Adversarial Risk Analysis as a Decomposition Method for Structured Expert Judgement Modelling

David Ríos Insua[1], David Banks[2], Jesús Ríos[3], Jorge González-Ortega[1]

[1] Instituto de Ciencias Matemáticas (CSIC-UAM-UC3M-UCM), Madrid, Spain
[2] Department of Statistical Science (Duke University), Durham, NC
[3] IBM Research Division (IBM), Yorktown Heights, NY

Abstract

We argue that adversarial risk analysis may be incorporated into the structured expert judgement modelling toolkit for cases in which we need to forecast the actions of competitors based on expert knowledge. This is relevant in areas such as cybersecurity, security, defence and business competition. As a consequence, we present a structured approach to facilitate the elicitation of probabilities over the actions of other intelligent agents by decomposing them into multiple, but simpler, assessments later combined together using a rationality model of the adversary to produce a final probabilistic forecast. We then illustrate key concepts and modelling strategies of this approach to support its implementation.

Keywords: Structured expert judgement, adversarial risk analysis, decomposition, security, cybersecurity.

1 Introduction

Structured Expert Judgement (SEJ) elicitation has a long history of successes, both in methodology and applications, many of them stemming from Roger Cooke's work, e.g. Cooke (1991) and Goossens et al. (1998). Hence, it has become a major ingredient within risk and decision analysis (Bedford and Cooke, 2011). A significant feature in the practice of these disciplines, as already acknowledged in the classic book by Raiffa (1968), is the emphasis in decomposing complex problems into smaller pieces that are easier to understand and recombining the piecewise solutions to tackle the global problem.

In particular, belief assessment benefits from decomposition, typically through the argument of *extending the conversation*. Direct elicitation of probabilities can be a very difficult cognitive task. For example, there may be many factors influencing the occurrence of an outcome of interest whose effects experts would have to identify and balance in their heads to produce a probability judgement. Thus, rather than directly assessing this probability (with a standard SEJ technique), one could find a conditioning partition and estimate the probabilities of the outcome given the corresponding events. From these, and the probabilities of the conditioning events, the law of total probability enables calculation of the unconditional probability of the outcome. Ravinder et al. (1988) and Andradottir and Bier (1997, 1998) provide a methodological framework to validate the advantages of this approach, empirically tested in e.g. MacGregor and Kleinmuntz (1994) and MacGregor (2001). Tetlock and Gardner (2015) call this approach *Fermitisation* and present it as a key strategy for the success of their super-forecasters, and SEJ at large. Decompositions uncover the complexity underlying a direct probability assessment, eliminating the burden on experts to perform sophisticated modelling in their heads. This simplifies complex cognitive tasks, reveals assumptions experts make in their judgements and mitigate their reliance on heuristics that can introduce bias, ensuring that they actually analyse the relevant problem (Montibeller and von Winterfeldt, 2015). Decompositions typically entail more assessments, though these tend to be simpler and more meaningful, leading to improved judgements and decisions. In turn, this would allow for better harnessing expert knowledge e.g. by assigning the proper expertise to the different sub-tasks of an assessment.

In many settings, specially in contexts such as security, counterterrorism or cybersecurity, experts will have to face adversarial problems in the sense that they need to deal with probabilities referring to actions carried out by opponents. As an example, in Chen et al. (2016) nearly 30% of the questions posed to experts somehow involved adversaries (e.g. *Will Syria use chemical or biological weapons before January 2013?*). Though we could think of using the standard SEJ tools as illustrated in other chapters in this volume, we present Adversarial Risk Analysis (ARA) as a decomposition strategy to support SEJ when forecasting adversarial actions. Regardless of the many issues associated with how an expert can translate domain knowledge into a probability, there is always the problem of how to best structure the elicitation process to get to a probability. When this is too difficult to assess but can be expressed as a combination of other simpler probabilities, decomposition becomes a critical part of the SEJ procedure. Our focus is on how ARA, as a structured SEJ technique, determines what the right questions to ask are and how experts' answers to these questions are combined to produce an adversarial probabilistic forecast.

After sketching the ARA approach to decomposition (Section 2), we show how this can actually improve expert assessment of opponent actions (Section 3). We then propose several ways to implement ARA in practice (Section 4), include a numerical example (Section 5), and end with a discussion (Section 6).

2 ARA as a SEJ Decomposition Method

ARA was originally introduced to deal with game theoretic problems studied from a Bayesian perspective, (Ríos Insua et al., 2009; Banks et al., 2015). It stems from the observation that common knowledge assumptions in standard game theoretic approaches based on Nash equilibria and their refinements do not hold in many applications, such as counterterrorism or cybersecurity, as competitors try to conceal information. Games are formulated in a Bayesian manner, as in Kadane and Larkey (1982) and Raiffa (2003), and operationalised by providing procedures to forecast the actions of the adversary.

To simplify the discussion, we consider the basic ARA approach through a sequential Defend-Attack game: agent D (she, defender) first makes her decision $d \in \mathcal{D}$, then agent A (he, attacker) observes d and chooses his alternative $a \in \mathcal{A}$. The outcome s of their interaction is a random variable S whose distribution depends upon d and a. As an example, imagine that a company deploys cybersecurity controls and then, having observed them, a cybercriminal decides whether to launch a cyber attack. The cost to the company would be a random variable that is conditioned upon both decisions (the controls deployed and the attack launched). The problem that agent D faces is depicted in the influence diagram in Figure 1.

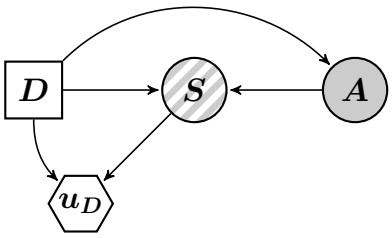

Figure 1: The decision problem as seen by D.

To solve it, she requires $p_D(s \mid d, a)$, which reflects her beliefs on the outcome given both agents' actions, and her utility function $u_D(d, s)$, modelling her preferences and risk attitudes over the consequences, which we assume depends on the outcome and the defence implemented. Besides, she needs the distribution $p_D(a \mid d)$, which is her assessment of the probability that A will choose action a after having observed her choice d. Once D has completed these judgements, she can compute the expected utility of

decision d as

$$\psi_D(d) = \int \left[\int u_D(d,s)\, p_D(s \mid d,a)\, \mathrm{d}s \right] p_D(a \mid d)\, \mathrm{d}a,$$

and seek for the optimal decision $d^* = \arg\max_{d \in \mathcal{D}} \psi_D(d)$.

