Adversarial Risk Analysis (Overview)

David Banks[1], Victor Gallego[2], Roi Naveiro[2], and David Ríos Insua[3,2]

[1]Duke University
[2]Instituto de Ciencias Matematicas, ICMAT-CSIC
[3]University of Shanghai for Science and Technology

Abstract

Adversarial risk analysis (ARA) is a relatively new area of research that informs decision-making when facing intelligent opponents and uncertain outcomes. It enables an analyst to express her Bayesian beliefs about an opponent's utilities, capabilities, probabilities and the type of strategic calculation that the opponent is using. Within that framework, the analyst then solves the problem from the perspective of the opponent while placing subjective probability distributions on all unknown quantities. This produces a distribution over the actions of the opponent that permits the analyst to maximize her expected utility. This overview covers conceptual, modeling, computational and applied issues in ARA.

1 Introduction

Adversarial risk analysis (ARA) guides decision-making when there are intelligent opponents and uncertain outcomes. It is a decision-theoretic alternative to classical game theory that uses Bayesian subjective distributions to model the goals, resources, beliefs, and reasoning of the opponent. Within this framework, the analyst solves the problem from the perspective of her opponent while placing subjective probability distributions on all unknown quantities. This provides a distribution over the actions of the opponent that enables her to maximize her expected utility, accounting for the uncertainty she has about the opponent.

Game theory is the standard approach to adversarial reasoning, and it has been applied, among many other areas, in politics (Brams, 2011), biology (Hammerstein and Selten, 1994), economics (Samuelson, 2016), social sciences (Shubik, 1982) and cybersecurity (Shiva et al., 2010). The cornerstone of game theory is the Nash equilibrium, in which no opponent can improve their outcome by any unilateral action. Of course, this methodology has been generalized and extended in many ways (Halpern, 2008). Nonetheless, the fundamental premises of game theory have been criticized (Hargreaves-Heap and Varoufakis, 2004; Young, 2004). The main concerns are:

- The classical formulation generally assumes that all participants in the game have the same beliefs about the other players, and that all players know those beliefs are known. This common knowledge assumption is frequently unrealistic. For example, in a three-person auction, it is quite possible for players A and B to have different distributions for the value to player C of the item on offer and that they will conceal that information.

- The field of behavioral economics has repeatedly demonstrated that humans do not act as game theory would prescribe (Camerer, 2003; Gintis, 2009), so it is a poor predictor of real-world decisions.

- Game theory generally finds pessimistic and expensive solutions; it assumes that the opponent will do the most damaging thing possible, and thus the analyst must invest in expensive protection. But often opponents are not so ruthless or so calculating.

- The game theory paradigm assumes that each opponent is hyperlogical and capable of infinite computation. Researchers have weakened these assumptions in many ways; e.g., through bounded rationality (Simon, 1955), prospect theory (Tversky and Kahneman, 1979), and computing constraints (Osborne and Rubinstein, 1998). But the core idea remains problematic.

- Although the judgments of a supported agent could be arguably well assessed, as explained in Keeney (2007), knowledge about the other agents' judgments is less precise, since it requires them to reveal beliefs and preferences. Such disclosure is unrealistic in cybersecurity, counterterrorism, and other domains, where information is concealed by adversaries.

- In spite of this, uncertainty in the adversary judgments is not frequently acknowledged and game theoretic solutions are often sensitive to such inputs as shown in examples in Ekin et al. (2019).

- Game solutions can have multiple equilibria with no clear criteria to choose among them (Raiffa et al., 2002). Opponent A may select a strategy corresponding to one equilibrium, while opponent B could play a strategy corresponding to a different equilibrium; see Russell (1998) for developments in coordination games.

- Except for toy problems, computing the game theoretic solution is difficult, often to the point of impracticability; see Nisan et al. (2007) for issues in computational game theory.

None of these criticisms is a sockdologer argument, and game theorists have worked hard to shore up these deficiencies. But although the equilibrium perspective is attractive, it is also compelling to maximize one's expected utility as in ARA.

Indeed, the ARA perspective is actually very natural. One builds a model for the thinking of one's opponent, and then maximizes expected utility under that model. For example, when someone asks the boss for a raise, the first step is to understand what the boss values and what criteria he will use. If that understanding is correct, the employee has a reasonable chance of success; but if the model for the boss's thought process is wrong, then the outcome will likely be disappointing.

One of the advantages of ARA is its ability to partition the uncertainty into three separate components (Merrick and Parnell, 2011). The first is *aleatory uncertainty*, which is the uncertainty in the outcome conditional on the choices of each the opponents; this is handled by conventional statistical risk analysis Bedford et al. (2001); Cox Jr (2013). The second component is *epistemic uncertainty*. Usually, one does not know an opponent's utility function, nor his assessment of the probability of outcomes conditional on the decisions that are made, nor even his resources; but a Bayesian should be comfortable making subjective probability assessments about each of these quantities. The third component is *concept uncertainty*. This reflects the fact that the analyst does not know how her opponent is making his decision. Perhaps he is a game theorist and seeks an equilibrium solution. Or perhaps he randomizes, or follows some other protocol.

To make this a little more concrete, consider a sealed bid auction between Daphne and Apollo, each of whom wants to own a first edition of the *Theory of Games and Economic Behavior*. Daphne's aleatory uncertainty is the value she receives conditional on her bid and Apollo's. If she has not been allowed to examine the book prior to the auction, then its condition is a random variable–perhaps it is old and torn, or perhaps it has marginalia written by John Nash, and both circumstances affect its value. Epistemic uncertainty arises because Daphne does not know the value of the book to Apollo, nor what he thinks is the probability that he will win with a bid of x dollars, nor how much money Apollo has. The concept uncertainty reflects the fact that Daphne does not know whether Apollo is determining his bid using classical game theory (Milgrom and Weber, 1982), or whether he is simply bidding some unknown fraction of his true top-dollar value, or using some other principle.

As we shall see, ARA can lead to mathematical formulations whose complexity is comparable to that found in game theory. There is no easy way to reason about other people's strategic thinking in realistic situations. Nonetheless, the decision-theoretic approach sidesteps some of the common criticisms of game theory.

2 Auctions

To illustrate how ARA works, we shall build out the discussion of auctions. It is convenient to assume that Daphne and Apollo have had the opportunity to examine the book before the sealed-bid auction, which simplifies things by eliminating the aleatory uncertainty. And suppose Daphne has a subjective distribution $F(x)$ about the probability she will win the auction with a bid of x. Then, if her true top-dollar value for the book is x_0, assuming risk neutrality, she maximizes her expected utility by bidding

$$x^* = \text{argmax}_{x \in \mathbb{R}^+} (x_0 - x) F(x).$$

Her utility is her profit $x_0 - x$, and the probability of realizing that profit is $F(x)$.

How does Daphne acquire her subjective probability $F(x)$ regarding her probability of winning? There are several ways, and these depend upon which solution concept she thinks Apollo will use. In the simplest, she models Apollo as non-strategic, and believes he will bid some unknown fraction of his true, also unknown, value. As a Bayesian, Daphne is comfortable placing a subjective distribution $G(p)$ on the fraction of the true value that Apollo will bid ($0 \leq p \leq 1$), and she can place a subjective distribution $H(v)$ on the value of the book to Apollo. Then, simple calculation shows that

$$F(x) = \mathbb{P}[PV \leq x] = \int_0^\infty \int_0^{x/v} g(p)h(v)dpdv \tag{1}$$

where $g(p)$ and $h(v)$ are the densities of $G(p)$ and $H(v)$, respectively.

