edges is total conflict events. Further, the structure is rooted in environmental processes. Apart from the direct causal path from the temperature node to battle events, all the other paths from the environmental processes to conflict events are indirect.

Because the population density, temperature and wheat production nodes have the highest number of incoming and outgoing edges, these nodes are pivotal to the connectedness of the empirical causal structure. This lends credence to the causal mediation of environmental processes on the conflict outcomes. In fact, without population density and wheat production, the environmental causes would be largely disconnected from the outcomes. Rather than treating this evidence as conclusive, the empirical structure is further used to conduct hypothesis testing.

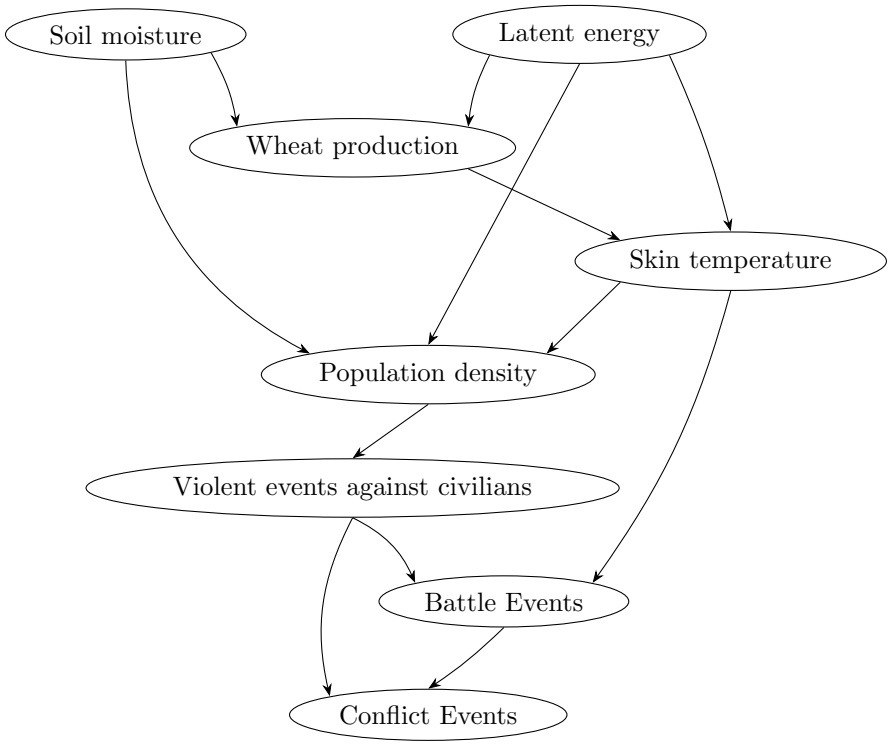

Figure 6.3: The empirical causal structure retrieved from the GES algorithm.

Causal hypotheses

The empirical causal structure can assist with the validation of the causal hypotheses of naturally caused armed conflict. With this structure, a spatially explicit structural equation model was fit. Continuous explanatory variables were z-score standardized, while count variables were left unstandardized to maintain policy-relevance. Causal effects of the explanatory variables on the count of conflict events were estimated.

The SESEM model showcased acceptable performance, with the comparative fit index (ranging 0-1) exceeding 0.9 and the standardized root mean square residual falling below 0.08 for almost all spatial distances [211]. These values are indicators of an acceptable model fit, underscoring the model's effectiveness in capturing the underlying data structure. While the root mean square error of approximation values are occasionally above 0.10, suggesting room for improvement, this is likely attributable to the limited number of observations available.

For each causal estimate, a hypothesis test was conducted to discern whether to attribute the estimate to a random error or not. Table 6.1 lists the causal estimates, their standard errors, and statistical significances.

The first set of hypotheses states that effects of environmental processes directly cause armed conflict outcomes. The changes in latent energy (H1 a), skin temperature (H1 b), and soil moisture (H1 c) were hypothesized to cause a change in the count of conflict events directly. The only environmental process with a direct causal path to conflict events is skin temperature. This path is mediated via battle events and the estimates, standard errors, and statistical significances can be estimated for each of the isolated paths. The results are in Table 6.2. Given the causal structure, the estimated causal effect of skin temperature on total conflict events mediated by battle events is positive for all spatial distances and 47.35 for the non-spillover fitted model, with at least a 0.1% statistical significance level. Since there is no evidence that warrants rejecting the null hypotheses relative to H1 a) and H1 c), hypothesis H1 b) is accepted and hypotheses H1 a) and H1 c) are not accepted.

Further, the second set of hypotheses states that causal effects of the environmental processes on conflict events are mediated. Latent heat was hypothesized to increase the count of total conflicts indirectly (H2 a). Since all effects from latent heat to conflict events are mediated, and the estimated effect of latent heat on the conflict events is positive and 0.1% statistically significant for all spatial distances, hypothesis H2 a) is accepted. Furthermore, soil moisture (H2 c) was hypothesized to cause a decrease in the count of conflict events indirectly. As there are no direct paths from

Table 6.1: Standardized causal estimates across distance bins with standard errors in brackets. All results are statistically significant at the 0.1% level.