This is a standard risk or decision analysis exercise except for the elicitation of $p_D(a \mid d)$, which entails strategic aspects. D could try to assess it from a standard belief elicitation perspective, as in Cooke (1991) or O'Hagan et al. (2006), but ARA usefully suggests a decomposition approach to such assessment that requires her to analyse the problem from A's perspective, as shown in the influence diagram in Figure 2.

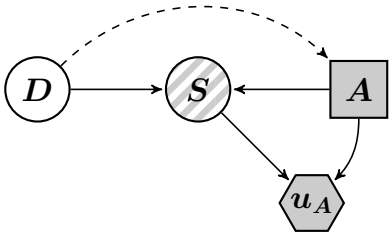

Figure 2: D's analysis of the decision problem as seen by A.

Thus, D puts herself in A's shoes. She would use all the information she can obtain about A's probabilities $p_A(s \mid d,a)$ and utilities $u_D(d,s)$, assuming he is an expected utility maximiser. Then, instead of using point estimates for p_A and u_A to find A's optimal response for a given d, her uncertainty about A's decision would derive from her uncertainty about (p_A, u_A), through a distribution F on the space of probabilities and utilities. This induces a distribution over A's expected utility, which for each d and a is

$$\Psi_A(d,a) = \int U_A(a,s)\, P_A(s \mid d,a)\, \mathrm{d}s,$$

where (P_A, U_A) follow the distribution of F. Then, D finds the required $p_D(a \mid d)$ as $\mathbb{P}_F\left[a = \arg\max_{x \in \mathcal{A}} \Psi_A(d,x)\right]$, in the discrete case and, analogously, in the continuous one. She could use Monte Carlo simulation to approximate $p_D(a \mid d)$, as shown in Sections 3 and 5.

Observe that the ARA approach weakens the standard, but unrealistic, common knowledge assumptions in game theoretic approaches (Hargreaves-Heap and Varoufakis, 2004), according to which the agents share information about their probabilities and utilities. In our case, not having common knowledge means that D does not know (p_A, u_A), and thus we model such uncertainty through F. The approach extends to simultaneous decision making problems, general interactions between both agents, multiple agents, agents who employ principles different than maximum expected utility, as well as to other contexts presented in Banks et al. (2015). Here we exclusively explore the relevance of ARA as part of the SEJ toolkit.

3 Assessing ARA decompositions

We hereafter study ARA as a decomposition approach through the sequential Defend-Attack model described above, comparing direct SEJ and the ARA decomposition.

3.1 Framework

As mentioned, there are two possible ways to assess the distribution $p_D(a \mid d)$:

- One could do it directly with standard SEJ procedures (Cooke, 1991). Denote such assessment by $p_D^{SEJ}(a \mid d)$.
- Otherwise, one could determine it indirectly through ARA as in Section 2. D would model her uncertainty about A's beliefs and preferences, represented by $(P_A, U_A) \sim F$, and then solve A's decision making problem using these random probabilities and utilities to estimate

$$p_D^{ARA}(a \mid d) = \mathbb{P}_F \left[a = \arg\max_{x \in \mathcal{A}} \int U_A(x, s)\, P_A(s \mid d, x)\, \mathrm{d}s \right].$$

To compare both approaches, we make three simplifying assumptions: (i) D has only two options, defend (d_1) or not (d_0); (ii) A can solely choose between attacking (a_1) or not (a_0); and (iii) if A decides to attack, the only two outcomes are success (s_1) or failure (s_0). For A, the problem can be viewed as the decision tree in Figure 3, with $d \in \{d_0, d_1\}$, which parallels the influence diagram in Figure 2. The ARA approach obtains the required conditional probabilities $p_D^{ARA}(a \mid d)$ by solving the decision tree using D's (random) assessments over A's inputs.

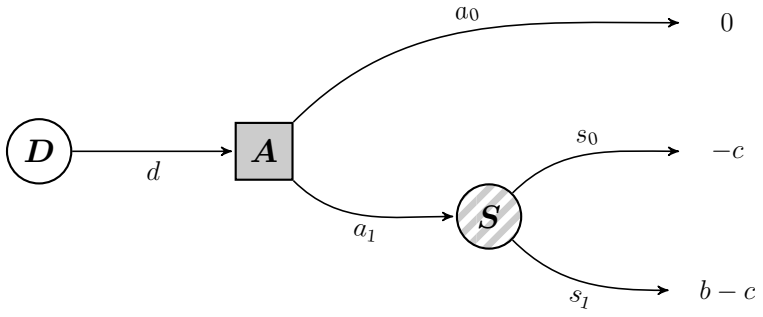

Figure 3: Decision tree representing A's problem. c represents the cost of implementing an attack; b, the benefit of a successful attack.

Suppose D thinks A bases his decision on a cost-benefit analysis. In that case, the consequences for A are described in Table 1. For this, D might use a multi-attribute value

model to decompose her judgement about A's valuation of consequences into simpler assessments regarding such costs and benefits. Later, she can aggregate these estimates as shown in the row *Profit* in Table 1, reflected in Figure 3.

	(Attack, Outcome) - (a, s)		
	(a_0, s_0)	(a_1, s_0)	(a_1, s_1)
Cost	0	c	c
Benefit	0	0	b
Profit	0	$-c$	$b - c$

Table 1: Cost-benefit analysis of A's consequences.

This requires D to assess two quantities: c and b, A's cost of undertaking an attack and his benefit if successful, respectively. We assume that $0 < c < b$, implying that attacking is more costly for A than not attacking, but potentially more beneficial; and that a successful attack is better for A than an unsuccessful one. Since D is generally uncertain about these quantities, she will provide probability distributions to model her beliefs about them. Suppose her self-elicitations lead to the uniform distributions

- A's cost of an attack: $c \sim \mathcal{U}(c_{\min}, c_{\max})$.
- A's benefit from a successful attack: $b \sim \mathcal{U}(b_{\min}, b_{\max})$.

These allow D to estimate the random values related to A's consequences in Table 1. We have assumed that D believes that A's costs and benefits are uniformly distributed and, very importantly, independent. However, in many cases there is dependence; e.g. a more costly attack is most likely correlated with larger benefits for A. In that case, one needs to model c and b jointly. For simplicity, this discussion assumes independence.