A second common solution concept Apollo might use is the Bayes Nash equilibrium (BNE). The BNE formulation makes a strong common knowledge (CK) assumption: both Apollo and Daphne have distributions H_A and H_D for each other's valuation, and each knows both distributions and knows that the other knows them. This leads to solving a system of first-order ordinary differential equations. For an asymmetric auction, when $H_A \neq H_D$, no general solution exists, although it is known that if H_A and H_D are differentiable then a unique solution exists and it is also differentiable (Lebrun, 1999). Previous attempts at solutions are based on the backshooting algorithm (Hubbard et al., 2013), but Fibich and Gavish (2011) have shown that all such solutions are inherently unstable. Au (2014) has an algorithm that succeeds, based on the limit of discretized bids and points of indifference.

From an ARA perspective, the CK assumption can be replaced by something more reasonable. Daphne has a subjective belief about the distribution H_A of Apollo's value, and she has a subjective belief about H_D, the distribution she believes he thinks is the distribution of her value. The H_A and H_D are epistemic uncertainties.

In this BNE framework, she thinks Apollo will solve a system of coupled equations:

$$\begin{aligned} \text{argmax}_{d \in \mathbb{R}^+} (D^* - d) F_A(d) &\sim F_D \\ \text{argmax}_{a \in \mathbb{R}^+} (A^* - a) F_D(a) &\sim F_A \end{aligned} \tag{2}$$

where $D^* \sim H_D$, $A^* \sim H_A$, and F_A and F_D are the distributions of Apollo's and Daphne's bids, respectively. The equilibrium solution gives F_A, her best guess, under the BNE solution concept, of the distribution for Apollo's bid. Now Daphne should step outside the BNE framework and solve $x^* = \arg\max_{x \in \mathbb{R}^+} (x_0 - x) F_A(x)$ where x_0 is Daphne's true value for the book, which of course is known to her.

For a two-person auction, this solution is the same as found from a BNE analysis. But with more than two persons, the solution will diverge: Daphne can believe their opponent will bid high, but also think that Apollo thinks he will bid low. This allows non-common knowledge, and to distinguish this case from the BNE, we call it a ”mirroring argument” since Daphne is trying to do the analysis she thinks her opponents are doing.

A third important solution concept is level-k thinking (Stahl and Wilson, 1995). A level-0 bidder is non-strategic. A level-1 bidder thinks the opponent is non-strategic, and optimizes her bid against a model for his behavior (as previously discussed). A level-2 bidder thinks her opponent is a level-1 thinker who is modeling her as a level-0 thinker. This can get complicated, but except in highly structured games such as chess, empirical studies indicate that people rarely go beyond level-2 thinking (Stahl and Wilson, 1994).

To illustrate, suppose Daphne is a level-2 bidder. She thinks Apollo is a level-1 bidder, who models her as a non-strategic bidder. Assume she believes that Apollo thinks her true value for the book has the uniform distribution on $[\$100,\ \$200]$ and that she bids a random fraction p of her value, where $F(p) = p^9$ of $0 \leq p \leq 1$. From (1), it follows that Apollo thinks her bid will have the distribution

$$F(d) = \left(\frac{d - 100}{200}\right)^9 \qquad 100 \leq d \leq 200.$$

Apollo needs to find his best response, by making the bid a^* such that

$$a^* = \text{argmax}_{a \in \mathbb{R}^+} (A_0 - a) F(a)$$

where A_0 is his true value for the book, which is a random variable to Daphne since she does not know it. Using basic calculus to find the maximum shows

$$0 = \frac{d}{da}(A_0 - a)F(a) = \frac{9a^8}{200^9}(A_0 - a) - \left(\frac{a}{200}\right)^9$$

so Apollo's bid should be 90% of his true value.

Daphne does not know Apollo's value, but as a Bayesian, she has a subjective distribution for it. Suppose she thinks it is the triangular distribution on on $[\$140,\ \$200]$ with a peak at \$170. Then it follows that she thinks his bid will have the triangular distribution on $[\$126,\ \$180]$ with a peak at \$153. If her true value for the book is \$175, then her expected profit is maximized with a bid of \$161.67, which completes the level-2 analysis.

3 ARA in general

In ARA one takes the side of one agent, using only her beliefs and knowledge, rather than assuming CK and trying to solve all of the agents' problems simultaneously. The selected agent must have (1) a subjective probability about the actions of each opponent, (2) subjective conditional probabilities about the outcome for every set of possible choices, and (3) accurate knowledge of her own utility function. Obviously, in practice a person will only have approximate knowledge of these quantities, but that is sufficient for many applications.

	1	**2**	**3**	**4**	**5**	**6**
0	u_A	$p_A(\cdot\|d,a)$	$\pi_A(d)$	u_D	$p_D(\cdot\|d,a)$	$\pi_D(a)$
1	U_A^1	$P_A^1(\cdot\|d,a)$	$\Pi_A^1(d)$	U_D^1	$P_D^1(\cdot\|d,a)$	$\Pi_D^1(a)$
2	U_A^2	$P_A^2(\cdot\|d,a)$	$\Pi_A^2(d)$	U_D^2	$P_D^2(\cdot\|d,a)$	$\Pi_D^2(a)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 1: Cognitive load for different kinds of ARA.

Thus Daphne believes Apollo has probability $\pi_D(a)$ of choosing action $a \in \mathcal{A}$. She has a subjective probability $p_D(s \,|\, d, a)$ for each possible outcome $s \in \mathcal{S}$ given every choice $(d, a) \in \mathcal{D} \times \mathcal{A}$, where $\mathcal{D}$ is Daphne's set of possible actions. And she knows her own utility $u_D(d, a, s)$ for each combination of outcome and pair of choices.

Daphne maximizes her expected utility by choosing the action d^* such that

$$\begin{aligned} d^* &= \text{argmax}_{d\in\mathcal{D}}\, \mathbb{E}_{\pi_D, p_D}[u_D(d, A, S)] \\ &= \text{argmax}_{d\in\mathcal{D}} \int_{s\in\mathcal{S}} \int_{a\in\mathcal{A}} u_D(d, a, s) p_D(s \,|\, d, a) \pi_D(a)\, da\, ds \end{aligned}$$

where A is the random action chosen by Apollo and S is the random outcome that results from choosing A and d.

In practice, the most difficult quantity to obtain is $\pi_D(a)$. The $p_D(s \,|\, d, a)$ is found by conventional risk analysis and $u_D(d, a, s)$ is a personal utility. Risk analysis and utility self-assessment are not easy, but both are mature fields and researchers know how to proceed.

Previously, we laid out ARA methods for obtaining $\pi_D(a)$ in auctions for the cases of the the non-strategic opponent, the Nash equilibrium seeking opponent, the opponent whose analysis mirrors that of the decision-maker, and the opponent who is a level-k thinker. Implementing these approaches imposes different cognitive loads upon the analyst.

Table 1 shows how the cognitive load depends upon the kind of ARA. Each row corresponds to a different level of reasoning in level-k thinking. It displays the quantities that Daphne must assess in order to implement a level-k analysis. Row 0 corresponds to the utilities and beliefs of Daphne and Apollo, as perceived by themselves. Subsequently, row k contains the additional utilities and probabilities that Daphne would have to assess in order to perform a level-k analysis. The upper case characters in rows 1 and greater indicate that these quantities are all random variables to Daphne.

