From risk analysis to adversarial risk analysis

Part VII. Adversarial risk analysis

David Ríos,
AXA-ICMAT Chair @ICMAT-CSIC and R. Academy

DRI. Aalto

Which is the best security resource allocation in a railway network?

Railway Network as stations, lines (&hotspots)

Threats: Pickpocketing, Fare evasion, Terrorism, …

Each element has a value

For each element, each threat, a predictive model of acts

Allocate security resources (constraints)

For each cell predict the impact of resource allocation

Optimal resource allocation

Which is the best security resource allocation in a railway network?

Railway Network as stations, lines (&hotspots)

Threats: Pickpocketing, Fare evasion, Terrorism, …

Each element has a value

For each element, each threat, a predictive model of acts

Allocate security resources (constraints)

For each cell predict the impact of resource allocation

Optimal resource allocation

NB1: Bad guys operate intelligent and organisedly!!!

Which is the best security resource allocation in a railway network?

Railway Network as stations, lines (&hotspots)

Threats: Pickpocketing, Fare evasion, Terrorism, …

Each element has a value

For each element, each threat, a predictive model of acts

Allocate security resources (constraints)

For each cell predict the impact of resource allocation

Optimal resource allocation

NB1: Bad guys operate intelligent and organisedly!!!

NB2: Different bad guys uncoordinated…

From RA to ARA…

Motivation

- ‘The World’s (23) Biggest Problems´ (Lomborg)
  - Arms proliferation
  - Conflicts
  - Corruption
  - Terrorism
  - Drugs
  - Money laundering
- One of H2020 priorities (Secure Societies FCT, BD, DS)

Motivation

- RA extended to include adversaries ready to increase our risks

- S-11, M-11 lead to large security investments globally, some of them criticised
- Many modelling efforts to efficiently allocate such resources
- Parnell et al (2008) NAS review
  - Standard reliability/risk approaches not take into account intentionality
  - Game theoretic approaches. Common knowledge assumption…
  - Decision analytic approaches. Forecasting the adversary action…
- Merrick, Parnell (2011) review approaches commenting favourably on ARA

ARA

- A framework to manage risks from actions of intelligent adversaries

  Banks, Rios, DRI Adversarial Risk Analysis (2015) Taylor Francis

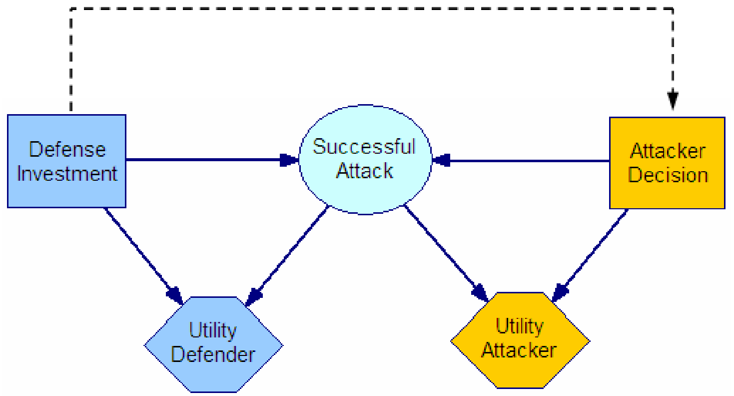
- One-sided prescriptive support
  - Use a SEU model
  - Treat the adversary's decision as uncertainties
- Method to predict adversary's actions
  - We assume the adversary is a *expected utility maximizer*
    - Model his decision problem
    - Assess his probabilities and utilities
    - Find his action of maximum expected utility

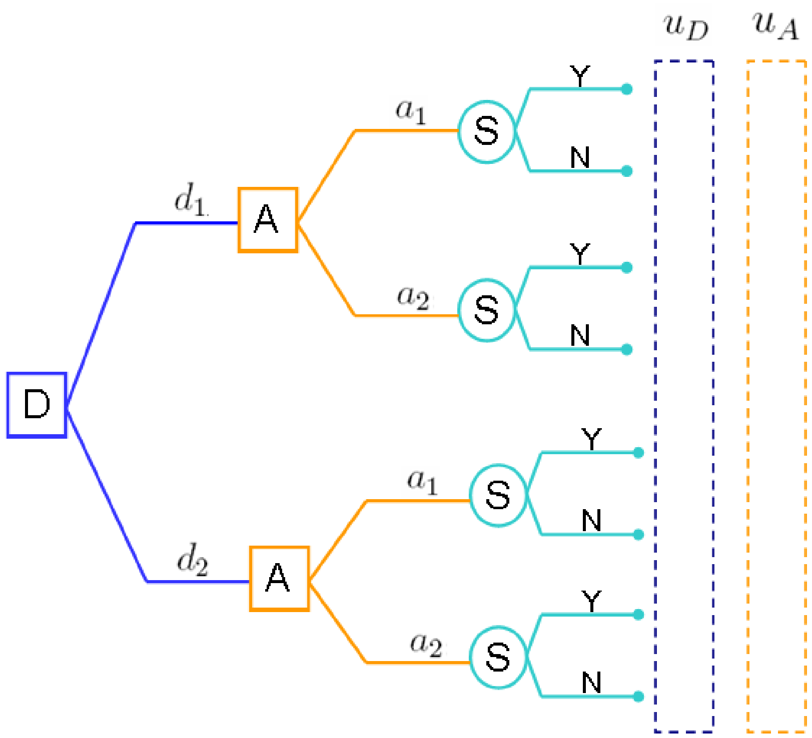
  (But other *descriptive* models are possible)
- Uncertainty in the Attacker's decision stems from
  - *our* uncertainty about his probabilities and utilities
  - but this leads to a hierarchy of nested decision problems