Distance	Wheat Production	Latent Heat	Soil Moisture	Skin Temperature	Population Density
0 km	14.02 (3.9)	24.27 (6.2)	-18.30 (4.6)	51.67 (9.5)	29.66 (5.8)
3–27 km	23.48 (4.1)	33.92 (5.4)	-20.72 (4.2)	64.16 (7.6)	22.61 (4.3)
27–40 km	10.69 (2.6)	21.83 (4.3)	-17.93 (3.1)	48.17 (7.4)	29.42 (4.5)
40–50 km	15.62 (3.4)	26.64 (5.2)	-19.95 (3.8)	56.04 (8.1)	26.65 (4.5)
50–59 km	12.38 (2.7)	23.22 (4.4)	-18.93 (3.4)	45.23 (6.5)	34.42 (4.7)
59–66 km	7.59 (2.2)	15.05 (3.5)	-10.41 (2.4)	36.99 (7.0)	16.52 (3.3)
66–73 km	13.99 (3.1)	28.43 (5.4)	-16.99 (3.8)	58.17 (7.8)	18.17 (4.1)
73–80 km	14.62 (3.7)	28.87 (5.5)	-19.53 (4.0)	71.40 (9.0)	32.45 (5.9)
80–86 km	11.18 (2.6)	15.61 (3.6)	-14.71 (2.9)	41.21 (7.0)	26.29 (4.0)
86–92 km	14.45 (2.7)	20.63 (4.1)	-16.43 (3.1)	47.49 (6.3)	23.16 (3.8)
92–99 km	12.00 (2.5)	21.88 (4.3)	-19.68 (3.5)	41.15 (6.0)	26.86 (4.0)
99–104 km	18.30 (3.7)	26.26 (5.2)	-25.15 (4.3)	63.40 (8.3)	43.84 (5.6)
104–110 km	23.17 (4.6)	37.04 (6.6)	-24.85 (4.8)	81.29 (9.4)	27.69 (5.1)
110–116 km	7.70 (2.0)	13.16 (3.1)	-14.46 (2.6)	30.19 (5.9)	30.58 (3.7)
116–121 km	11.67 (2.7)	23.80 (4.8)	-14.44 (3.0)	45.79 (7.4)	29.06 (4.3)
121–126 km	6.92 (1.7)	13.87 (3.0)	-16.06 (2.7)	33.26 (5.4)	27.72 (3.4)
126–132 km	16.08 (3.0)	23.86 (4.5)	-16.46 (3.2)	50.07 (6.8)	21.62 (3.5)
132–137 km	8.66 (2.3)	15.39 (4.0)	-13.12 (2.8)	33.65 (6.8)	18.19 (3.4)
137–142 km	11.02 (2.5)	16.71 (3.8)	-14.94 (2.9)	41.28 (7.0)	22.61 (3.6)
142–147 km	15.89 (3.2)	20.92 (4.6)	-17.00 (3.6)	54.80 (7.8)	31.78 (4.6)
147–152 km	9.75 (2.2)	18.46 (4.0)	-11.33 (2.5)	30.66 (5.7)	15.23 (3.5)
152–158 km	11.11 (2.5)	17.22 (4.0)	-14.67 (3.0)	34.63 (6.3)	20.02 (3.3)
158–163 km	14.61 (3.0)	27.24 (5.2)	-13.79 (3.3)	56.28 (7.8)	25.70 (4.7)
163–168 km	14.06 (2.8)	22.30 (4.6)	-18.70 (3.6)	43.42 (6.9)	29.26 (4.0)
168–173 km	11.76 (2.7)	18.74 (4.5)	-15.62 (3.3)	32.77 (6.6)	28.39 (4.2)
173–178 km	13.12 (3.0)	23.36 (4.8)	-15.54 (3.3)	54.18 (7.6)	24.45 (4.3)
178–183 km	8.16 (2.4)	12.68 (3.4)	-12.00 (2.6)	41.87 (8.0)	24.57 (4.2)
183–188 km	17.17 (3.4)	22.95 (5.0)	-20.85 (4.1)	48.57 (7.6)	39.79 (5.4)
188–193 km	14.64 (2.9)	25.01 (5.1)	-20.75 (4.1)	50.37 (6.9)	19.87 (3.7)
193–198 km	15.00 (3.2)	23.31 (4.9)	-17.27 (3.6)	48.33 (7.7)	24.65 (4.8)

soil moisture to conflict events, their causal effects on conflict events can only be indirect. Given the indirect paths from soil moisture to armed conflict activity, ceteris paribus, a one unit increase in soil moisture causes a decrease in the counted conflict events at all spatial bins, including a -18.30 decrease for the non-spillover fitted model. Therefore, hypothesis (H2 c) is also accepted. Isolating the indirect paths from the skin temperature (H2 b) to armed conflict activity via population density, ceteris paribus,

Table 6.2: Causal effects of skin temperature on conflict across distance bins. Values are standardized estimates with standard errors in brackets. Asterisks denote statistical significance at 5% (*), 1% (**), and 0.1% (***) levels.