If D believes that A is risk neutral (i.e. seeking to maximise expected profit), she would now elicit her beliefs about A's impression on his probability of success. Otherwise, beforehand, she would have to model A's risk attitudes. She could do that by eliciting a utility function over profits for him and model his risk attitude as shown in Section 4.2 and exemplified in Section 5, where her uncertainty about the attacker risk attitude is captured through a probability distribution over the risk aversion coefficient of a parametric utility function. Alternatively, because there are just three possible outcomes for A (no attack, failed attack, successful attack), D may directly assess her belief about his utility for each of them. Without loss of generality, utilities of 0 and 1 can be respectively assigned to the worst and best consequences for A. Since D believes that

$-c < 0 < b - c$, $u_A(-c) = 0$ and $u_A(b - c) = 1$, even if she does not know the exact values of b and c. Thus, she just needs to elicit her distribution for $u_A(0) = u$, knowing that $0 < u < 1$, though being uncertain of A's exact value of u. Recall that this could be elicited as the probability at which A is indifferent between getting profit 0 for sure and a lottery ticket in which he wins $b - c$ with probability u and looses c with probability $1 - u$. This way, D could elicit a distribution for the random variable U_A that represents her full uncertainty over A's utility u.

Having done this, D would also need to assess A's beliefs about his chance of success, determined by $p_A(s_1 \,|\, d_0, a_1) = \pi_{d_0}$ and $p_A(s_1 \,|\, d_1, a_1) = \pi_{d_1}$. She should model her uncertainty about these with random probabilities $\pi_{d_0} \sim P_A^{d_0}$ and $\pi_{d_1} \sim P_A^{d_1}$, with $\pi_{d_1} < \pi_{d_0}$ to ensure that defending (d_1) reduces the chance of a successful attack. Then, based on the above assessments, for each $d \in \{d_0, d_1\}$, D can compute A's random expected utilities as

$$\Psi(d, a_0) = u_A(0) = u \sim U_A,$$

$$\Psi(d, a_1) = u_A(b - c) \times p_A(s_1 \,|\, d, a_1) + u_A(0) \times p_A(s_0 \,|\, d, a_1) = \pi_d \sim P_A^d,$$

and the ARA probabilities of attack, given the implemented defence, through

$$p_D^{ARA}(a_1 \,|\, d) = \mathbb{P}_{(U_A, P_A^d)}\left(u < \pi_d\right). \tag{1}$$

These probabilities represent the defender's ARA probabilistic predictions of how A will respond to each of her possible choices. As an example, suppose that we assess these distributions as $U_A \sim \mathcal{B}e(1, 2)$ (beta) and $P_A^{d_0} \sim \mathcal{U}(0.5, 1)$ and $P_A^{d_1} \sim \mathcal{U}(0.1, 0.4)$. Then, using Monte Carlo (MC) simulation we estimate the attack probabilities as $\hat{p}_D^{ARA}(a_1 \,|\, d_0) \approx 0.92$ and $\hat{p}_D^{ARA}(a_1 \,|\, d_1) \approx 0.43$ (based on an MC sample size of 10^6). In this case, choosing to defend (d_1) acts as a deterrent for A to attack (a_1).

3.2 Comparison

We now address whether this ARA decomposition approach leads to improved attack probability estimates over those obtained by direct SEJ methods. Adopting a normative viewpoint, we show through simulation that under certain conditions the variance of the ARA estimates are smaller than those of the SEJ estimates.

In our case, due to the assumptions behind expression (1), we have no reason to believe that D finds one attack distribution more (or less) likely than another, except that an attack is more likely when no defence is attempted. That is, $p_{d_0}^{SEJ} \geq p_{d_1}^{SEJ}$ where $p_{d_i}^{SEJ} = p_D^{SEJ}(a_1 \,|\, d_i)$, $i = 1, 2$. Thus, as a high-entropy benchmark, we assume that $p_{d_0}^{SEJ}$, $p_{d_1}^{SEJ}$ are uniformly distributed over the set $\{0 \leq p_{d_1}^{SEJ} \leq p_{d_0}^{SEJ} \leq 1\}$, whose

variance-covariance matrix is analytically computed as

$$\begin{pmatrix} \frac{1}{18} & \frac{1}{36} \\ \frac{1}{36} & \frac{1}{18} \end{pmatrix} \approx \begin{pmatrix} 5.56 & 2.78 \\ 2.78 & 5.56 \end{pmatrix} \cdot 10^{-2}. \tag{2}$$

In turn, D's assessment of the ARA attack probabilities involves eliciting distributions $(U_A, P_A^{d_0}, P_A^{d_1})$. It is reasonable to assume that u is independent of π_{d_0} and π_{d_1}. Since the support of all three random variables is $[0, 1]$, an equitable framework for the benchmark may assume that $U_A \sim \mathcal{U}(0, 1)$ and $(P_A^{d_0}, P_A^{d_1})$ are uniformly distributed over the set $\{0 \leq \pi_{d_1} \leq \pi_{d_0} \leq 1\}$. We computed 10^4 MC estimates of the attack probabilities using these distributions, each based on an MC sample size of 10^4, leading to a variance-covariance matrix for $p_{d_0}^{ARA}$ and $p_{d_1}^{ARA}$ of

$$\begin{pmatrix} 2.24 & 1.10 \\ 1.10 & 2.22 \end{pmatrix} \cdot 10^{-5}. \tag{3}$$

Thus, as a result of the decomposition approach inherent to the ARA methodology, both variances and the covariance in the ARA approach (3) are significantly smaller than those in the benchmark (2), providing a more precise assessment.

Yet, typically, one would have more information about $(U_A, P_A^{d_0}, P_A^{d_1})$. For example, suppose D believes that the mean values of the three random variables are $E[U_A] = \frac{2}{5}$, $E[P_A^{d_0}] = \frac{2}{3}$ and $E[P_A^{d_1}] = \frac{1}{3}$. If she assumes they all are uniformly distributed with maximum variance, then $U_A \sim \mathcal{U}(0, \frac{4}{5})$, $P_A^{d_0} \sim \mathcal{U}(\frac{1}{3}, 1)$ and $P_A^{d_1} \sim \mathcal{U}(0, \frac{2}{3})$ (with $\pi_{d_1} \leq \pi_{d_0}$). In this case, the estimated variance-covariance matrix for $p_{d_0}^{ARA}$ and $p_{d_1}^{ARA}$ is

$$\begin{pmatrix} 1.42 & 0.65 \\ 0.65 & 2.35 \end{pmatrix} \cdot 10^{-5}.$$

Compared to (3), these assumptions reduce the variance for $p_{d_0}^{ARA}$ and the covariance, although slightly increase the variance of $p_{d_1}^{ARA}$. Finally, if the random variables followed beta distributions with common variance $\frac{1}{10}$, then $U_A \sim \mathcal{B}e(0.56, 0.84)$, $P_A^{d_0} \sim \mathcal{B}e(0.81, 0.41)$ and $P_A^{d_1} \sim \mathcal{B}e(0.41, 0.81)$ (and $\pi_{d_1} \leq \pi_{d_0}$), and the variance-covariance matrix for $p_{d_0}^{ARA}$ and $p_{d_1}^{ARA}$ is

$$\begin{pmatrix} 1.52 & 0.64 \\ 0.64 & 2.25 \end{pmatrix} \cdot 10^{-5}.$$

Again, the covariance matrix is significantly more precise than the benchmark.