  (random, noninformative, level-k, heuristic, mirroring argument,…) vs (common knowledge)

- Kadane, Larkey (1982), Raiffa (1982, 2002)
- Lippman, McCardle (2012)
- Stahl and Wilson (1994, 1995) D. Wolpert (2012)
- Rothkopf (2007)

First Defender, afterwards Attacker





$$a^*(d) = \mathrm{argmax}_{a \in \mathcal{A}} \psi_A(d,a), \forall d \in \mathcal{D}$$

$$d^* = \mathrm{argmax}_{d \in \mathcal{D}} \psi_D(d, a^*(d))$$

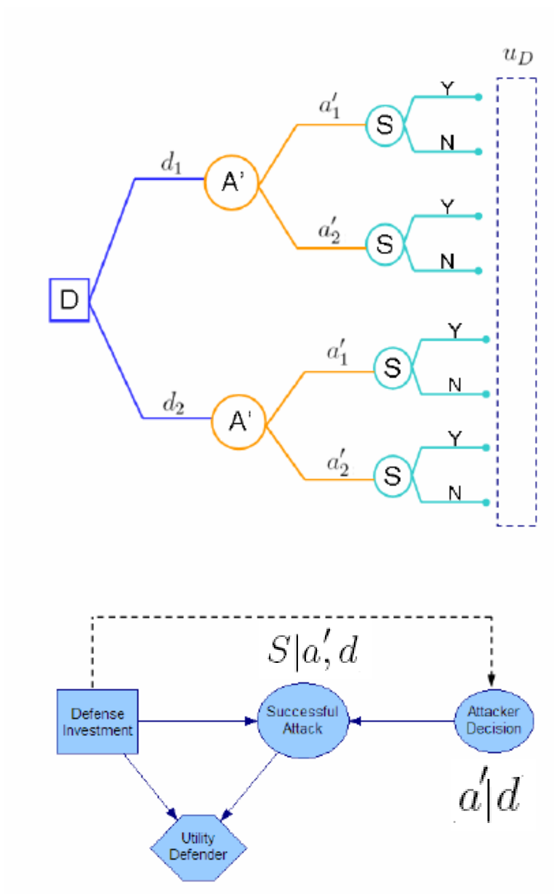
Nash Solution, SPE: $(d^*, a^*(d^*))$

DRI. Aalto

Standard Game Theory Analysis

Supporting the Defender

Defender problem



Defender's solution

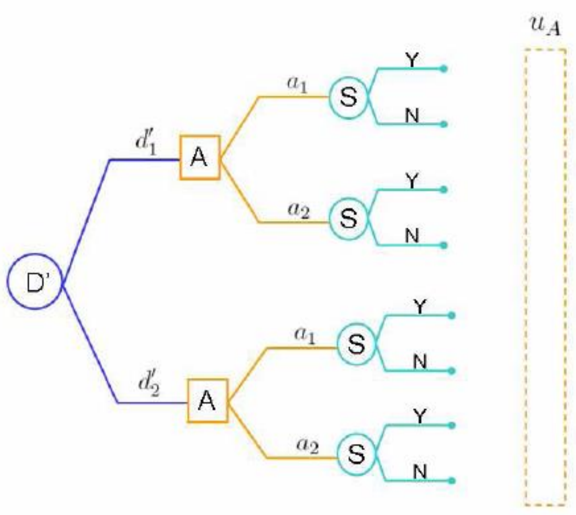
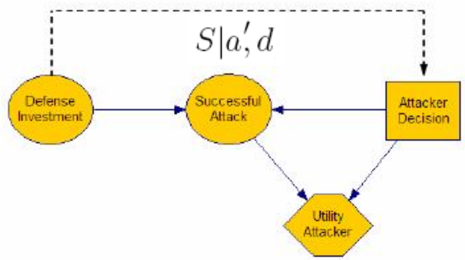
$$\psi_D(d,a') = u_D(d, S=Y)\, p_D(S=Y \mid X_D = d, X'_A = a') + u_D(d, S=N)\, p_D(S=N \mid X_D = d, X'_A = a')$$

$$\psi_D(d) = \psi_D(d,a'_1)\, p_D(a'_1 \mid d) \;+\; \psi_D(d,a'_2)\, p_D(a'_2 \mid d)$$

$$d^* = \arg\max_{d \in X_D} \psi_D(d)$$

Modeling input: $p_D(S|a',d)$ $p_D(a'|d)$??