The first column contains the utility functions she ascribes to Apollo. The second column contains the conditional probabilities of the outcome, given her choice and Apollo's, that she ascribes to Apollo. The third column contains what she thinks is Apollo's distribution over her choice. The fourth column contains what Daphne believes are the utility functions that Apollo ascribes to her. The fifth column contains the probabilities of the outcome, conditional on both her action and Apollo's, that she believes Apollo ascribes to her. The sixth column contains her opinion of what Apollo thinks is her distribution for he will do.

In terms of the table, different solution concepts require information in different cells:

- Traditional game theory requires cells (0,1), (0,2), (0,4), (0,5) and assumes that these are CK.
- The non-strategic adversary analysis requires cells (0,4), (0,5) and (0,6), where the (0,6) cell is assessed from historical data and/or expert opinion.
- When the adversary seeks a Nash equilibrium solution, the analysis requires cells (0,4), (0,5) and (1,1), (1,2), (1,4), and (1,5). It uses these last four cells to infer cell (0,6).
- The level-k adversary approach requires cells (0,4), (0,5) and:
  - for a level-1 analysis, cells (1,1), (1,2) and (1,3) can produce (0,6);
  - for a level-2 analysis, cells (2,4), (2,5) and (2,6) produce (1,3), which, with (1,1), and (1,2) then produce (0,6);
  - and so forth for larger k.
- The mirror equilibrium approach requires cells (0,4), (0,5) and uses a consistency condition between (1,1), (1,2), (1,3) and (1,4), (1,5) and (1,6) to produce (0,6).

The main message is that all of these methods entail significant effort.

4 Basic concepts in ARA

In this section, we compare game-theoretic and ARA concepts in two template models. Their basic structures may be simplified or made more complex, by removing or adding nodes. Assume there is a defender (D) who chooses her decision $d \in \mathcal{D}$ and an attacker (A) who chooses his attack $a \in \mathcal{A}$. In principle, the agents are assumed to act so as to maximize expected utility (or minimize expected loss) (French and Rios Insua, 2000). Bi-agent influence diagrams (BAIDS) (Banks et al., 2015) describe the problems: the circular nodes represent uncertainties, the hexagonal show utilities, and square nodes indicate decisions. The arrows point to decision nodes (meaning that decisions are made given the values of preceding nodes) or chance and utility nodes (these events or consequences are influenced by predecessors). Colored nodes indicate relevance to just one of the agents (white, defender; gray, attacker), striped nodes are relevant to both agents.

4.1 Simultaneous defend-attack games

First consider simultaneous games—the agents decide their actions without knowing the action chosen by each other. Its template is shown in Fig. 1.

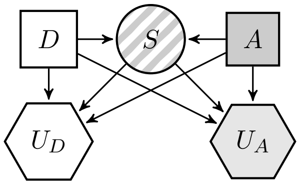


Figure 1: Basic two player simultaneous defend-attack game.

The consequences for both agents depend upon the outcome of the attack, $s \in S$. Each agent has their own assessment on the probability of s, which depends on d and a, respectively called $p_D(s|d,a)$ and $p_A(s|d,a)$. Similarly, their utility functions are $u_D(d,a,s)$ and $u_A(d,a,s)$. Both agents A and D know the expected utility that a pair (d,a) would provide them, $\psi_A(d,a) = \int u_A(d,a,s) p_A(s|d,a)\, \mathrm{d}s$, and $\psi_D(d,a) = \int u_D(d,a,s) p_D(s|d,a)\, \mathrm{d}s$. A Nash equilibrium (NE) (d^*, a^*) in this game satisfies $\psi_D(d^*,a^*) \geq \psi_D(d,a^*)\ \ \forall\, d \in \mathcal{D}$ and $\psi_A(d^*,a^*) \geq \psi_A(d^*,a)\ \ \forall\, a \in \mathcal{A}$.

If utilities and probabilities are not CK, one may model the game as one with incomplete information (Harsanyi, 1967) using the notion of types: each player will have a type known to him but not to the opponent, and thus is private information. The type $\tau_i \in T_i$ determines the agent's utility $u_i(d,a,s,\tau_i)$ and probability $p_i(s|d,a,\tau_i)$, $i \in \{A, D\}$. Harsanyi proposes Bayes-Nash equilibria (BNE) as solution concept, still under a strong CK assumption: the adversaries' beliefs about types are CK through a common prior $\pi(\tau_D, \tau_A)$ (moreover, the players' beliefs about other uncertainties in the problem are also CK). Define strategy functions by associating a decision with each type, $d : \tau_D \to d(\tau_D) \in \mathcal{D}$, and $a : \tau_A \to a(\tau_A) \in \mathcal{A}$. Agent D's expected utility associated with a pair of strategies (d,a), given her type $\tau_D \in T_D$, is

$$\psi_D(d(\tau_D), a, \tau_D) = \int\int u_D(d(\tau_D), a, s, \tau_D) p_D(s|d(\tau_D), a(\tau_A), \tau_D) \pi(\tau_A|\tau_D)\, \mathrm{d}\tau_A \mathrm{d}s.$$

Similarly, we compute the attacker's expected utility $\psi_A(d, a(\tau_A), \tau_A)$. Then, a BNE is a pair (d^*, a^*) of strategy functions satisfying

$$\begin{aligned} \psi_D(d^*(\tau_D), a^*, \tau_D) &\geq \psi_D(d(\tau_D), a^*, \tau_D), \quad \forall\, \tau_D \\ \psi_A(d^*, a^*(\tau_A), \tau_A) &\geq \psi_A(d^*, a(\tau_A), \tau_A), \quad \forall\, \tau_A \end{aligned}$$

for every d and every a, respectively.

However, the common prior assumptions are still unrealistic (Antos and Pfeffer, 2010), e.g. in security contexts. We thus weaken them when supporting D, who should maximize her expected utility through

$$d^* = \operatorname*{arg\,max}_{d \in \mathcal{D}} \int\int u_D(d,a,s) p_D(s \mid d,a) \pi_D(a)\, \mathrm{d}s\, \mathrm{d}a = \operatorname*{arg\,max}_{d \in \mathcal{D}} \int \psi_D(d,a) \pi_D(a)\, \mathrm{d}a, \tag{3}$$

where $\pi_D(a)$ models her beliefs about the attacker's decision a. To assess such probability distribution, suppose D thinks that A maximizes expected utility, so that he seeks $a^* = \arg\max_{a \in \mathcal{A}} \int \left[\int u_A(d,a,s) p_A(s|d,a) \, \mathrm{d}s \right] \pi_A(d) \, \mathrm{d}d$. She will typically be uncertain about A's required inputs (u_A, p_A, π_A). If one models all information available to her about these elements through a subjective probability distribution $F \sim (U_A, P_A, \Pi_A)$, mimicking (3), one propagates the uncertainty into the distribution of

$$A \mid D \sim \arg\max_{a \in \mathcal{A}} \int \left[\int U_A(d,a,s) \; P_A(s \mid d,a) \, \mathrm{d}s \right] \Pi_A(D = d) \, \mathrm{d}d. \tag{4}$$