Distance	Temperature Population Civilians Battles Conflicts	Temperature Battles Conflicts	Temperature Population Civilians Conflicts
0 km	1.39 (0.83)	47.35*** (9.36)	2.93* (1.25)
3–27 km	0.13 (0.17)	63.34*** (7.50)	0.69 (0.72)
27–40 km	1.29* (0.65)	44.88** (7.31)	2.00** (0.82)
40–50 km	0.68 (0.52)	52.46*** (8.03)	2.91** (1.02)
50–59 km	0.52 (0.42)	41.22*** (6.28)	3.49** (1.28)
59–66 km	0.19 (0.26)	35.14*** (6.93)	1.66* (0.70)
66–73 km	0.87 (0.54)	55.63*** (7.76)	1.66*** (0.54)
73–80 km	0.58 (0.52)	68.62*** (8.88)	2.20* (1.04)
80–86 km	0.30 (0.48)	37.21*** (6.74)	3.70*** (1.18)
86–92 km	0.36 (0.27)	45.67*** (6.25)	1.46 (0.75)
92–99 km	2.05** (0.74)	36.79*** (5.86)	2.30*** (0.64)
99–104 km	1.31 (0.71)	58.73*** (8.06)	3.36* (1.41)
104–110 km	0.58 (0.56)	77.91*** (9.29)	2.79** (1.06)
110–116 km	2.15** (0.76)	24.85*** (5.65)	3.19*** (0.97)
116–121 km	0.95 (0.52)	42.72*** (7.28)	2.12* (0.98)
121–126 km	1.73* (0.70)	28.14*** (5.34)	3.39*** (1.01)
126–132 km	1.13* (0.51)	42.72*** (7.28)	1.82** (0.69)
132–137 km	0.82* (0.39)	30.58*** (6.75)	2.25*** (0.76)
137–142 km	1.53* (0.61)	37.31*** (6.87)	2.45*** (0.74)
142–147 km	1.58 (0.72)	49.34*** (7.62)	3.88*** (1.10)
147–152 km	0.35 (0.24)	29.54*** (5.69)	0.76 (0.49)
152–158 km	0.97* (0.45)	31.38*** (6.18)	2.28*** (0.71)
158–163 km	0.63 (0.51)	53.10*** (7.72)	2.55*** (0.90)
163–168 km	1.04 (0.56)	39.63*** (6.68)	3.04** (1.14)
168–173 km	0.71 (0.48)	30.67*** (6.47)	1.39 (0.83)
173–178 km	1.96* (0.87)	49.37*** (7.56)	2.85*** (0.78)
178–183 km	1.14 (0.65)	38.80*** (7.92)	1.93* (0.82)
183–188 km	1.84* (0.84)	43.89*** (7.36)	2.84** (1.08)
188–193 km	0.52 (0.50)	47.06*** (6.84)	2.79*** (0.85)
193–198 km	1.79* (0.90)	44.92*** (7.56)	2.12** (0.74)

computing causal effects did not yield significant results for all the spatial distances as can be observed in Table 6.2. Therefore, hypothesis H2 b) is not accepted.

Finally, the third set of hypotheses states that causal effects of environmental conditions on conflict events are agriculturally and demographically mediated. Whereas wheat production was hypothesized to cause a decrease in the count of conflict events

(H3 a), population density was hypothesized to cause an increase (H3 b). Given the indirect paths from soil moisture and latent energy to armed conflict activity, ceteris paribus, a one unit increase in wheat production causes an increase in counted conflict events for all spatial distances, including a 14.02 unit increase at 0.1% statistical significance level for the non-spillover model. Given the indirect paths from soil moisture and latent energy to armed conflict activity, ceteris paribus, a one unit increase in population density causes a 29.66 increase in the number of counted conflict events at 0.1% statistical significance level for the non-spatial model and similar significant positive estimates for other spatial distances (see Table 6.1). This evidence leads to the rejection of the null hypotheses relative to the H3 a) and H3 b) hypotheses, whereas hypothesis H3 a) is rejected and H3 b) is accepted at all spatial bins.

6.1.4 Conclusion and Future Work

Armed conflict research is advanced by applying causal methodology to non-experimental data, enabling the retrieval of empirical causal structures, identification of causal paths, and estimation of effects. The research confirms that environmental processes, particularly soil moisture and latent energy, affect armed conflict through demographic and agricultural mediators, addressing key gaps in environmental security literature. Finally, the findings support the design of targeted policy interventions by identifying causal mechanisms and mediators amenable to strategic action. In the context of environmental security, this enables preventive strategies that interrupt causal paths from environmental stressors to conflict outcomes, such as reducing population density through social or migration policies, or mitigating environmental pressures via investments in hydrological infrastructure.

Future research should assess the robustness of the causal findings by relaxing the causal sufficiency assumption and testing for sensitivity to unobserved confounders, particularly social and political variables such as power-sharing, intergroup animosities, and horizontal inequality [261, 3, 188]. This can be achieved using existing causal frameworks designed to account for unobserved confounding [33, 35, 140].

Additionally, the current analysis assumes linear structural relations in the SEM, justified by the need for interpretability, estimation stability, and suitability as a first-order approximation in data-constrained settings. Future research should extend this by applying more flexible, non-linear SEMs to test the robustness of findings and uncover potentially richer causal dynamics.

6.2 Hybrid Threats

Hybrid threats, defined as the coordinated use of violent and non-violent means to exploit vulnerabilities and influence adversaries below the threshold of armed conflict, pose an escalating challenge in an era of growing global interconnectedness. In response, states have implemented a broad spectrum of potential counter-hybrid measures, including economic sanctions, cyber defense strategies, information campaigns, and diplomatic initiatives. However, the effectiveness of these measures remains uncertain due to the complex and opaque nature of hybrid threats, which often operate across multiple domains and take place below the threshold of detection and attribution [125]. Therefore, researchers have resorted to modeling approaches.

While attempting to model hybrid threat dynamics, some authors have turned to game theory to examine strategic interactions among rival states in an effort to overcome the paucity of information available [25, 18]. Others have incorporated scarce data sources into Bayesian modeling techniques [60, 24] with the aim of refining domain knowledge with available data. Although current game-theoretic approaches struggle to capture the complexities and uncertainties inherent in hybrid threat dynamics, Bayesian modeling techniques, while effective at handling uncertainty, fall short in representing strategic interactions.

This section proposes an integrated probabilistic and game-theoretic framework to assess counter-hybrid threat measures under conditions of deep uncertainty. Drawing on influence diagrams to model uncertainties in threat detection, attribution, and mitigation as probabilistic relations [108], the approach is extended to a multi-agent setting [128], enabling the inclusion of strategic interactions. The model captures both cognitive and psychological aspects of deterrence through probabilistic reasoning [30], and strategic decision-making via game-theoretic structures. The interaction between two state-like agents is modeled such that the defender's pay-off reflects the trade-off between the costs of countermeasures and the potential damage from hybrid attacks, whether successfully deterred or not. Optimal countermeasures are identified by maximizing expected pay-offs, while strategic equilibria are derived from the adversary's strategic responses.