Supporting the Defender: The assessment problem

Defender's view of Attacker problem	Elicitation of $p_D(a'|d)$
 	Assume A is a EU maximizer D's beliefs about $(\widehat{u}_A, \widehat{p}_A) \sim F$ $\hat{\psi}_A(d', a) = \hat{u}_A(a, S = Y)\, \hat{p}_A(S = Y \mid X'_D = d',\ X_A = a) + \hat{u}_A(a, S = N)\, \hat{p}_A(S = N \mid X'_D = d',\ X_A = a)$ $\hat{\psi}_A \sim \widehat{\Psi}_A$ $p_D(a'|d) = Pr\Big[a' = \arg\max_{x \in X'_A} \widehat{\Psi}_A(d, x)\Big]$ <u>MC simulation</u> $\hat{p}_D(a|d) \approx n^{-1} \sum_i \#\{a = \mathrm{argmax}_{x \in \mathcal{A}} \hat{\psi}^i_A(x, d)\}$ where $\hat{\psi}^i_A \sim \widehat{\Psi}_A,\ i = 1, \ldots, n$ DRI. Aalto

Sequential D-A

1. Assess (p_D, u_D) from the Defender

2. Assess $F = (P_A, U_A)$, describing the Defender's uncertainty about (p_A, u_A)

3. For each d, simulate to assess $p_D(A|d)$ as follows:

   (a) Generate $(p_A^i, u_A^i) \sim F$, $i = 1, \ldots, n$

   Solve $a_i^*(d) = \mathrm{argmax}_{a \in \mathcal{A}}\ \psi_A^i(d, a)$

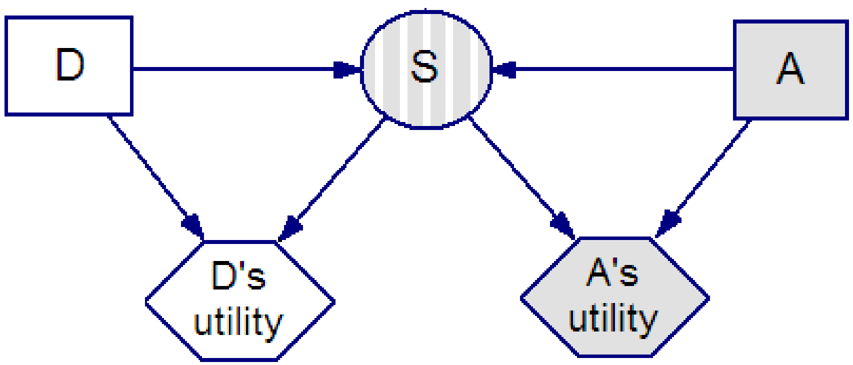
   (b) Approximate $\hat{p}_D(A = a|d) = \#\{a = a_i^*(d)\}/n$