For further insights, assume that the direct elicitation process incorporates additional information, so that $p_{d_0}^{SEJ}$ and $p_{d_1}^{SEJ}$ are now uniformly distributed over the set $\{\varepsilon \leq p_{d_1}^{SEJ} \leq p_{d_0}^{SEJ} \leq 1 - \varepsilon\}$, requiring $0 \leq \varepsilon \leq \frac{1}{2}$ to be defined. Then, the variance-covariance

matrix for $p_{d_0}^{SEJ}$ and $p_{d_1}^{SEJ}$ is

$$\begin{pmatrix} \frac{(1-2\varepsilon)^2}{18} & \frac{(1-2\varepsilon)^2}{36} \\ \frac{(1-2\varepsilon)^2}{36} & \frac{(1-2\varepsilon)^2}{18} \end{pmatrix}. \tag{4}$$

From (3) and (4), we see that one must take $\varepsilon > 0.49$, a very precise assessment, so that the corresponding variance-covariance matrix of $p_{d_0}^{SEJ}$ and $p_{d_1}^{SEJ}$ becomes less variable than $p_{d_0}^{ARA}$ and $p_{d_1}^{ARA}$.

All these comparisons indicate that although the ARA approach requires more assessments to obtain the relevant probabilities of the adversarial actions, ARA tends to provide more precise estimates. However, if the direct information is very precise, then direct elicitation can outperform ARA in terms of reduced variance for the relevant probabilities.

4 ARA Modelling Strategies

We have shown that the ARA decomposition can have advantages over the plain SEJ approach. Consequently, it is worth describing how to implement it. We thus present a catalogue of strategies to model the random probabilities and utilities necessary to put ARA into practice.

4.1 Random probabilities

We focus first on D's assessments over A's perspective of the different random events involved in the problem, that is, the random probabilities. To fix ideas, assume we have a single chance node S which depends on both D's and A's choices. Our task is to develop a (random) distribution $P_A(s \mid d, a)$ that reflects D's uncertainty about A's prospect of S. We distinguish three cases. In all of them, as shown in Section 4.1.1, Bayesian updating could be used to dynamically adjust the assessed priors as data accumulates, thus attaining subsequent random posterior distributions that better reflect D's information and perspective over A's uncertainty.

4.1.1 Probability of a single event

Suppose first that the chance node S consists of a single event which may ($s = 1$) or not ($s = 0$) happen. Then, $p_A(s \mid d, a)$ is completely determined by $p_A(s = 1 \mid d, a)$, for each pair (d, a), as $p_A(s = 0 \mid d, a) = 1 - p_A(s = 1 \mid d, a)$.

One possibility would be to base $P_A(s = 1 \mid d, a)$ on an estimate π_D of $p_A(s = 1 \mid d, a)$, with some uncertainty around it. This may be accomplished in several ways. We could do it through a uniform distribution $\mathcal{U}(\pi_D - \mu, \pi_D + \mu)$ centred around π_D in which the

parameter μ would have to be assessed also. For example, if we get that the expected variance of the distribution is ν, we get $\mu = \sqrt{3\nu}$. Another option would be to use a beta distribution $\mathcal{B}e(\alpha, \beta)$ in which π_D may be regarded as the mean (or the median or the mode) of the distribution and we would have to assess the parameters α and β to shape the distribution, e.g. based on a further assessment of the variance ν. This would lead, when π_D is the mean, to

$$\alpha = \frac{\pi_D}{\nu}\left(\pi_D\left(1-\pi_D\right)-\nu\right), \qquad \beta = \frac{1-\pi_D}{\nu}\left(\pi_D\left(1-\pi_D\right)-\nu\right)$$

Note that when D thinks that A has information similar to hers, an adequate best guess for π_D could be based on her own assessment $p_D(s = 1 \,|\, d, a)$.

If the possible occurrence of event s were to be repeated over time, random prior distributions could be reassessed by means of Bayesian updating. Consider, for example, the second case in which a beta distribution $\mathcal{B}e(\alpha, \beta)$ is used. If event s has had y opportunities to happen and materialises only z of them, our random posterior would be $\mathcal{B}e(\alpha + z, \beta + y - z)$.

4.1.2 Probabilities of multiple events

We assume now that the chance node S includes N events $\{s_1, \ldots, s_N\}$. In this case, probabilities $p_A(s = s_1 \,|\, d, a), \ldots, p_A(s = s_{N-1} \,|\, d, a)$ determine $p_A(s \,|\, d, a)$ completely, for each pair (d, a), as $p_A(s = s_N \,|\, d, a) = 1 - \sum_{n=1}^{N-1} p_A(s = s_n \,|\, d, a)$. Therefore, we only need to model $P_A(s = s_1 \,|\, d, a), \ldots, P_A(s = s_{N-1} \,|\, d, a)$, which we jointly designate $P_A(s \,|\, d, a)$.

In line with the previous case, we could base $P_A(s \,|\, d, a)$ on a best guess $\pi_D(s)$, for example $p_D(s \,|\, d, a)$ when D believes that A has similar information, with some uncertainty around it. We could use a parametric probability distribution, randomising each of its parameters much as we have done in the preceding subsection. In this manner, for each pair d and a, we could estimate $\pi_{D,n}$ of $p_A(s = s_n \,|\, d, a)$ $\forall n \in \{1, \ldots, N-1\}$ and, then, incorporate the uncertainty through a uniform $\mathcal{U}(\pi_{D,n} - \mu_n, \pi_{D,n} + \mu_n)$ or a beta distribution $\mathcal{B}e(\alpha_n, \beta_n)$ centred around $\pi_{D,n}$, making sure that their sum does not exceed 1.