In order to test the modeling approach, a cyber threat scenario on critical infrastructure was developed, inspired by real-world incidents. Policy experts and available literature were consulted to identify relevant countermeasures to this cyber threat and collate estimates of the cost, damage mitigation ability, and deterrence ability of each of the counter-hybrid measures. These estimates provided a basis for analyzing counter-

hybrid measures and allowed for gauging their effectiveness across different scenarios, including scenarios where the adversary engages in strategic competition. To validate the proposed approach, the findings were contextualized within the framework of existing studies, and sensitivity analyses were conducted to identify and quantify the most influential variables driving the model's outcomes.

While Section 6.2.1 outlines the methodology for the models considered and elicitation of expert knowledge, Section 6.2.2 highlights the experimental design of the considered hybrid threat scenario. The results are introduced in Section 6.2.3, and some concluding remarks are given in Section 6.2.4.

6.2.1 Methodology

This section outlines the underlying mechanism of the simulation model, along with the process of eliciting inputs required to run simulations using the model. The full scope of the proposed method, including the extent of expert involvement, is illustrated in Figure 6.4.

The model considers the behavior of two agents possessing the characteristics of sovereign states, following the two-agent approach of Balcaen et al. [25] and Attiah et al. [18]. On one side, agent A aims to pursue its strategic objectives using hybrid attacks. On the other side, agent B wishes to protect its national interests and deter

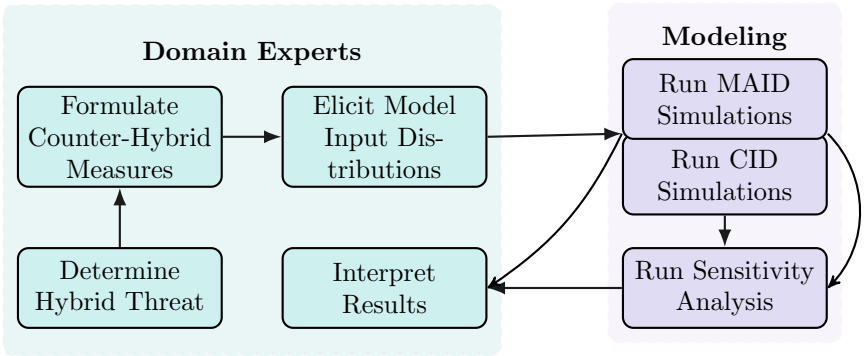

Figure 6.4: The figure illustrates the processes involved in counter-hybrid threat analysis. Initially, domain experts identify the hybrid threat and develop corresponding counter-hybrid measures. They also provide key input parameters, which are used to construct probabilistic input distributions for the model. Samples from these distributions are used to run simulations with the causal influence diagram (CID) as well as the multi-agent influence diagram (MAID) model. Finally, a sensitivity analysis is performed and the model results are interpreted and compared with existing studies.

and defend against hybrid attacks. Agent B, the defender, chooses a counter-hybrid posture to deter or dissuade agent A from carrying out a hybrid operation. For this reason, the strategy is referred to as a counter-hybrid measure. To this purpose, agent B explores available counter-hybrid measures to dissuade the adversary from carrying out hybrid attacks by altering the cost-benefit calculus [38]. The defender may also adopt measures - such as the enhancement of detection and/or attribution capabilities [212] - that would boost resilience and mitigate the potential impact of hybrid conducts [82].

Both the counter-hybrid measure and the hybrid attack bear direct costs. Such costs represent not only the resource costs but also costs involving, for instance, political capital to rally domestic and international support as well as potential costs associated with escalation. The interaction between agents A and B is of a zero-sum nature. Probabilities are used to reflect the considerable degree of uncertainty over the value of key variables that lead to different outcomes. Examples are uncertainty associated with the impact of counter-hybrid measures on the strategic calculus of agent A, as well as with detection, attribution, and the mitigatory impact of the counter-hybrid measure.

First, a causal influence diagram approach is introduced, enabling the optimization of counter-hybrid deterrence strategies when the adversary's responsiveness is estimated probabilistically. This approach is then extended into a multi-agent influ-

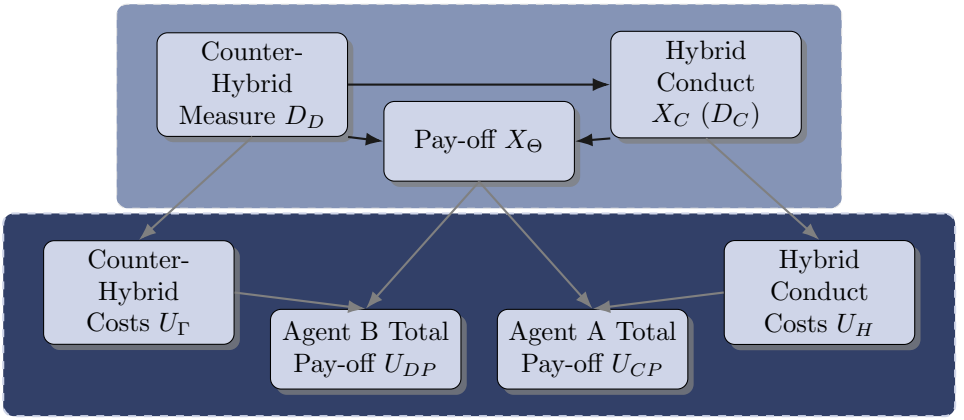

Figure 6.5: (Multi-agent) influence diagram encoding hybrid threat modeling. While the bottom background layer groups the deterministic variables, the top layer represents the probabilistic variables. Probabilistic relations are displayed by black arrows and deterministic relations by grey arrows.

ence diagram by modeling both agents as players within a game-theoretical framework, allowing for the analysis of game equilibria.