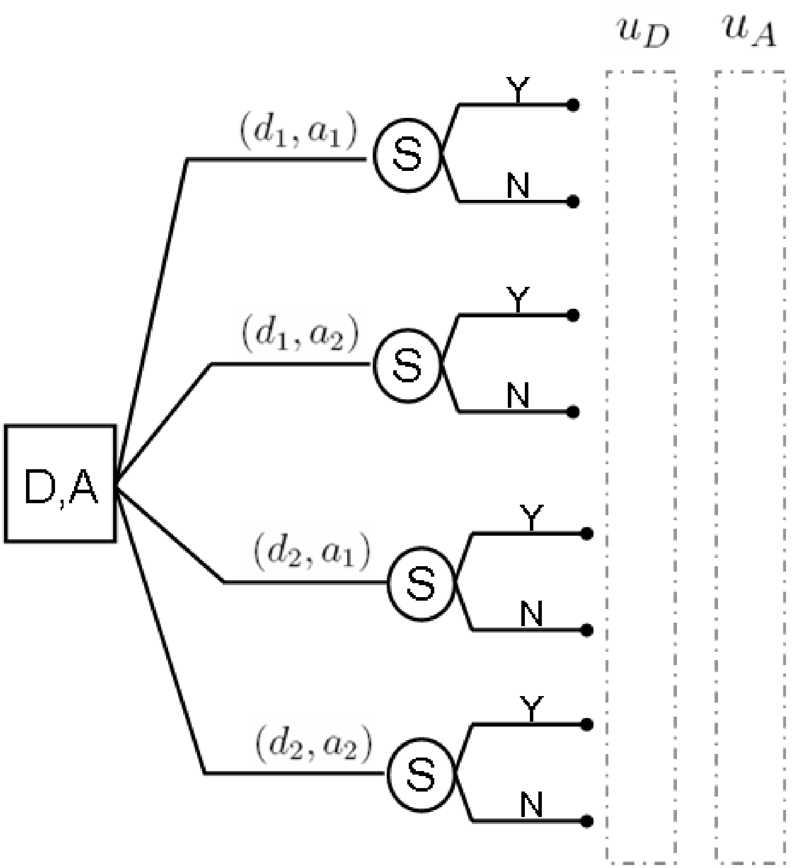
4. Solve the Defender's problem

$$d^* = \mathrm{argmax}_{d \in \mathcal{D}}\ \psi_D(d, a_1)\ \hat{p}_D(A = a_1|d)\ +\ \psi_D(d, a_2)\ \hat{p}_D(A = a_2|d)$$

Simultaneous games

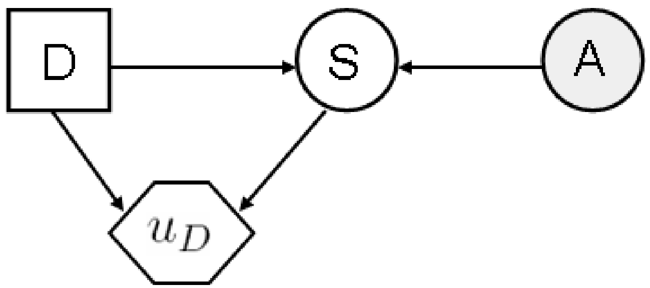
- Decisions are made without knowing each other's decisions