Usually, to estimate $\pi_D(a)$ we use Monte Carlo simulation, drawing K samples from F $\{u_A^k, p_A^k, \pi_A^k\}_{k=1}^K \sim F$, finding for each of them

$$A_k^* = \arg\max_{a \in \mathcal{A}} \int \left[\int u_A^k(d,a,s) \; p_A^k(s \mid d,a) \, \mathrm{d}s \right] \pi_A^k(d) \, \mathrm{d}d.$$

and approximating $\pi_D(a)$ using the empirical frequencies

$$\widehat{\pi}_D(d) = \frac{\# \{A_k^* = d\}}{K}$$

U_A and P_A may be directly elicited from D. However, eliciting Π_A may require deeper analysis and level-k thinking as she needs to model how A analyzes D's problem (this is why we condition on the distribution of D in (4)). This entails computing $D \mid A^1 \sim \arg\max_{d \in \mathcal{D}} \int \int U_D(d,a,s) \;\; P_D(s \mid d,a) \, \mathrm{d}s \; \Pi_D(A^1 = a) \, \mathrm{d}a$, assuming she is able to assess: (U_D, P_D), representing her knowledge about how A estimates $u_D(d,a,s)$ and $p_D(s|d,a)$; and $\Pi_D(A^1)$, her beliefs about A's estimate of the probability of the defender playing action d. A^1 is A's random decision within D's second level of recursive thinking. Again, eliciting this last element may require further thinking from D, leading to a recursion of nested models, related to the level-k thinking concept in Stahl and Wilson (1994). This recursion might stop at a level in which D lacks the information necessary to assess the corresponding distributions. At that point, she could assign a non-informative distribution (French and Rios Insua, 2000). Further details are in Rios and Rios Insua (2012).

In general, and particularly in simultaneous games, ARA and game theory use different conditions and assumptions, so it is natural that they lead to different solutions see, for instance, a cybersecurity example in Rios Insua et al. (2019). As shown in Ekin et al. (2019), ARA solutions tend to be more robust than NE. This is related with the fact that NE solutions are based on point estimates of the attacker's judgments, whereas ARA solutions take account of the uncertainty about such estimates, leading to solutions that better reflect the analyst's beliefs. Also, in the particular case of simultaneous games, BNE and ARA solutions do not necessarily coincide.

4.2 Sequential defend-attack games

Consider now sequential games. The defender chooses her decision d and then A, after observing d, chooses his attack a. The BAID for this game is shown in Figure 2. The arc from D to A indicates that D's choice is known to A. This situation has been called the sequential Defend-Attack (Brown et al., 2006) game or Stackelberg game (Gibbons, 1992).

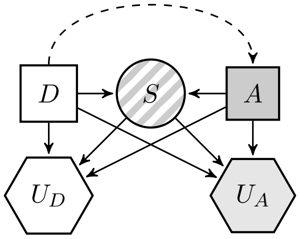


Figure 2: Basic two player sequential defend-attack game

The game-theoretic solution does not require A to know D's judgments, as he observes her decision. However, D must know those of A, which is the CK condition in this case. To solve, one computes both agents' expected utilities at node S: $\psi_A(d,a)$ and $\psi_D(d,a)$. and finds $a^*(d) = \arg\max_{a\in\mathcal{A}}\ \psi_A(d,a)$, A's best response to D's action d. Then, D's optimal action is $d^*_{\text{GT}} = \arg\max_{d\in\mathcal{D}}\ \psi_D(d, a^*(d))$. The pair $\left(d^*_{\text{GT}},\ a^*(d^*_{\text{GT}})\right)$ is a Nash equilibrium and, indeed, a sub-game perfect equilibrium.

The CK condition is weakened if we assume only partial information, leading to games under incomplete information (Harsanyi, 1967). In this case, to model D's uncertainty about A's elements, the attacker is assumed to belong to a certain type τ_A, unknown to the defender. A's optimal response to d depends on his type, and is denoted as $a^*(d,\tau_A)$. If we model D's uncertainty about τ_A through a prior $\pi(\tau_A)$, clearly, D's optimal action is

$$\arg\max_{d\in\mathcal{D}} \int \psi_D[d, a^*(d,\tau_A)]\pi(\tau_A)\,\mathrm{d}\tau_A. \tag{5}$$

The ARA approach to sequential games is different in its formulation. Given the lack of CK, D is uncertain about the attacker's response to her action d and models this uncertainty through the distribution $p_D(a|d)$. Then, her expected utility would be $\psi_D(d) = \int \psi_D(d,a)\,p_D(a|d)\,\mathrm{d}a$ with optimal decision $d^*_{\text{ARA}} = \arg\max_{d\in\mathcal{D}} \psi_D(d)$.

To elicit $p_D(a|d)$, D benefits from modeling A's problem. For this, she would use all information available about u_A and p_A to model her uncertainty about these elements through a distribution $F = (U_A, P_A)$ over the space of utilities and probabilities. This induces a distribution over A's expected utility, being his random expected utility is $\Psi_A(d,a) = \int U_A(d,a,s)P_A(s|d,a)\,\mathrm{d}s$. Then, D finds $p_D(a|d) = \mathbb{P}_F\left[a = \arg\max_{x\in\mathcal{A}} \Psi_A(x,d)\right]$. In general, one can use Monte Carlo (MC) simulation to approximate $p_D(a|d)$.

As in simultaneous games, NE and ARA solutions are different in this sequential setting, as they are based in different assumptions. However, BNE and ARA solutions coincide in this case, although their operational interpretations are quite different. As mentioned, computing the ARA solutions requires eliciting $p_D(a|d)$, which in turn requires modelling D's uncertainty about A's elements through a distribution $F = (U_A, P_A)$ over the space of utilities and probabilities. Without loss of generality, assume that both U_A and P_A are defined over a common probability space $(\Omega, \mathcal{A}, \mathcal{P})$ with atomic elements $\omega \in \Omega$ (Chung, 2001). This induces a distribution over the Attacker's expected utility $\psi_A(d, a)$, being its random expected utility $\Psi_A^\omega(d, a) = \int U_A^\omega(d, a, s) P_A^\omega(s|d, a) \, \mathrm{d}s$. In turn, this induces a random optimal alternative defined through $A^*(d)^\omega = \arg\max_{x \in \mathcal{A}} \Psi_A^\omega(d, x)$. Then, the defender would find

$$p_D(a|d) = \mathbb{P}_F \left[A^*(d) = a \right] = \mathcal{P}(\omega : A^*(d)^\omega = a). \tag{6}$$

Now, ω and $\mathcal{P}$ could be reinterpreted, respectively, as the type τ_A and the common prior π in Harsanyi's doctrine. Obviously, P_A^ω and U_A^ω correspond to A's probability and utility given his type, respectively; and $A^*(d)^\omega = a^*(d, \tau_A)$ in this new interpretation. With this in mind, the connection between BNE and ARA is straightforward. Notice that we can rewrite the BNE solution for the defender (5) as

$$\arg\max_{d \in \mathcal{D}} \int \left[\int \psi_D(d, a) \delta(a - a^*(d, \tau_A)) \, \mathrm{d}a \right] \pi(\tau_A) \, \mathrm{d}\tau_A,$$

where $\delta(\cdot)$ is the Dirac delta function. If we change the order of the integrals on a and τ_A we get

$$\arg\max_{d \in \mathcal{D}} \int \psi_D(d, a) \left[\int \delta(a - a^*(d, \tau_A)) \pi(\tau_A) \, \mathrm{d}\tau_A \right] \mathrm{d}a. \tag{7}$$

Clearly,

$$\int \delta(a - a^*(d, \tau_A)) \pi(\tau_A) \, \mathrm{d}\tau_A = \mathcal{P}(\tau_A : a^*(d, \tau_A) = a) = p_D(a|d), \tag{8}$$

where the last step follows from (6). Inserting (8) into (7), we recover the ARA formulation of the sequential game. Thus, in the sequential Defend-Attack game, one can interpret the ARA approach in terms of Harsanyi's, although the underlying principles are different. The take home message is that ARA provides a formal decision theoretic based approach to specify the types τ_A and the prior $\pi(\tau_A)$, thus facilitating the implementation of Bayesian games as we discuss next in the general case of facing more than one opponent.