A more effective way would model $P_A(s \,|\, d, a)$ as a Dirichlet distribution with mean $\pi_D(s)$ and parameters assessed based on one further judgement concerning, e.g. the variance of one of the probabilities. To do this, for each pair (d, a) we would obtain from D an estimate $\pi_{D,n}$ of $p_A(s = s_n \,|\, d, a)$ $\forall n \in \{1, \ldots, N\}$ and associate random variables S_n such that $E[S_n] = \pi_{D,n}$. Their joint distribution could then be described as Dirichlet, $(S_1, \ldots, S_N) \sim \mathcal{D}ir(\alpha)$, with parameters $\alpha = (\alpha_1, \ldots, \alpha_N)$. If $\hat{\alpha} = \sum_{n=1}^{N} \alpha_n$, it follows

that

$$E[S_n] = \frac{\alpha_n}{\hat{\alpha}}, \qquad Var[S_n] = \frac{\alpha_n(\hat{\alpha} - \alpha_n)}{\hat{\alpha}^2(\hat{\alpha} + 1)};$$

and it suffices to elicit one value, e.g. $Var[S_1]$, to calculate the required α_n parameters.

4.1.3 The continuous case

We consider now the case in which the chance node S involves a continuous set of events. Techniques are similar to those described to assess the probabilities of multiple events. We could base $P_A(s \mid d, a)$ on a guess $\pi_D(s)$, say $p_D(s \mid d, a)$, with some uncertainty around it. For example, this may be achieved by means of a Dirichlet process, with base distribution $\pi_D(s)$ and concentration parameter ρ as perceived by D, which allows to sample approximate distributions of $P_A(s \mid d, a)$. Other non-parametric approaches such as hierarchical Pitman-Yor processes (Teh and Jordan, 2010) could be used with reference to the above guess.

4.2 Random utilities

We draw now attention over D's beliefs on A's preference assessments over the consequences of the decisions, that is, the random utilities. We shall usually have some information about A's multiple interests. For example, when dealing with terrorism cases, Keeney (2007) and Keeney and von Winterfeldt (2010) present extensive classifications of criteria amongst which to choose. Keeney (2007) then advocates that standard utility methods may be adopted by interviewing experts in the problem at hand, therefore developing utility functions modelling A's preferences. However, note that such preferences are not directly elicited from A, but rather through a surrogate. Thus, intrinsically, there is uncertainty about A's preferences.

An alternative approach, illustrated in Banks et al. (2015), is to aggregate the objectives with a weighted measurable value function, as in Dyer and Sarin (1979). As an example, we could consider an additive value function for A in which his objectives $v_1, \ldots, v_R$ are aggregated using weights $w_1, \ldots, w_R \geq 0$, $\sum_{r=1}^{R} w_r = 1$ as $v_A = \sum_{r=1}^{R} w_r\, v_r$. The uncertainty about the weights could be modelled using a Dirichlet distribution, as in Section 4.1.2, so that we may estimate their value and then associate random variables W_r such that $E[W_r] = w_r$, their joint distribution being Dirichlet, $(W_1, \ldots, W_R) \sim \mathcal{D}ir(\alpha)$, with parameters $\alpha = (\alpha_1, \ldots, \alpha_R)$ with one further judgement, e.g. fixing the variance of one of the parameters. Finally, using the relative risk aversion concept (Dyer and Sarin, 1982), we could assume different risk attitudes when modelling A's utility function. Continuing the example and assuming an exponential utility function, we may transform the (random) value function $V_A = \sum_{r=1}^{R} W_r\, v_r$

into one of the three following utilities depending on A's risk attitude: *risk aversion*, $U_A = 1 - \exp(-\lambda V_A + c), \quad \lambda > 0$; *risk neutrality*, $U_A = V_A + c$; or *risk proneness*, $U_A = \exp(\lambda V_A + c), \quad \lambda > 0$. Further uncertainty about the risk coefficient λ and the adjusting constant c may be modelled, e.g. through uniform distributions $\Lambda \sim \mathcal{U}(\lambda_1, \lambda_2)$ and $C \sim \mathcal{U}(c_1, c_2)$. In any case, to determine all the required distributions, we may ask experts to directly elaborate such distributions or request them to provide point estimates of the weights and coefficients and build the distributions from these.

An alternative to building a distribution over A's preferences is described in Wang and Bier (2013). As before, suppose that they are represented through a multi-attribute utility function, which involves the above attributes $v_1, \ldots, v_R$ as well as an unobserved one v_0. For simplicity, consider A's utility to be linear in the attributes. Then we ask several experts to provide rank orders of A's action valuations and derive probability distributions that can match those orderings to obtain the (random) weights $(W_0, W_1, \ldots, W_R)$ for his utility function. For this, we consider as input such rankings and as output a distribution over A's preferences (expected utilities) for which two methods are suggested. One is an adaptation of probabilistic inversion (Neslo et al., 2008); essentially, it identifies a probability distribution Q over the space of all possible attribute weights $(W_0, W_1, \ldots, W_R)$ that can match the empirical distribution matrix of expert rankings with minimum Kullback-Leibler divergence to a predetermined (e.g. non-informative, Dirichlet) starting probability measure Q_0. The other one uses Bayesian density estimation (Müller et al., 2015) based on a prior distribution Q_p (e.g. chosen in accordance to a Dirichlet process with base distribution Q_0) over the space of attribute weights $(W_0, W_1, \ldots, W_R)$ and treating the expert rankings as observations to update that prior leading to a posterior distribution Q, obtained through Gibbs sampling.

5 A Numerical Example

As an illustration, consider a sequential defend-attack cybersecurity problem. A user (D, defender) needs to make a connection to a site, either through a safe, but costly, route (d_0) or through a cheaper, but more dangerous protocol. In the latter case, she may use a security key, rendering the protocol less dangerous. While using the dangerous protocol, whether unprotected (d_1) or protected by a security key (d_2), the defender may be the target of a cybercriminal (A, attacker) who may decide to attack (a_1) or not (a_0). The case may be viewed through the game tree in Figure 4.