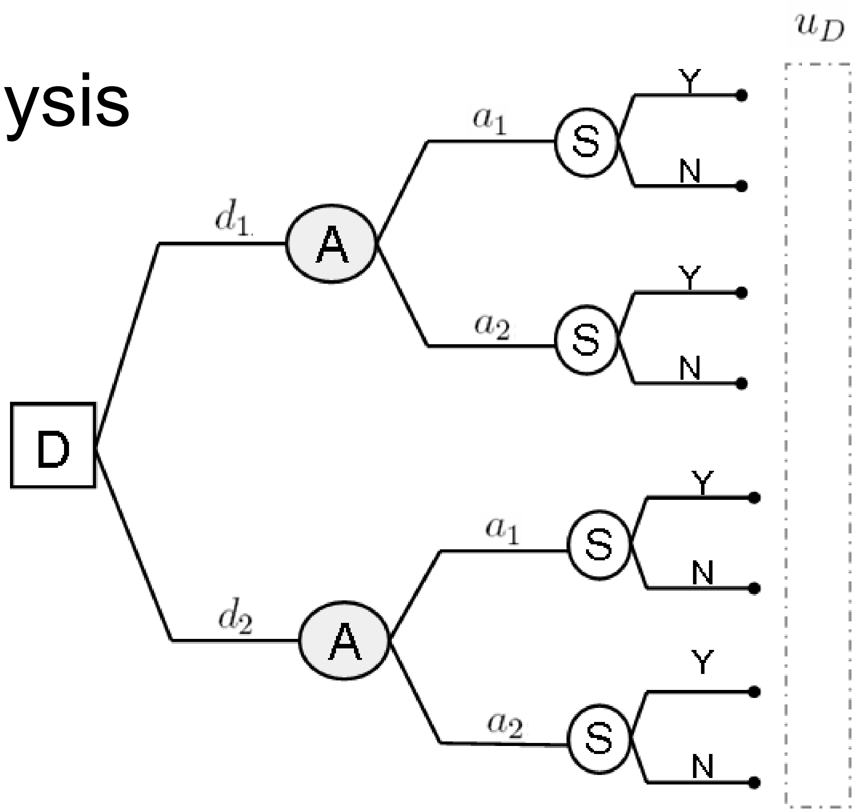




Supporting the Defender

- Defender's decision analysis



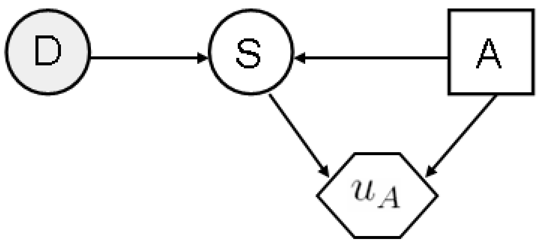


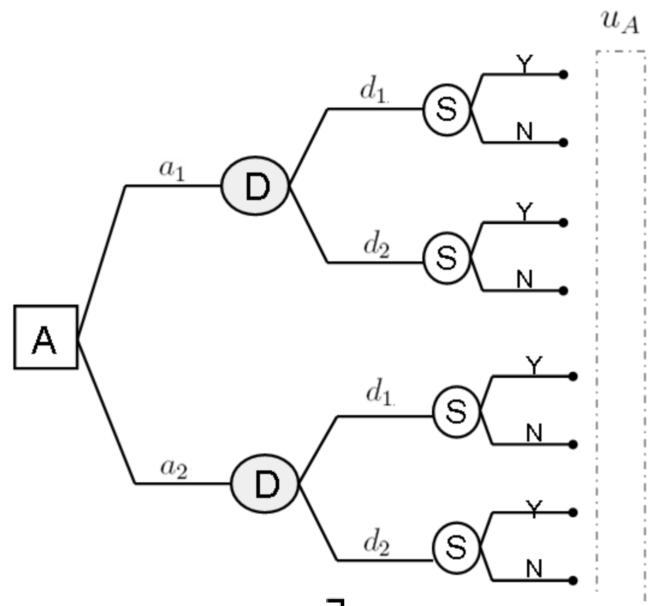
$$d^* = \operatorname{argmax}_{d \in \mathcal{D}} \sum_{a \in \mathcal{A}} \left[\sum_{s \in \{0,1\}} u_D(d, s)\, p_D(S = s \mid d, a) \right] \pi_D(A = a)$$

How to assess it ??

Assessing $\pi_D(A = a)$

- Attacker's decision analysis as seen by the Defender





$$a^* = \operatorname{argmax}_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}} \left[\sum_{s \in \{0,1\}} u_A(a, s)\ p_A(S = s \mid d, a) \right] \pi_A(D = d)$$

$$(u_A, p_A, \pi_A) \sim (U_A, P_A, \Pi_A)$$

$$A \mid D \sim \operatorname{argmax}_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}} \left[\sum_{s \in \{0,1\}} U_A(a, s)\ P_A(S = s \mid d, a) \right] \Pi_A(D = d)$$

The assessment problem

- To predict Attacker's decision

  The Defender needs to solve Attacker's decision problem

  She needs to assess (u_A, p_A, π_A)

- Her beliefs about (u_A, p_A, π_A) are modeled through a probability distribution (U_A, P_A, Π_A)

- The assessment of $\Pi_A(D = d)$ requires deeper analysis
  - D's analysis of A's analysis of D's problem

- It leads to an infinite regress thinking-about-what-the-other-is-thinking-about…

Hierarchy of nested models

Repeat

Find $\Pi_{D^{i-1}}(A^i)$ by solving

$$A^i \mid D^i \sim \mathrm{argmax}_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}} \left[\sum_{s \in \{0,1\}} U_A^i(a,s)\ P_A^i(S = s \mid d, a) \right] \Pi_{A^i}(D^i = d)$$

$$\text{where } (U_A^i, P_A^i) \sim F^i$$

Find $\Pi_{A^i}(D^i)$ by solving

$$D^i \mid A^{i+1} \sim \mathrm{argmax}_{d \in \mathcal{D}} \sum_{a \in \mathcal{A}} \left[\sum_{s \in \{0,1\}} U_D^i(d,s)\ P_D^i(S = s \mid d, a) \right] \Pi_{D^i}(A^{i+1} = a)$$

$$\text{where } (U_D^i, P_D^i) \sim G^i$$

$i = i + 1$

Stop when the Defender has no more information about utilities and probabilities at some level of the recursive analysis. Level-k thinking