5 Game Theory and ARA: Further relations

ARA operationalizes the Bayesian approach to games (Kadane and Larkey, 1982; Raiffa, 1982), constructing a procedure to make probabilistic predictions about the decisions of an opponent. This section details the relationship between ARA and core foundational concepts in game theory.

We consider again two agents, a defender D and attacker A. To simplify the discussion, assume there is certainty about the actions' consequences and the utilities the agents obtain (should the outcome conditional on the actions be random, one would operate with expected utilities after modeling uncertainty). Each agent has a finite set of possible actions, denoted $\mathcal{D}$ and $\mathcal{A}$ respectively. We discuss the case in which both agents simultaneously implement their actions and obtain their respective utilities $u_i(d, a)$, $i \in \{A, D\}$, which they want to maximize. Assume one can build a distribution $\pi_D(a)$ that expresses D's beliefs about A's decision. Then, the defender should solve

$$d^* = \arg\max_{d \in \mathcal{D}} \sum_{a \in \mathcal{A}} u_D(d, a)\pi_D(a). \tag{9}$$

This is just the Bayesian approach to games, initiated by Kadane and Larkey (1982), Raiffa (1982) and Raiffa et al. (2002) in non-constructive ways. It has been criticized by Harsanyi (1982) and Myerson (1997), among others. Before ARA, there was no formal methodology allowing the analyst to encode her subjective probabilities on the actions of the other agents. Of course, one could claim just deploying standard probability elicitation methods for such purpose (A. et al., 2006), however Ríos Insua et al. (2019) demonstrate the benefits of the decomposition adopted in ARA in adversarial elicitation.

Armed with this setting, we discuss the ARA perspective of several key concepts in game theory, beyond the relations with Nash and Bayes Nash equilibria concepts outlined above.

ARA solutions and dominance. A major tenet in game theory is that the actions proposed for the agents should be non-dominated. It is easy to show that ARA actions are non-dominated under fairly general conditions (Esteban et al., 2020).

Proposition 1 *If $\pi_D(a) > 0$ for every $a \in \mathcal{A}$, then the ARA action d^* defined in (9) is non-dominated.*

ARA solutions and iterative dominance Besides dominance, another key rationality concept in game theory is iterative dominance (Hargreaves-Heap and Varoufakis, 2004). Note, however, that if the opponent's payoff is unknown to the defender, due to the absence of CK, then the agent using ARA will not know which are the non-dominated actions of the rest. However, she can consider possible rankings of the corresponding random utilities and use those to her advantage. The strongest case is when the attacker always gets more utility with one of his actions than with any other, as perceived by the ARA analyst.

Consider the defender's analysis of her opponent, then (Esteban et al., 2020)

Proposition 2 *Let $U_A(d, a)$ be the (random) utility that A receives, according to D, when he implements his a and D plays d. Suppose that there are two actions a and a' for the attacker such that*

$$\max \text{supp}[U_A(d, a)] < \min \text{supp}[U_A(d, a')],$$

for every d. Then, $\pi_D(a) = 0$.

As a consequence, one can perform an ARA analogue of iterative dominance as follows. First, eliminate the dominated actions of the defender by means of Proposition 1; then, eliminate actions of the attacker using Proposition 2. Once completed, try again to eliminate actions from the defender, and repeat until no alternative can be eliminated.

ARA can mimic less demanding conditions than the dominance in Proposition 2, employing ideas from stochastic dominance (Levy, 1998), including *state dominance*, *first order stochastic dominance* or *second order stochastic dominance*.

ARA and ficticious play One approach to forming D's subjective probabilities about the actions of her opponent is to use data from previous interactions among the agents, when relevant history is available. For example, she might be a level-1 thinker who believes A is non-strategic (a.k.a level-0 thinker). In that case, she can use a Dirichlet-multinomial model, e.g. Gelman et al. (2020), and the expected predictive distribution over A's i-th action a_i would be

$$\pi_D(a_i) = \frac{\alpha_i + x_i}{\sum_{j=1}^{|\mathcal{A}|}(\alpha_j + x_j)}, \tag{10}$$

where $|\mathcal{A}|$ is the cardinality of $\mathcal{A}$, α_i is the Dirichlet prior parameter over the probability of a_i and x_i is the number of times that A previously implemented such action. This approach is related with the fictitious play methodology in game theory, summarized in Menache and Ozdaglar (2011). Under appropriate conditions (Brown, 1951), if all players use fictitious play they converge to a NE as the number of plays grows to infinity. But ARA uses it for just one player, it incorporates prior information and we work with a small number of iterations. The approach is easily extended to cases in which agents have longer memories.

ARA and level-1 thinking ARA can aid in eliciting the analyst's beliefs about the other agent by thinking about the problem he would solve. This would be

$$a^* = \arg\max_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}} u_A(d,a)\pi_A(d),$$

where $\pi_A(d)$ represents A's beliefs about D's actions. Since CK does not hold, D does not know u_A nor π_A, but she can describe that uncertainty with a random utility U_A and a random probability Π_A, as previously discussed. This leads her to compute the random optimal action

$$A^* = \arg\max_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}} U_A(d,a)\Pi_A(d). \tag{11}$$

Then, the first agent would find

$$\pi_D(a) = \Pr(A^* = a).$$

Typically, as mentioned in Section 4.1, this would be estimated by simulation, sampling from U_A and Π_A, computing the corresponding optimal action and estimating its probability through empirical frequencies.

ARA and level-k thinking. Assessing the random distribution $\Pi_A(d)$ entails strategic thinking since D needs to understand how A will model her decision problem. This can lead to a hierarchy of decision-making problems as in Section 4.1. This analysis continues recursively, creating the hierarchy of nested decision models. Thus, D selects her action based upon a chain of reasoning of the form "I think that A thinks that I think..." which will go k levels deep, depending on how sophisticated she believes her opponent is. This is the level-k thinking approach in Stahl and Wilson (1995), who make it operational by dynamic programming. However it seems more natural to stop the iteration when no more information is reasonably available, and then use a non-informative prior for the random probabilities and utilities.

6 Further topics

We cover now additional important topics in ARA. First of all, we focus on cases in which other rationalities rather than maximum expected utility is assumed on behalf of the adversaries. In addition, we have just deal with interactions between two agents. Next, we discuss how the ARA methodology could be extended to multiagent setting. We also discuss more complex interactions between two agents than the ones introduced in Sections 4.1 and 4.2. We end the section discussing computational aspects of ARA.