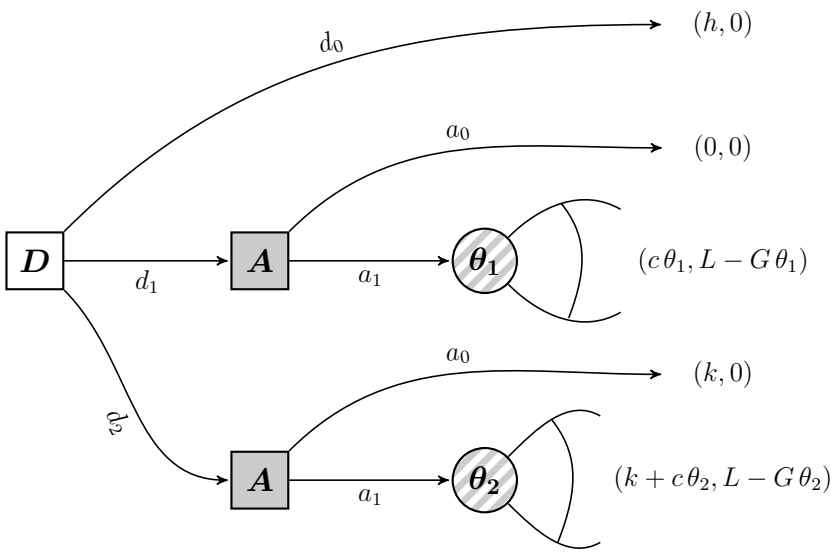

Figure 4: Game tree for the cybersecurity routing problem (losses). Outcomes after $\theta_i, i = 1, 2$ are continuous.

The following parameters are used: (i) h is the cost of using the expensive protocol; (ii) θ_1 is the fraction of assets lost by the defender when attacked and unprotected; (iii) θ_2 is the fraction of assets lost by the defender when attacked but protected; (iv) k is the security key's cost; (v) c is the defender's scaling cost relative to the fraction of assets lost; (vi) L is the uncertain cost of an attack; and (vii) G is the uncertain cybercriminal's scaling gain relative to the fraction of assets lost by the defender. Table 2 (respectively, Table 3) displays the defender's (respectively, attacker's) consequences, expressed as costs, for the various defend and attack possibilities, reflected in the tree

		Attack	
		$\boldsymbol{a_0}$	$\boldsymbol{a_1}$
Defence	$\boldsymbol{d_0}$	h	—
	$\boldsymbol{d_1}$	0	$c\,\theta_1$
	$\boldsymbol{d_2}$	k	$k + c\,\theta_2$

Table 2: Defender's loss function.

		Attack	
		$\boldsymbol{a_0}$	$\boldsymbol{a_1}$
Defence	$\boldsymbol{d_0}$	0	—
	$\boldsymbol{d_1}$	0	$L - G\,\theta_1$
	$\boldsymbol{d_2}$	0	$L - G\,\theta_2$

Table 3: Attacker's loss function.

The defender believes that the asset fractions θ_i follow distributions $p_D(\theta_i \,|\, d_i, a_1)$ with $\theta_i \sim \mathcal{B}e(\alpha_i^D, \beta_i^D)$, $i = 1, 2$. She is risk averse and her utility function is strategically equivalent to $1 - e^{\lambda_D\, x}$, where x is her cost and $\lambda_D > 0$ her risk aversion coefficient. She

expects θ_1 to be greater than θ_2 (but not necessarily), reflected in the choice of the beta parameters, with $E[\theta_1] = \frac{\alpha_1^D}{\alpha_1^D+\beta_1^D} > \frac{\alpha_2^D}{\alpha_2^D+\beta_2^D} = E[\theta_2]$. Table 4 provides the defender's expected utilities u_D under the various interaction scenarios.

		Attack	
		$\boldsymbol{a_0}$	$\boldsymbol{a_1}$
Defence	$\boldsymbol{d_0}$	$1 - e^{\lambda_D h}$	—
	$\boldsymbol{d_1}$	0	$1 - \int e^{\lambda_D c\theta_1} p_D(\theta_1)\,\mathrm{d}\theta_1$
	$\boldsymbol{d_2}$	$1 - e^{\lambda_D k}$	$1 - \int e^{\lambda_D (k+c\theta_2)} p_D(\theta_2)\,\mathrm{d}\theta_2$

Table 4: Defender's expected utility.

Suppose we assess from the defender the following parameter values (with standard elicitation techniques): (i) a protocol cost $h = 150,000$ €; (ii) a security key cost $k = 50,000$ €; (iii) a scaling cost $c = 200,000$ €; (iv) a risk aversion coefficient $\lambda_D = 3 \cdot 10^{-5}$; (v) the distribution $\theta_1 \sim \mathcal{B}e(\alpha_1^D, \beta_1^D)$ with expected fraction (mean) of 0.6 of the assets lost and standard deviation 0.15 when attacked and unprotected, leading to $\alpha_1^D = 0.36$ and $\beta_1^D = 0.24$; (vi) and the distribution $\theta_2 \sim \mathcal{B}e(\alpha_2^D, \beta_2^D)$ with expected fraction (mean) of 0.3 of the assets lost and standard deviation 0.07 when attacked but protected, leading to $\alpha_2^D = 0.6$ and $\beta_2^D = 1.4$. These are standard decision analytic assessments and the resulting problem faced by her is described in the decision tree in Figure 5.

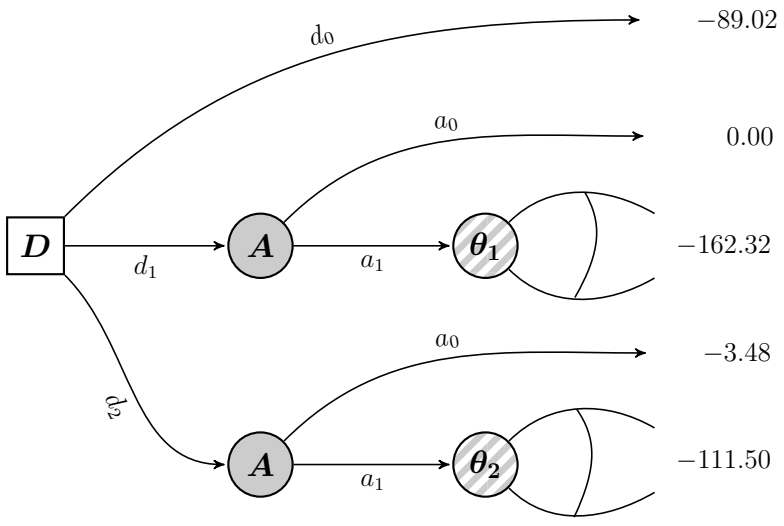

Figure 5: Decision tree representing the defender's problem (expected utilities).