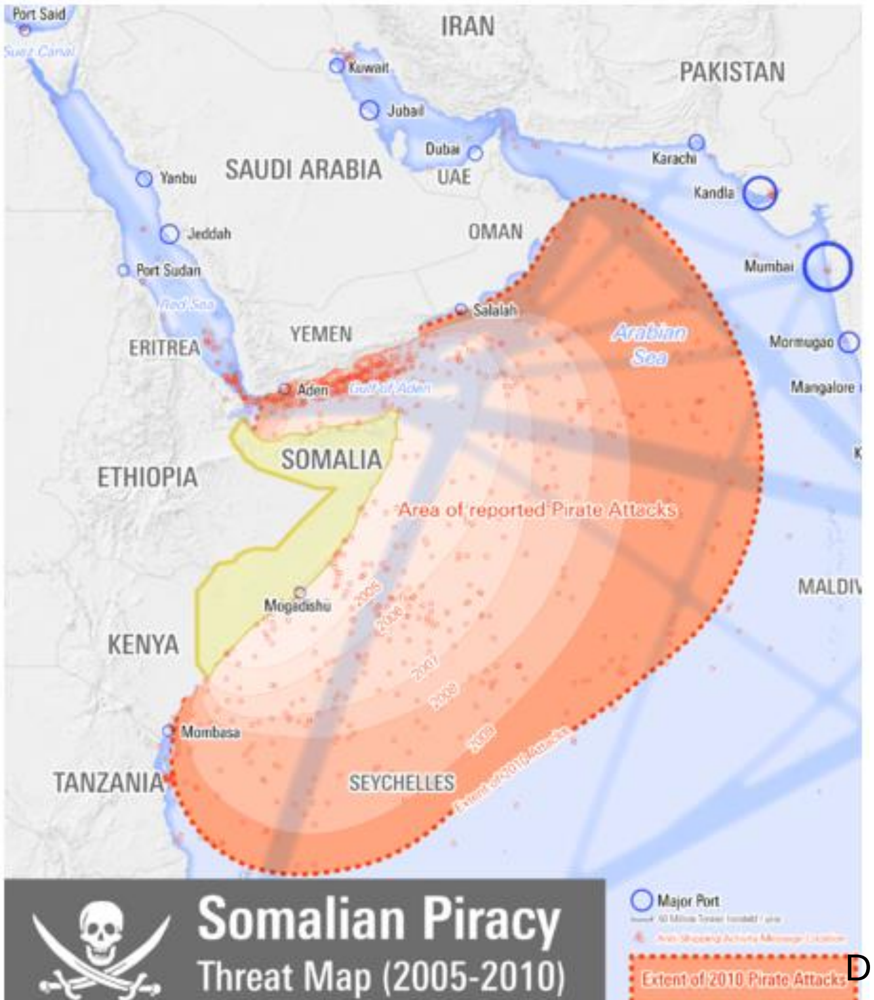
Opponent modeling

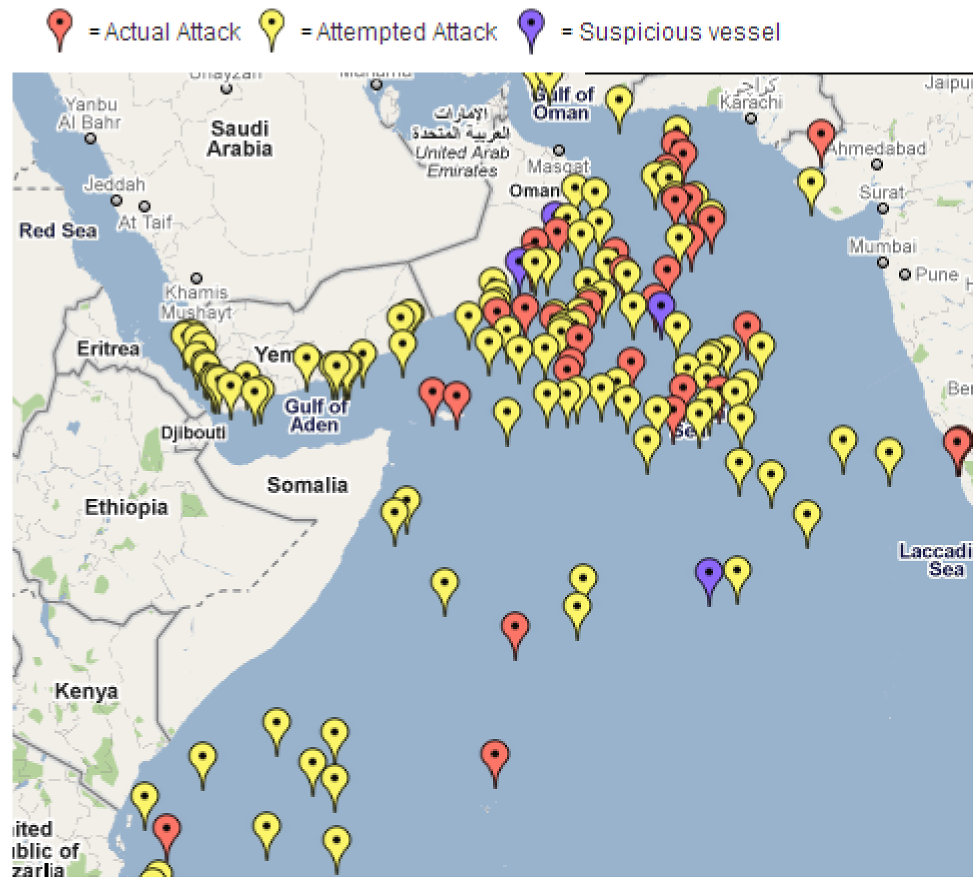
- Non strategic
- Nasheq
- Level-k
- Mirroreq
- Prospectmax

- Reconcile them through a mixture

DRI, Banks, Rios (2015) RA

Piracy in Somalia





Piracy and armed robbery incidents reported to the IMB Piracy Reporting Centre 2011