Optimizing counter-hybrid strategy

A Bayesian network modeling technique is often used to account for the combination of probabilistic and deterministic relationships [24, 268, 260, 123, 256]. The truncated factorization of Chapter 3 is applied to factorize the joint probability distribution efficiently. Furthermore, the Bayesian networks will be extended to causal influence diagrams as in Chapter 4 to further dissect the nodes of the graph into random variables, utility nodes, and decision nodes and allow for causal interventions. In the first proposed modeling approach, the defender preemptively commits to a selected counter-hybrid posture to deter or dissuade agent A from conducting a hybrid operation, represented by a decision node. The adversary's response is driven by the estimated probability of successful dissuasion.

More formally, let D_D be the decision variable describing the counter-hybrid measure, which is followed by X_C, the random variable for the hybrid conduct. The available strategies and state spaces for D_D and X_C are $\Omega_{D_D} = \{d_1, \ldots, d_n\}$ and $\Omega_{X_C} = \{c_1, \ldots, c_m\}$ respectively. Each combination of counter-hybrid measure and hybrid operation variables leads to a discrete pay-off random variable X_Θ with state space $\Omega_{X_\Theta} = \{\theta_1, \ldots, \theta_k\}$ representing the interaction pay-off structure excluding direct costs. Because the pay-off structure from hybrid and counter-hybrid interaction are separated from the direct cost of hybrid operation and counter-hybrid measure, it is assumed that the pay-off structure has a zero-sum game component, meaning a positive value $\theta \in \Omega_{X_\Theta}$ corresponds to the gain of agent B and the loss of agent A, while a negative value for $\theta \in \Omega_{X_\Theta}$ represents a gain for agent A and the loss for agent B. The direct costs for the counter-hybrid measure and the hybrid attack are drawn from a cost probability function and denoted by U_Γ and U_H respectively, where the costs for the counter-hybrid measure has state space $\Omega_{U_\Gamma} = \{\gamma_1, \ldots, \gamma_n\}$ with $\gamma_i \geq 0$ for $i = 1, \ldots, n$ and the costs for the hybrid attack has state space $\Omega_{U_H} = \{\eta_1, \ldots, \eta_m\}$ with $\eta_i \geq 0$ for $i = 1, \ldots, m$. The total pay-off can finally be calculated by $U_{DP} = X_\Theta - U_\Gamma$ for agent B and $U_{CP} = -X_\Theta - U_H$ for agent A. The causal influence diagram that corresponds with these relations is depicted in Figure 6.5.[1]

[1]While the cost variables are priorly drawn from probability distributions, they become deterministic in the influence diagram, as only one cost value corresponds to a (counter-) hybrid measure per experiment.

Note that factorizing the direct costs in the initial pay-off node could have made the influence diagram purely probabilistic. However, for clarification purposes, direct costs have been separated from interaction costs. All probability distributions in the influence diagram are typically assumed to be categorical, which makes the conditional probabilities expressable via conditional probability tables.

Suppose that agent B has access to the potential costs of counter-hybrid measures $\Omega_{U_\Gamma} = \{\gamma_1, \ldots, \gamma_n\}$, their deterrence capacity $P(c \mid d)$ and their ability to mitigate potential damages $P(\theta \mid d, c)$. Agent B's objective is to compute the counter-hybrid measure that maximizes its total pay-off under the assumed probability distributions [9]. Specifically, the goal is to find the intervention $do(D = d)$ that maximizes agent B's expected total pay-off. Since the counter-hybrid measure node has no incoming arrows (see Figure 6.5), intervening on a variable is the same as conditioning on this variable [178]:

$$\begin{aligned}
&\max_{d_1,\ldots,d_n} \mathbb{E}[dp \mid do(D_D = d)] \\
&= \max_{d_1,\ldots,d_n} \sum_c \sum_\theta \sum_\gamma dp P(c, \theta, \gamma, dp, \mid do(D_D = d)) \\
&= \max_{d_1,\ldots,d_n} \sum_c \sum_\theta \sum_\gamma dp P(c \mid d) P(\theta \mid d, c) P(\gamma \mid d) P(dp \mid \gamma, \theta).
\end{aligned}$$

This can be formulated as an integer linear program (ILP):

$$\begin{aligned}
&\max \sum_{h=1}^{k} \sum_{j=1}^{m} \sum_{i=1}^{n} \theta_h w_{ijh} q_{ij} p_i + \sum_{i=1}^{n} \gamma_i p_i \\
\text{subject to} \quad &\sum_{i=1}^{n} p_i = 1 \\
&p_i \in \{0, 1\} \quad i = 1, \ldots, n
\end{aligned}$$

Subgame perfect equilibrium

In optimizing the counter-hybrid strategy, probabilities are used to estimate the likelihood of successfully deterring the adversary after each measure. These probabilities are determined ex-ante, meaning they are drawn before a counter-hybrid measure is chosen. As a result, they do not account for any short-term pay-off adjustments that occur during the interaction that eventually set the outcome in equilibrium. To this end, the causal influence diagram approach is extended to multi-agent influence diagrams [128, 90] to account for these strategic considerations, and the notion of

equilibrium is addressed using the causal games that emerge from such diagrams.

Formally, the hybrid conduct node X_C of Figure 6.5 is no longer modeled probabilistically but is instead represented as a decision node D_C. Additionally, the distinction between the two agents, A and B, is made explicit by assigning decision and utility nodes to each respective player. Specifically, decision node D_D and utility nodes U_{DP} and U_Γ are assigned to agent B, while decision node D_C and utility nodes U_{CP} and U_H are associated with agent A.