The expected utility of the first alternative (d_0, use the expensive protocol) may be directly estimated as

$$\psi_D(d_0) = 1 - e^{\lambda_D\, h} \approx -89.02,$$

since there is no chance of attack in this scenario. However, those of the other two alternatives have the form

$$\psi_D(d_i) = \sum_{j=0}^{1} p_D(a_j \,|\, d_i)\, u_D(d_i, a_j), \quad i = 1, 2;$$

where $u_D(d_i, a_j)$ may be obtained from Table 4 with the specific values indicated in Figure 5. Thus, we need to assess the attack probabilities $p_D(a_1 \,|\, d_i)$ (and $p_D(a_0 \,|\, d_i) = 1 - p_D(a_1 \,|\, d_i), i = 1, 2$) and we adopt an ARA approach to assess them.

The attacker has different beliefs about θ_i, $p_A(\theta_i \,|\, d_i, a_1)$, with $\theta_i \sim \mathcal{B}e(\alpha_i^A, \beta_i^A)$, $i = 1, 2$; the defender's uncertainty about α_i^A and β_i^A inducing its randomness. He is risk prone and his utility function is strategically equivalent to $e^{-\Lambda_A\, x} - 1$, where x is his cost and $\Lambda_A > 0$ his uncertain risk proneness coefficient. Table 5 provides the attacker's random expected utilities respectively, U_A under the various interaction scenarios.

		Attack	
		$\boldsymbol{a_0}$	$\boldsymbol{a_1}$
Defence	$\boldsymbol{d_0}$	0	—
	$\boldsymbol{d_1}$	0	$\int e^{\Lambda_A (G\,\theta_1 - L)}\, P_A(\theta_1)\, \mathrm{d}\theta_1 - 1$
	$\boldsymbol{d_2}$	0	$\int e^{\Lambda_A (G\,\theta_2 - L)}\, P_A(\theta_2)\, \mathrm{d}\theta_2 - 1$

Table 5: Attacker's random expected utility.

Suppose that, in line with Section 4, we assess that: (i) $L \sim \mathcal{U}(10^4, 2 \cdot 10^4)$ with an expected cost of 15,000 €; (ii) $G \sim \mathcal{U}(10^4, 5 \cdot 10^4)$ with an expected scaling gain of 30,000 €; (iii) $\Lambda_A \sim \mathcal{U}(10^{-4}, 2 \cdot 10^{-4})$ with an expectation of $1.5 \cdot 10^{-4}$; (iv) the distribution $\theta_1 \sim \mathcal{B}e(\alpha_1^A, \beta_1^A)$ has a expected fraction (mean) of 0.6 assets lost when the defender is attacked but protected, with $\alpha_1^A \sim \mathcal{U}(5, 7)$ and $\beta_1^A \sim \mathcal{U}(3, 5)$; and (v) the distribution $\theta_2 \sim \mathcal{B}e(\alpha_2^A, \beta_2^A)$ has a expected fraction (mean) of 0.3 assets lost when the defender is attacked but protected, with $\alpha_2^A \sim \mathcal{U}(2, 4)$ and $\beta_2^A \sim \mathcal{U}(6, 8)$. We may then use Algorithm 1 to estimate the required probabilities $\hat{p}_D(a_1 \,|\, d)$, where $\Psi_A^n(d_i, a)$ designates the expected utility that the cybercriminal obtains when the defender implements d, he chooses action a and the sampled parameters are $l^n, g^n, \lambda_A^n, \alpha_i^{A,n}$ and $\beta_i^{A,n}$.

Algorithm 1 Numerical example: Simulation of $\hat{p}_D(a_1 \mid d)$

Data: Number of iterations N.

Set $p_1, p_2 = 0$.
For $n = 1$ **to** N **do**
  Draw l^n from $\mathcal{U}(10^4, 2 \cdot 10^4)$, g^n from $\mathcal{U}(10^4, 5 \cdot 10^4)$.
  Draw λ_A^n from $\mathcal{U}(10^{-4}, 2 \cdot 10^{-4})$.
  Draw $\alpha_1^{A,n}$ from $\mathcal{U}(2, 7)$, $\beta_1^{A,n}$ from $\mathcal{U}(1, 5)$.
  Draw $\alpha_2^{A,n}$ from $\mathcal{U}(0, 3)$, $\beta_2^{A,n}$ from $\mathcal{U}(1, 6)$.
  For $i = 1$ **to** 2 **do**
    $\Psi_A^n(d_i, a_0) = 0$.
    $\Psi_A^n(d_i, a_1) = \int e^{\lambda_A^n (g^n \theta_i - l^n)} \dfrac{\theta_i^{\alpha_i^{A,n}-1} (1-\theta_i)^{\beta_i^{A,n}-1}}{\mathrm{Beta}(\alpha_i^{A,n}, \beta_i^{A,n})} \, \mathrm{d}\theta_i - 1$.
    If $\Psi_A^n(d_i, a_1) \geq \Psi_A^n(d_i, a_0)$ **then**
      $p_i = p_i + 1$.
    End If
  End For
End For
For $i = 1$ **to** 2 **do**
  $\hat{p}(a_1 \mid d_i) = p_i / N$.
End For

In our case, with $N = 10^6$, we obtain $\hat{p}(a_1 \mid d_1) = 0.66$ (and, consequently, $\hat{p}(a_0 \mid d_1) = 0.34$). Similarly, $\hat{p}(a_1 \mid d_2) = 0.23$ (and $\hat{p}(a_0 \mid d_2) = 0.77$). Then, we have $\psi_D(d_0) = -89.02$, $\psi_D(d_1) = -107.13$ and $\psi_D(d_2) = -28.32$. Thus, the optimal cyberdefense is $d^*_{ARA} = d_2$, that is, employing the dangerous protocol protected by the security key.

6 Discussion

ARA is an emergent paradigm when supporting a decision maker who faces adversaries so that the attained consequences are random and depend on the actions of all participating agents. We have illustrated the relevance of such approach as a decomposition method to forecast adversarial actions in competitive contexts, therefore being of relevance to the SEJ toolkit. We have also presented key implementation strategies. We have limited the analysis to the simpler sequential case, but ideas extend to simultaneous problems, albeit with technical difficulties, due to the belief recursions typical of level-k thinking.

As usual, in applications this tool could be combined with other SEJ strategies.