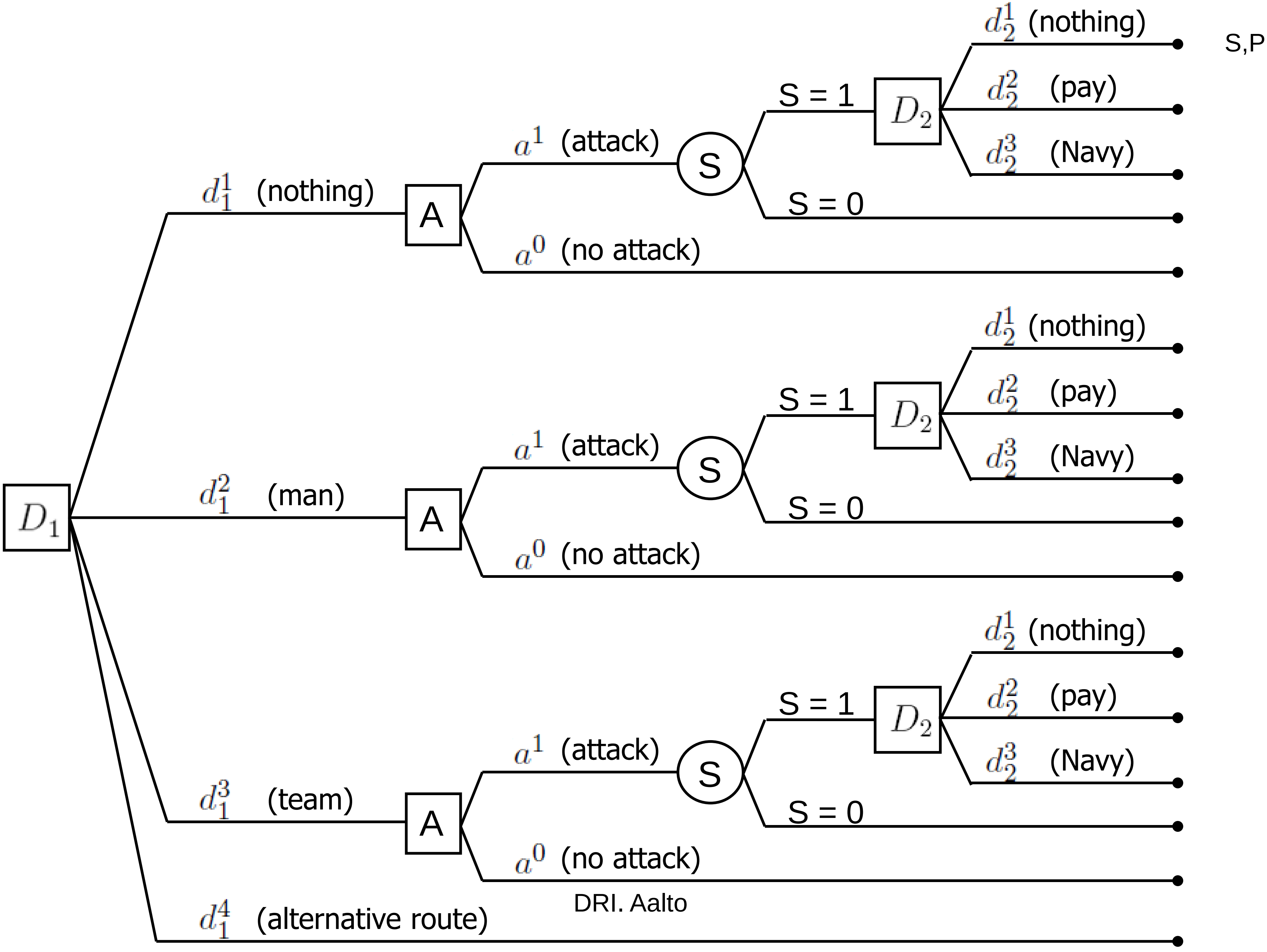
The Defend–Attack–Defend model

- Two intelligent players
  - Defender and Attacker

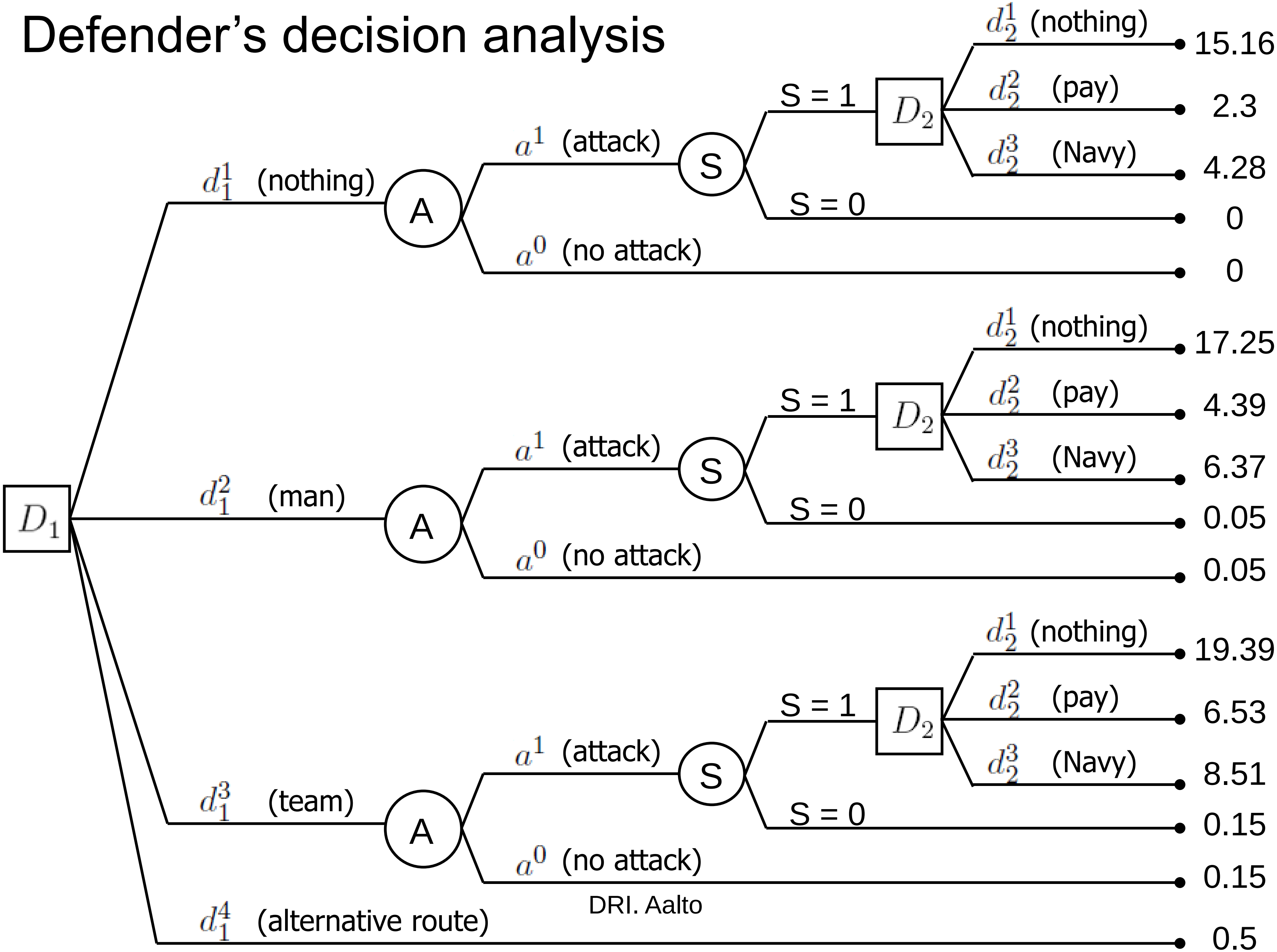
- Sequential moves
  - First, Defender moves
  - Afterwards, Attacker knowing Defender's move
  - Afterwards, Defender again responding to attack

The Somali Pirates Case: Problem formulation

- Two players
  - Defender: Ship owner
  - Attacker: Pirates
- Defender first move
  - Do nothing
  - Private protection with an armed person
  - Private protection with a team of two armed persons
  - Go through the Cape of Good Hope avoiding the Somali coast