6.1 Facing adversaries with other rationalities

Throughout the discussion, we have emphasised that all agents seek to maximize expected utility. This is reasonable for the supported agent, as we have her available for dialogue, but not so clear for the adversaries, who may be interested in hiding and concealing information. But it is possible to use prospect theory instead of utility functions, or other rationalities reflected in Section 3, and all the previous results follow in analogy. Another nice point is the analyst does not have to suppose her opponent is using a single solution concept. For example, she might believe her opponent has probability $1/3$ of seeking a BNE, $1/3$ of being a level-0 thinker, and $1/3$ of being a level-1 thinker. This creates a mixture model that expresses her full uncertainty, and ARA can be used to solve this more complex problem.

6.2 Facing more than one adversary

Consider a community of n agents. The i-th agent has a finite set $\mathcal{A}_i = \{a_i^1, \ldots, a_i^{m_i}\}$ of possible actions, $i = 1, \ldots, n$. The agents simultaneously implement their actions $a = (a_1, \ldots, a_n)$ and obtain their respective utilities $u_i(a)$, $i = 1, \ldots, n$, which they want to maximize. Assume one can build a distribution $\pi_1(a_2, \ldots, a_n) = \pi_1(a_{\text{-}1})$ that expresses the first agent's beliefs about the decisions made by the other agents. In ARA, the first agent should solve

$$a_1^* = \arg\max_{a_1 \in \mathcal{A}_1} \sum_{j=1}^{N_1} u_1(a_1, a_{\text{-}1}^j)\, \pi_1(a_{\text{-}1}^j),$$

where N_1 is the number of possible combinations of opponents' actions. The reviewed approaches to form this probabilities in the two agent case, can be extended to the multi-agent setting.

For instance, when performing fictitious play, as in the two agents case, agent one could use a Dirichlet-multinomial model where now the expected predictive distribution over all agents' actions excluding the first one would be

$$\pi_1(a_{\text{-}1}^k) = \frac{\alpha_k + x_k}{\sum_{j=1}^{N_1} (\alpha_j + x_j)}, \quad k = 1, \ldots, N_1$$

being α_k the Dirichlet prior parameter over the probability of $a_{\text{-}1}^k$ and x_k is the number of times that the other agents previously implemented such action.

The level-1 thinking approach can also be extended to the multi-agent system. This entails that agent 1 thinks about the problem that each of the other agents would solve. For the i-th agent, it is

$$a_i^* = \operatorname*{arg\,max}_{a_i \in \mathcal{A}_i} \sum_{j=1}^{N_i} u_i(a_i, a_{\text{-}i}^j)\, \pi_i(a_{\text{-}i}^j),$$

where N_i is the number of possible combinations of agent i opponents' actions and $\pi_i(a_{\text{-}i}^j)$ represents the beliefs of the i-th agent about the actions of the others. Since CK does not hold, the i-th agent does not know u_i nor π_i, but she can describe that uncertainty with a random utility U_i and a random probability Π_i, as previously discussed. This leads her to compute the random optimal action

$$A_i^* = \operatorname*{arg\,max}_{x \in \mathcal{A}_i} \sum_{j=1}^{N_i} U_i(a_i, a_{\text{-}i}^j)\, \Pi_i(a_{\text{-}i}^j).$$

Then, π_1 would be assessed using

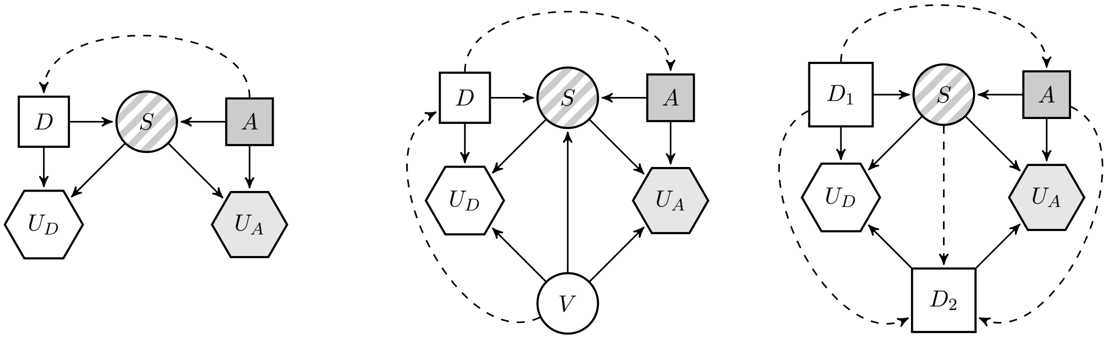
$$\pi_1(a_i^j) = \Pr(A_i^* = a_i^j), \quad j = 1, \ldots, m_i.$$

Extending further steps in the level-k thinking approach to the case of several opponents is theoretically straightforward, but comes at a higher computational cost. An example of the ARA framework deployed in reinforcement learning scenarios with multiple agents can be seen in Gallego et al. (2019b). To mitigate the computational cost entailed by the large size of the sets of N_1 and N_i above, we could assume conditional independence of the actions of opponents.

6.3 Facing more complex interactions

Previously, this paper introduced BAIDs for a simple simultaneous game and a simple sequential game. Figure 3a reflects a template for the sequential defend-attack game leading to the development of a contingency plan. But ARA can treat more challenging situations. Figure 3b generalizes the sequential defend-attack case of Figure 2 to situations in which the defender has private information V not shared with the attacker. Arc $V - D$ shows that information is known by D when she makes her decision; the lack of the arc $V - A$, shows that this information is not known by A when making his decision. The uncertainty about the outcome s depends on the actions by A and D, as well as on V.

Finally, Figure 3c depicts sequential defend-attack-defend games. In these, the defender moves first choosing d_1, then, the attacker chooses a after having observed d_1, and finally the defender, knowing the outcome s, her previous decision d_1 and the attacker's a, chooses her final defense (mitigation) d_2.



(a) Sequential attack-defend.

(b) Sequential defend-attack with private information.

(c) Sequential defend-attack-defend.

Figure 3: Bi-agent influence diagrams for ARA templates.

Beyond these game structures, there can be more general and complex interactions between agents. Such BAIDS can be analyzed using methods described in González-Ortega et al. (2019).

6.4 An ARA computational pipeline

Based on the BAID game templates, here is a general computational pipeline for ARA.

1. Modeling system threats. Represent the attacker's problem from the defender's perspective through an influence diagram (ID). The attacker's key features are his goals, knowledge and capabilities. Assessing these require determining the actions he may undertake and the utility that he perceives when performing them, given a defender's strategy. The output is the set of attacker's decision nodes, together with the value node and arcs indicating how his utility depends on his decisions and those of the defender. Assessing the attacker's knowledge entails looking for relevant information that he may have when performing the attack, and his degree of knowledge about this information, as we do not assume CK. This entails not only a modeling activity, but also a security assessment of the system to determine which of its elements are accessible to the attacker. The outputs are the uncertainty nodes of the attacker ID, the arcs connecting them and those ending in the decision nodes, indicating the information available to A when attacking. Finally, identifying his capabilities requires determining which part of the defender problem the attacker

has influence upon. This enables the attacker ID to connect with that of the defender.
2. Simulating attacks. Based on step 1, a mechanism is required to simulate reasonable attacks. The state of the art solution assumes that the attacker seeks an NE, given strong CK hypothesis. The ARA methodology relaxes such assumptions, through a procedure to simulate adversarial decisions. Starting with the adversary model in step 1, the uncertainty about his probabilities and utilities is propagated to his problem and leads to the corresponding random optimal adversarial decisions which provide the required simulation.
3. Adopting defenses. Augment the defender's problem by incorporating the attacker's problem produced in step 1. As output, generate a BAID reflecting the dual confrontation. Finally, solve the defender's problem by maximizing her subjective expected utility, integrating out all attacker decisions, which are random from the defender's perspective given the lack of CK. In general, the corresponding integrals are approximated through MC, simulating attacks consistent with our knowledge about the attacker using the mechanism of step 2.