A solution concept is necessary that pinpoints a subset of possible outcomes when agents act rationally. As explained in Chapter 4, the subgame perfect equilibrium is a natural solution concept in the causal games that emerge from MAIDS, which helps eliminate non-credible threats. In the context of the hybrid threat game, non-credible threat equilibria emerge when the attacker threatens to conduct a hybrid operation despite it not being in their best interest in terms of pay-off.

Probability distributions and elicitation

Filling the influence diagram with accurate conditional probabilities is widely recognized as a challenging task [123] and a rigorous elicitation process should be developed to ensure the highest degree of accuracy in the inputs. To maintain a realistic perspective in the estimates, the cost, potential deterrence capacity, (either denial or punishment), and resilience-enhancing ability of each counter-hybrid measure are estimated on the basis of an in-depth literature review complemented with a mini-Delphi approach with seven (junior) analysts with a background in strategic studies. This resulted in probability distributions from which samples were drawn to conduct the experiments. Initially, the parameters of these distributions were inspired by a literature review. Subsequently, analysts made a one-time adjustment to the parameters, informed by visualizations of the resulting distributions. The specifics of these probability distributions per variable are summarized in Table 6.3 while the exact parameters are available in Appendix C.1. Values that are likely to be drawn from these probability distributions indicate that they align closely with consensus in the literature and the outcomes of the mini-Delphi survey, while values unlikely to be drawn correspond to values that are less in alignment. By repeatedly sampling input variables from these distributions independently, a total of 1000 experimental scenarios were generated. These experiments can be considered semi-synthetic due to the absence of a rigorous, standardized method for constructing the prior distributions for these estimates, as described in Section 4.2.4, requiring reliance on the constructed probabilistic representations.

Table 6.3: Values drawn from probability distributions. Costs of counter-hybrid measures and damaging impacts of hybrid attacks are expressed in US million dollars. While the costs of counter-hybrid measures and damaging impacts of hybrid attacks are drawn from variants of the normal distribution, the probability values for the ability to deter and the ability to mitigate damaging impacts are drawn from their corresponding conjugate priors (Beta and Dirichlet, respectively).

Value	Meaning	Probability Distribution
θ_h	Potential damage of a hybrid attack	Drawn from different truncated normal distributions for each category of damaging impacts [170]
γ_i	Cost for conducting counter-hybrid measure d_i	Drawn from different truncated normal distributions for each counter-hybrid measure d_i [170]
q_{ij}	Probability of the adversary conducting hybrid operation c_j after counter-hybrid measure d_i	Drawn from different Beta distributions for each counter-hybrid measure
w_{ijh}	Probability of potential damage θ_h after the adversary conducts hybrid conduct c_j and defender counter-hybrid measure d_i	Drawn from different Dirichlet distributions for each counter-hybrid measure and hybrid operation combination [228]

Despite the semi-synthetic nature of the experiments, all the counter-hybrid measures considered have been derived from real-world examples and their impacts have been scored by experts, ensuring reflection of real-world variability and available empirical evidence. As the modeling approach enables the exploration of dynamics that cannot be empirically tested in the real world, the semi-synthetic nature of the data is a necessary instrument to conduct this analysis. Furthermore, the flexibility of the proposed framework ensures its applicability to other domains and hybrid threat types, as the underlying principles and interactions are generalizable beyond the specific scenarios tested. This adaptability enhances its utility in addressing a broad spectrum of hybrid threat challenges.

6.2.2 Experimental Design

A scenario is considered in which the defending agent B fears that revisionist agent A attempts to destabilize and harm agent B through hybrid attacks. In particular, the defender is aware of agent A's offensive capabilities in the cyber and information domains and is concerned that the latter will carry out a high-scale cyber-attack against its critical infrastructures, such as power plants, and grids, water management facilities, ports, the healthcare system and/or other essential services. Offensive cyber

operations constitute a clear example of a hybrid threat below the threshold of large-scale armed conflict. Indeed, cyber operations have become more prevalent in recent years due to the technical, physical and logical layers of cyberspace and the pervasive use of networks and technologies in our daily life [14].[2] Furthermore, offensive cyber operations may well produce material consequences resulting in considerable physical damage such as for instance in the case of Stuxnet in Iran (2009), Shamoon in Saudi Arabia (2012) or NotPetya in over sixty countries around the world (2017).

An anonymized list of plausible hybrid actions was constructed based on a series of real-world malicious cyber operations drawn from the updated datasets compiled by Valeriano and Maness [243], Stirparo et al. [234], and the Council on Foreign Relations,[2] as well as on a review of the relevant literature. In addition, given the exponential development of new technologies and the evolving dynamics in current conflicts, this was complemented with expert imagination, in an effort to anticipate potential courses of action (and response), and key variables in the cyber domain were distilled. Plausible counter-hybrid responses were identified, with experts selecting the top five cross-domain measures against malicious cyber attacks, summarized in Table 6.4. These include both in-domain (cyberspace) and out-domain responses, such as law enforcement, norm development, public diplomacy, and economic sanctions. Viewed through cumulative deterrence, some measures aim to mitigate damage, while others seek to deter aggression by altering adversaries' cost-benefit assessments. Alternatively, the defender can choose to refrain from engaging in counter-hybrid measures. Appendix C.1 contains the specifications of the measures, including probability distributions that will be used to distill the costs, the probability of successful deterrence, and the probability of mitigating damaging effects for each of the different counter-cyber measures.