For example, when assessing $p_D(a|d)$ we could use extending the conversation through $\sum_i p_D(a|b_i, d)p_D(b_i)$ and then assess the $p_D(a|b_i, d)$ probabilities through ARA. Similarly, throughout the discussion we have assumed just one single expert available to provide the $p(a|d)$ probabilities through ARA. In practice, several experts might be available and we could aggregate their ARA probabilities through e.g. Cooke's classical method (Cooke, 1991). Diverse adversarial rationalities, such as non-strategic or prospect-maximising players, could be handled by means of mixtures.

The ARA decomposition strategy breaks down an attack probability assessment into (random) multi-attribute utility and probability assessments for the adversary. This approach may lead to more precise probabilities than the ones that would have been directly obtained and, also, that the corresponding increased number of necessary judgements are cognitively easier. Behavioural experiments will be conducted to validate these ideas.

Acknowledgements

The work of DRI is supported by the Spanish Ministry of Economy and Innovation programs MTM2014-56949-C3-1-R and MTM2017-86875-C3-1-R AEI/FEDER, UE, the ESF-COST Action IS1304 on Expert Judgement and the AXA-ICMAT Chair on Adversarial Risk Analysis. JGO's research is financed by the Spanish Ministry of Economy and Competitiveness under FPI SO grant agreement BES-2015-072892. This work has also been partially supported by the Spanish Ministry of Economy and Competitiveness, through the "Severo Ochoa" Program for Centers of Excellence in R&D (SEV-2015-0554) and project MTM2015-72907-EXP, received funding from the European Union's H2020 Program for Research, Technological Development and Demonstration, under grant agreement no. 740920 (CYBECO) and based upon work partially supported by the US National Science Foundation through grant DMS-163851.

References

S. Andradottir and V. M. Bier. Choosing the number of conditioning events in judgemental forecasting. *Journal of Forecasting*, 16(4): 255–286, 1997.

S. Andradottir and V. M. Bier. An analysis of decomposition for subjective estimation in decision analysis. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 28(4): 443–453, 1998.

D. Banks, J. Ríos and D. Ríos Insua. *Adversarial Risk Analysis*. CRC Press, Boca Raton, FL, 2015.

T. Bedford and R. M. Cooke. *Probabilistic Risk Analysis: Foundations and Methods* (first published 2001). Cambridge University Press, Cambridge, United Kingdom, 2011.

E. Chen, D. V. Budescu, S. K. Lakshmikanth, B. A. Mellers and P. E. Tetlock. Validating the contribution-weighted model: Robustness and cost-benefit analyses. *Decision Analysis*, 13(3): 128-152, 2016.

R. T. Clemen and T. Reilly. *Making Hard Decisions with DecisionTools*. Cengage Learning, Mason, OH, 2013.

R. M. Cooke. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford University Press, New York, NY, 1991.

J. S. Dyer and R. K. Sarin. Group preference aggregation rules based on strength of preference. *Management Science*, 25(9): 822–832, 1979.

J. S. Dyer and R. K. Sarin. Relative risk aversion. *Management Science*, 28(8): 875–886, 1982.

S. French and D. Ríos Insua. *Statistical Decision Theory*. Wiley, New York, NY, 2000.

J. González-Ortega, V. Radovic and D. Ríos Insua. Utility elicitation. In *Elicitation: The Science and Art of Structuring Judgement* (pp. 241–264), Springer, New York, NY, 2018.

L. H. J. Goossens, R. M. Cooke and B. C. P. Kraan. Evaluation of weighting schemes for expert judgement studies. *Proceedings of the Fourth International Conference on Probabilistic Safety Assessment and Management*: 1937–1942, 1998.

S. Hargreaves-Heap and Y. Varoufakis. *Game Theory: A Critical Introduction* (first published 1995). Routledge, New York, NY, 2004.

J. B. Kadane and P. D. Larkey. Subjective probability and the theory of games. *Management Science*, 28(2): 113–120, 1982.

G. L. Keeney and D. von Winterfeldt. Identifying and structuring the objectives of terrorists. *Risk Analysis*, 30(12): 1803–1816, 2010.

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

D. G. MacGregor. Decomposition for judgmental forecasting and estimation. In *Principles of Forecasting* (pp. 107–123), Springer, Boston, MA, 2001.

D. G. MacGregor and J. S. Armstrong. Judgmental decomposition: When does it work? *International Journal of Forecasting*, 10(4): 495–506, 1994.

G. Montibeller and D. von Winterfeldt. Biases and debiasing in multi-criteria decision analysis. *IEEE 2015 48th Hawaii International Conference on System Sciences*: 1218–1226, 2015.

P. Müller, F.A. Quintana, A. Jara and T. Hanson. *Bayesian Nonparametric Data Analysis*. Springer, Cham, Switzerland, 2015.

R. Neslo, F. Micheli, C. V. Kappel, K. A. Selkoe, B. S. Halpern and R. M. Cooke. Modeling stakeholder preferences with probabilistic inversion: Application to prioritizing marine ecosystem vulnerabilities. In *Real-Time and Deliberative Decision Making* (pp. 265–284), Springer, Dordrecht, Netherlands, 2008.

A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley and T. Rakow. Uncertain Judgements: Eliciting Experts' Probabilities. Wiley, Chichester, UK, 2006.

H. Raiffa Decision Analysis: Introductory Lectures on Choices Under Uncertainty. Addison-Wesley, Menlo Park, CA, 1968.

H. Raiffa. *The Art and Science of Negotiation* (first published 1982). Harvard University Press, Cambridge, MA, 2003. arth

H. V. Ravinder, D. N. Kleinmuntz and J. S. Dyer. The reliability of subjective probabilities obtained through decomposition. *Management Science*, 34(2): 186–199, 1988.

D. Ríos Insua, J. Ríos and D. Banks. Adversarial risk analysis. *Journal of the American Statistical Association*, 104(486): 841–854, 2009.

Y. W. Teh and M. I. Jordan. Hierarchical Bayesian nonparametric models with applications. In *Cambridge Series in Statistical and Probabilistic Mathematics: Bayesian nonparametrics* (pp. 158–207), Cambridge University Press, New York, NY, 2010.

P. E. Tetlock and D. Gardner. *Superforecasting: The Art and Science of Prediction* (2015 ed.). Broadway Books, New York, NY, 2015.

C. Wang and V. M. Bier. Expert elicitation of adversary preferences using ordinal judgments. *Operations Research*, 61(2): 372–385, 2013.