Thus, from a computational point of view, given a defense policy, one simulates from the attacker's problem to forecast the attacks and that enables the defender problem to find her optimal policies. One can improve this general approach by taking advantage of the dynamic and informational structure of the BAID underlying the problem. One alternates between simulation and optimization as described in González-Ortega et al. (2019). But one still needs to simulate and optimize which can be computationally intensive. One possibility to alleviate this is to jointly the simulate and the optimize with the aid of the augmented simulation approach Ekin et al. (2019).

7 Applications

Section 2 introduced basic concepts in ARA through an auction example. Further details may be seen in Banks et al. (2015). Beyond that, applications in other areas abound. Many traditional applications of game theory in which CK conditions are debatable could be revisited from the ARA perspective. Important ARA applications are in defense, counterterrorism, security, cybersecurity risk analysis.

Regarding defense, Wang and Banks (2011) applies ARA to the problem of selecting a route through a network where an opponent chooses vertices for ambush. The methodology in that paper could be applied to convoy routing problems when there may be improvised explosive devices and imperfect information about their locations. Sevillano et al. (2012) uses ARA to support the owner of a ship in managing piracy risk; it models the situation using a defend-attack-defend game, in which CK is not assumed. Roponen and Salo (2016) considers ARA to enhance combat models and Roponen et al. (2020) uses it to protect from unmanned aerial vehicles.

The area of counterterrorism has also seen relevant applications of ARA. Rios and Insua (2012) provide general ARA models including defend-attack models, sequential defend-attack-defend models, and sequential defend-attack models with private information. These may be used as building blocks for more specific risk analysis of counterterrorism problems. For instance, Gil and Parra-Arnau (2019) studies counter-terrorist online surveillance from an ARA perspective, focusing on the problem of monitoring a set of websites through classification of profiles suspected of carrying out terrorist attacks. Rios Insua et al. (2016) deals with critical infrastructure protection studying security resource allocation decision processes for an organization which faces multiple threats over multiple sites, specifically using a railways network.

Regarding urban security, applications in resource allocation have attracted interest. For example, Gil et al. (2016) uses ARA to allocate security resources to protect urban spaces, taking their spatial structure into account. In cybersecurity, Rios Insua et al. (2019) provides a comprehensive framework for cybersecurity risk analysis, covering the presence of both intentional and non-intentional threats and the use of insurance as part of the security portfolio. Naveiro et al. (2019) applies ARA to the emerging field of adversarial machine learning. In particular, it shows how to protect statistical classification systems from attackers trying to fool them by intentionally modifying input data.

Finally, it is worth mentioning ARA applications in social robotics, such as the one in Esteban and Insua (2014). They illustrate how to use ARA to support the decision making of an autonomous agent that interacts with other agents and people in a competitive environment

8 Conclusions

This has been a brief overview on adversarial risk analysis, its key concepts, methods and applications. But there remain many open topics.

First, a promising line for future research is to extend ARA methodology to deal with repeated play. The structure of these problems allows learning from past adversarial actions via Bayesian updating of the defender's beliefs about her opponent's unknowns. These models could be especially useful in games in which actions are taken in continuous time, also known as differential games (Dockner et al., 2000). In addition, interest in ARA in multi-agent reinforcement learning has recently risen (Gallego et al., 2019a). Research in multi-agent reinforcement learning has mainly focused in modeling the whole system through Markov games. However, the problem of supporting a single agent facing one or more opponents in a reinforcement learning setting is largely unexplored.

As we have seen, ARA relaxes crucial assumptions such as CK, thereby increasing realism. However, this comes at involved computations. Research on efficient approaches for ARA is crucial. For instance, exploring gradient-based techniques for sequential games is a fruitful line of research. Naveiro and Insua (2019) provides an efficient solution method for sequential defend-attack games under the game-theoretic paradigm which could be extended to deal with the ARA setting. In addition, ARA entails simulating from the attacker's problem to forecast attacks and then optimizes for the defender to find her decision. This two-stage computation is demanding and single stage methods could reduce computation. Initial ideas based on augmented probability simulation are in Ekin et al. (2019).

Finally, much research can be conducted in the applied side. For instance, regarding applications of ARA to adversarial machine learning problems, Naveiro et al. (2019) aims at robustifying classification systems against adversarial threats. Extensions of such techniques to regression or time series problems would be interesting. A broad overview on adversarial machine learning is in Ríos Insua et al. (2019). Malware detection, fake news detection and autonomous vehicles security are just a few examples of important societal applications of AML.

Research Resources

This work was partially supported by the National Science Foundation under Grant DMS-1638521 to the Statistical and Applied Mathematical Sciences Institute and a BBVA Foundation project. Roi Naveiro also acknowledges support of the Spanish Ministry for his grant FPU15-03636. V.G. acknowledges also acknowledges support of the Spanish Ministry for his grant FPU16-05034. David Rios Insua is grateful to the MTM2017-86875-C3-1-R AEI/ FEDER EU project, the AXA-ICMAT Chair in Adversarial Risk Analysis and the EU's Horizon 2020 project 815003 Trustonomy.

Acknowledgments

We are grateful to discussions with J. Rios, A. Salo, J. Roponen, C. Joshi, F. Ruggeri, T. Ekin, P. Esteban, C. Gil, S. Liu and J. Ortega.

References

A., O., C., B., A., D., J., E., P., G., J., J., J., O., and T., R. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. Wiley.

Antos, D. and Pfeffer, A. (2010). Representing bayesian games without a common prior. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1*, AAMAS '10, pages 1457–1458.

Au, T. C. (2014). *Topics in Computational Advertising*. PhD thesis, Duke University.

Banks, D. L., Aliaga, J. M. R., and Insua, D. R. (2015). *Adversarial Risk Analysis*. Chapman and Hall/CRC.

Bedford, T., Cooke, R., et al. (2001). *Probabilistic risk analysis: foundations and methods*. Cambridge University Press.

Brams, S. J. (2011). *Game theory and politics*. Courier Corporation.

Brown, G., Carlyle, M., Salmerón, J., and Wood, K. (2006). Defending critical infrastructure. *Interfaces*, 36(6):530–544.

Brown, G. W. (1951). Iterative solution of games by fictitious play. *Activity Analysis of Production and Allocation*, pages 374–376.

Camerer, C. F. (2003). Strategizing in the brain. *Science*, 300(5626):1673–1675.

Chung, K. L. (2001). *A Course in Probability Theory*. Ac Press.

Cox Jr, L. A. (2013). Improving causal inferences in risk analysis. *Risk analysis*, 33(10):1762–1771.

Dockner, E. J., Jorgensen, S., Long, N. V., and Sorger, G. (2000). *Differential Games in Economics and Management Science*. Cambridge University Press.

Ekin, T., Naveiro, R., Torres-Barrán, A., and Ríos-Insua, D. (2019). Augmented probability simulation methods for non-cooperative games. *arXiv preprint arXiv:1910.04574*.