Estimating the exact costs and damages from cyber attacks on critical infrastructure is challenging [88], but experts agree that the defender's capacity to detect and recover from such attacks significantly shapes their overall impact [163, 237]. Accordingly, three categories of impact are considered: θ_1 denotes severe disruption due to ineffective detection and recovery; θ_2 reflects substantial losses with partial or delayed mitigation; and θ_3 captures limited or negligible effects resulting from effective preventive or responsive measures. As affected entities rarely disclose precise impact data, potential impacts θ_1, θ_2, and θ_3 are sampled from heavy-tailed half-normal distributions, following Lis and Mendel [146], with specifications provided in Appendix C.1 to

[2]Council of Foreign Relations, "Cyber Operations Tracker," accessed December 1, 2022, `https://www.cfr.org/cyber-operations/`

Table 6.4: Five Counter-Hybrid Measures

Domain	Title	Capability	Rationale
Military	Active intelligence sharing	Actively share your collective intelligence with allies.	Active intelligence sharing is useful for bringing your allies on board for political, public or other types of collective attribution.
Cyber	Boost cyber resilience at the wider societal level	Introduce legislation or collaboration that requires individuals and companies to adopt sufficient levels of cyber resilience, based on the specific risk exposure of the subject.	Boosting cyber resilience at the broader societal level is one of the core tenets of a whole-of-society approach to hybrid/cyber threats.
Cyber	Employ offensive cyber capabilities	Use offensive cyber operations in order to undermine the target.	It is often the case that cyber attacks are not directly attributable. In the short term, it provides a covert way to influence the target. The effect of an offensive cyber attack is scalable.
Legal	Market restrictions	Introduce legislation to restrict an opponent from accessing your market in a specific sector (such as ICT).	Such a measure reduces the possibilities for an adversary to exploit vulnerabilities but it may also clash with other legal commitments (see international trade commitments against market restrictions based on nationality).
Diplomatic	Open deterrence messaging through strategic communications	Communicate one's strategic posture in order to convince a target to comply with one's strategic aims.	Being transparent with the hostile actor regarding one's own strategic strengths and possible actions. This increases the possibility of a better-informed decision by the hostile actor.

capture the high variance and uncertainty inherent to such estimates.

6.2.3 Results

In this section, the results are discussed based on the experimental setup of the previous section. First, the results of optimizing for the counter-hybrid measure are presented

using estimated deterrence probabilities. This is followed by an analysis of the subgame perfect equilibria, where the decision to conduct the hybrid attack is modeled as the agent's strategic choice.[3]

For each of the 1000 experiments, the effectiveness of the counter-hybrid measure is ranked in terms of total pay-off for defending agent B from least optimal to most optimal. A count plot of the rank of the counter-hybrid measures for all experiments is displayed in Figure 6.6.

In summary, despite the high cost of imposing market restrictions (d_4), they are often deemed the most optimal counter-hybrid measure due to the potential to mitigate damages and to deny the adversary's ability to carry out attacks. Intelligence sharing (d_1), valued for its cost-effectiveness, plays a crucial role in mitigating attack damage by enabling timely defensive actions and fostering political support for collective responses. While offensive cyber operations (d_3) could disrupt enemy capabilities, they carry high risks of escalation and have variable success rates. Boosting cyber resilience (d_2), though costly, is consistently rated effective for both damage mitigation and deterrence. Open deterrence communication (d_5) hinges on the adversary's responsiveness to threats, requiring detection, attribution, and communication capabilities to be successful. Lastly, in very rare draws, abstaining from counter-hybrid

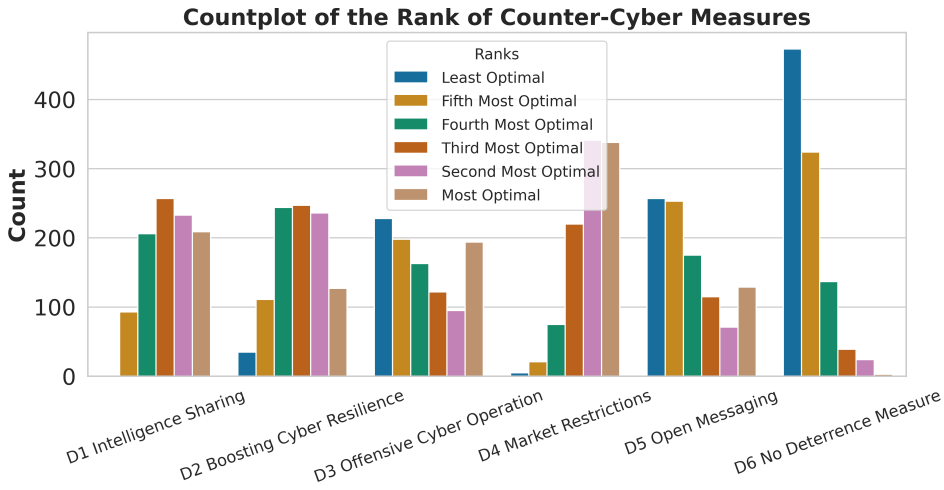

Figure 6.6: The count of the specific rank that each of the counter-hybrid measures is computed to attain.

[3]The modeling effort is publicly available at `https://github.com/HCSS-Data-Lab/Hybrid-Threat-Implementation`.

measures (d_6) emerges as the most effective strategy.

To derive more meaningful insights, the focus is shifted from the specific outcomes of individual measures to the broader results that can be drawn from the overarching characteristics of these measures, especially considering that a well-designed elicitation protocol would significantly enhance the interest and reliability of individual results, while the same overarching characteristics would prevail.

Overall, the measures vary in several ways: some measures rely on their ability to dissuade the adversary through punishment (open messaging, offensive cyber operations), others count on the ability to mitigate the potential damage when confronted with an attack (intelligence sharing), and there are also measures that are a mixture of both deterrence by denial and enhancing resilience (boosting cyber resilience and imposing market restrictions). While these characteristics are distributed evenly among the optimal measures, the measure designated as optimal in most cases - i.e., imposing market restrictions - is also the most versatile one with respect to both dissuasion as well as resilience enhancement. In addition, the variance of the cost, ability to mitigate the damage and ability to deter are different for each of the measures as their impact is mediated by favoring conditions. For instance, while deterrence by punishment measures (open messaging and offensive cyber operation) can be very effective counter-hybrid measures, they are also among the most ineffective measures for some experiments as illustrated by Figure 6.6. This is because they rely heavily on their effect on the adversary's strategic calculus. When dissuasion is not successful, they do not contribute to mitigating the damaging impact of hybrid conduct, leaving the defender exposed.