Esteban, P., Liu, S., Rios Insua, D., and Ortega, J. (2020). Competition and cooperation in a community of autonomous agents. *Autonomous Robots*, 44:533–546.

Esteban, P. G. and Insua, D. R. (2014). Supporting an autonomous social agent within a competitive environment. *Cybernetics and Systems*, 45(3):241–253.

Fibich, G. and Gavish, N. (2011). Numerical simulations of asymmetric first-price auctions. *Games and Economic Behavior*, 73(2):479–495.

French, S. and Rios Insua, D. (2000). *Statistical Decision Theory*. Wiley.

Gallego, V., Naveiro, R., and Insua, D. R. (2019a). Reinforcement learning under threats. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9939–9940.

Gallego, V., Naveiro, R., Insua, D. R., and Oteiza, D. G.-U. (2019b). Opponent aware reinforcement learning. *arXiv preprint arXiv:1908.08773*.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2020). *Bayesian Data Analysis*. CRC press.

Gibbons, R. (1992). *A Primer in Game Theory*. Harvester Wheatsheaf.

Gil, C. and Parra-Arnau, J. (2019). An adversarial-risk-analysis approach to counterterrorist online surveillance. *Sensors*, 19(3).

Gil, C., Rios Insua, D., and Rios, J. (2016). Adversarial risk analysis for urban security resource allocation. *Risk Analysis*, 36(4):727–741.

Gintis, H. (2009). *The Bounds of Reason: Game Theory and the Unification of Behavioural Sciences.* Princeton University Press.

González-Ortega, J., Insua, D. R., and Cano, J. (2019). Adversarial risk analysis for bi-agent influence diagrams: An algorithmic approach. *European Journal of Operational Research*, 273(3):1085–1096.

Halpern, J. Y. (2008). Beyond nash equilibrium: Solution concepts for the 21st century. In *Proceedings of the twenty-seventh ACM symposium on Principles of distributed computing*, pages 1–10.

Hammerstein, P. and Selten, R. (1994). Game theory and evolutionary biology. *Handbook of game theory with economic applications*, 2:929–993.

Hargreaves-Heap, S. and Varoufakis, Y. (2004). *Game Theory: A Critical Introduction*. Taylor & Francis.

Harsanyi, J. (1982). Comment—subjective probability and the theory of games: Comments on kadane and larkey's paper. *Management Science*, 28(2):120–124.

Harsanyi, J. C. (1967). Games with incomplete information played by "Bayesian" players, I–III Part I. The basic model. *Management Science*, 14(3):159–182.

Hubbard, T. P., Kirkegaard, R., and Paarsch, H. J. (2013). Using economic theory to guide numerical analysis: Solving for equilibria in models of asymmetric first-price auctions. *Computational Economics*, 42(2):241–266.

Kadane, J. B. and Larkey, P. D. (1982). Subjective probability and the theory of games. *Management Science*, 28:113–120.

Keeney, R. (2007). Modeling values for anti-terrorism analysis. *Risk Analysis*, 27(3):585–596.

Lebrun, B. (1999). First-price auction in the asymmetric *N* bidder case. *International Economic Review*, 40(1):125–142.

Levy, H. (1998). *Stochastic Dominance: Investment Decision Making under Uncertainty*. Springer.

Menache, I. and Ozdaglar, A. (2011). Network games: Theory, models, and dynamics. *Synthesis Lectures on Communication Networks*, 4(1):1–159.

Merrick, J. and Parnell, G. S. (2011). A comparative analysis of pra and intelligent adversary methods for counterterrorism risk management. *Risk Analysis: An International Journal*, 31(9):1488–1510.

Milgrom, P. R. and Weber, R. J. (1982). A theory of auctions and competitive bidding. *Econometrica: Journal of the Econometric Society*, pages 1089–1122.

Myerson, R. B. (1997). *Game Theory: Analysis of Conflict*. Harvard University Press.

Naveiro, R. and Insua, D. R. (2019). Gradient methods for solving stackelberg games. In *International Conference on Algorithmic DecisionTheory*, pages 126–140. Springer.

Naveiro, R., Redondo, A., Insua, D. R., and Ruggeri, F. (2019). Adversarial classification: An adversarial risk analysis approach. *International Journal of Approximate Reasoning*, 113:133–148.

Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V. (2007). *Algorithmic Theory*. Cambridge Univeristy Press.

Osborne, M. J. and Rubinstein, A. (1998). Games with procedurally rational players. *American Economic Review*, pages 834–847.

Raiffa, H. (1982). *The Art and Science of Negotiation*. Harvard University Press.

Raiffa, H., Richardson, J., Metcalfe, D., et al. (2002). *Negotiation analysis: The science and art of collaborative decision making*. Harvard University Press.

Rios, J. and Insua, D. R. (2012). Adversarial risk analysis for counterterrorism modeling. *Risk Analysis: An International Journal*, 32(5):894–915.

Rios, J. and Rios Insua, D. (2012). Adevrsarial risk analysis for counterterrorism modeling. *Risk Analysis*, 32(5):894–915.

Ríos Insua, D., Banks, D., Ríos, J., and Ortega, J. (2019). *Adversarial Risk Analysis as an Expert Judgement Methodology*. Springer International Publishing.

Rios Insua, D., Cano, J., Pellot, M., and Ortega, R. (2016). Multi-threat multi-site protection: a security case study. *European Journal Operaional Research*, 252(3):888–899.

Rios Insua, D., Couce-Vieira, A., Rubio, J. A., Pieters, W., Labunets, K., and G. Rasines, D. (2019). An adversarial risk analysis framework for cybersecurity. *Risk Analysis*.

Ríos Insua, D., Naveiro, R., Gallego, V., and Poulos, J. (2019). Adversarial Machine Learning: Perspectives from adversarial risk analysis. *arXiv e-prints*, page arXiv:1908.06901.

Roponen, J., Rios Insua, D., and Salo, A. (2020). Adversarial risk analysis under partial information. *European Journal of Operational Research*.

Roponen, J. and Salo, A. (2016). Adversarial risk analysis for enhancing combat simulation models. *Journal of Military Studies*, 6, 2:1–22.

Russell, C. (1998). *Coordination Games*. Cambridge University Press.

Samuelson, L. (2016). Game theory in economics and beyond. *Journal of Economic Perspectives*, 30(4):107–30.

Sevillano, J. C., Insua, D. R., and Rios, J. (2012). Adversarial Risk Analysis: The Somali Pirates Case. *Decision Analysis*, 9(2):86–95.

Shiva, S., Roy, S., and Dasgupta, D. (2010). Game theory for cyber security. In *Proceedings of the Sixth Annual Workshop on Cyber Security and Information Intelligence Research*, pages 1–4.

Shubik, M. (1982). *Game theory in the social sciences: Concepts and solutions*.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, 69(1):99–118.

Stahl, D. O. and Wilson, P. W. (1994). Experimental evidence on players' models of other players. *Journal of economic behavior & organization*, 25(3):309–327.

Stahl, D. O. and Wilson, P. W. (1995). On players models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1):218–254.

Tversky, A. and Kahneman, D. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291.

Wang, S. and Banks, D. (2011). Network routing for insurgency: An adversarial risk analysis framework. *Naval Research Logistics (NRL)*, 58(6):595–607.

Young, H. (2004). *Strategic Learning and its Limits*. Oxford UP, first edition.