In order to test how variations in input parameters influence the output of the model, sensitivity analyses were conducted using the state-of-the-art tool of van Stein et al. [233] for each of the counter-hybrid measures. Figure 6.7 presents the SHAP summary plot for imposing market restrictions, which serves as a representative example of the broader sensitivity analysis conducted. As can be observed from the figure, the probability of successfully deterring the adversary q_{14}, as well as the costs of the measure γ_4 are the most influential factors for the effectiveness of the counter-hybrid measure. While the measure's ability to mitigate the negative impact is comparatively less influential overall, its role in increasing the probability of a negligible impact (ω_{413}) or reducing the likelihood of a severe impact (ω_{411}) remains a significantly important factor in evaluating its effectiveness. As the effectiveness of the measure is strongly influenced by its ability to dissuade the adversary, there is a need to consider this factor not only probabilistically but also within a game-theoretic framework. By doing

so, the subgame perfect equilibria were computed as detailed in the previous section, providing a more comprehensive evaluation of the measures in a strategic context.

These subgame perfect equilibria for the same 1000 experiments are displayed in Table 6.5, reflecting outcomes where agents seek to optimize their pay-off rationally. The low occurrence of hybrid attacks indicates that the chosen counter-hybrid measures focus on deterring the adversary rather than mitigating the consequences of a hybrid attack. Given that the adversary's strategic calculus is assumed to be known, the defending agent can strategically select the most cost-effective counter-hybrid measure that successfully deters the adversary from launching an attack. This explains why intelligence sharing is preferred over market restrictions when both measures suffice to deter the adversary. Moreover, when the strategic calculus is such that the

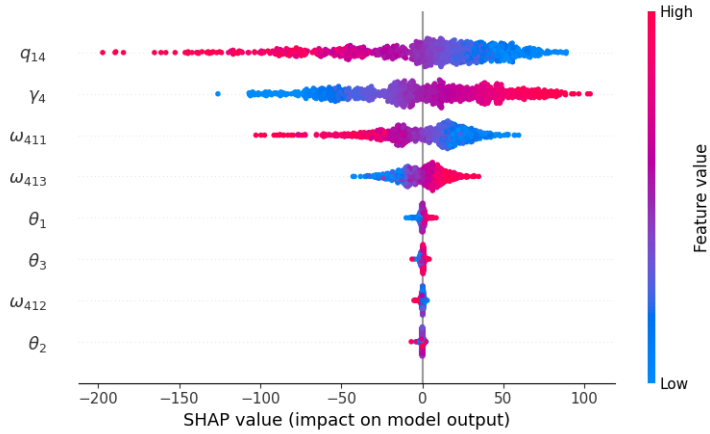

Figure 6.7: SHAP summary plot for counter-hybrid measure imposing market restrictions: The y-axis represents the features ranked by their importance to the model output. The x-axis shows the SHAP value, indicating the magnitude and direction of each feature's impact on the model output. The color gradient reflects the feature values.

Table 6.5: Subgame Perfect Equilibria Outcome Occurences

Counter-hybrid Measures	Hybrid Operation	
	Attack	No Attack
D1 Intelligence Sharing	15	697
D2 Boosting Cyber Resilience	1	98
D3 Offensive Cyber Operation	3	76
D4 Market Restrictions	0	110
D5 Open Messaging	0	0
D6 No Deterrence Measure	0	0

adversary is likely to proceed with a hybrid operation regardless of the counter-hybrid measure, the subgame perfect equilibria suggest that the defending agent should commit to cost-efficient counter-hybrid measures, such as intelligence sharing, to mitigate the impact of the attack.

6.2.4 Conclusion and Future Work

This section proposed novel approaches to evaluate cross-domain counter-hybrid measures by balancing cost, deterrence, and damage mitigation under uncertainty, modeled probabilistically and game-theoretically. The evaluation included 1000 scenarios of malicious cyber operations, incorporating countermeasures derived from literature and a mini-Delphi survey.

While general validation of the results remains challenging due to the ambiguous nature of hybrid threats and the reluctance of targeted parties to disclose information, contextualizing the findings within established models and prior studies provides valuable insights. The results confirm that the most effective counter-hybrid measures tend to be the most costly, supporting previous claims that hybrid threats can economically strain defenders [25], and underscoring the importance of prioritizing cost-effective, cross-domain strategies [18]. The analysis also supports democratic deterrence models advocating a whole-of-society approach [242], particularly highlighting the utility of non-military measures like market restrictions and intelligence sharing [258]. Finally, the sensitivity analysis reinforces the findings on existing modeling efforts on deterrence in the cyber realm [130], emphasizing that the effectiveness of such threats is heavily contingent upon the adversary's susceptibility to countermeasures. Deterrence by punishment measures, for instance, only work well when the adversary is responsive to such measures [158]. The modeling exercise indicates that even a small enhancement in understanding the aggressor's plausible receptiveness to counter-hybrid measures could lead to a significant enhancement in the assessment of the effectiveness of measures. This implies that resources spent on anticipating the adversary's reaction to possible counter-hybrid measures are conditional on the effect of the counter-hybrid measures.

Two key directions for future research are identified. First, applying elicitation methods for probabilities and utilities from Chapters 3 and 4, in direct consultation with policy-makers, would improve the precision of model inputs. Second, extending experiments to cross-domain hybrid threat scenarios beyond the cyber domain would increase realism by better capturing the complexity of modern hybrid threats.