- Attacker’s move
  - Attack or not to attack the Defender’s ship
- Defender response to an eventual kidnapping
  - Do nothing
  - Pay the ransom
  - Ask the Navy for support to release the boat and crew


S,P
D_1
d_1^1 (nothing)
d_1^2 (man)
d_1^3 (team)
d_1^4 (alternative route)
A
a^1 (attack)
a^0 (no attack)
S
S = 1
S = 0
D_2
d_2^1 (nothing)
d_2^2 (pay)
d_2^3 (Navy)
DRI. Aalto

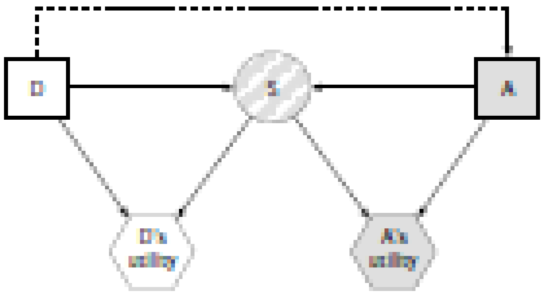
Defender's decision analysis



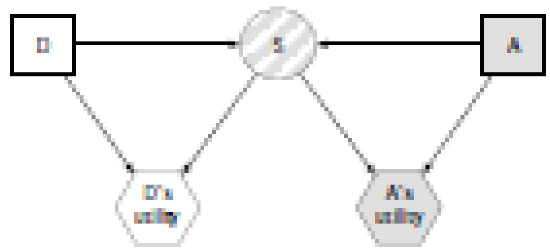
ARA: Cases

Problem	Defender	Attacker	Specificities	Template
ATC protection	Airport authority	Terrorist	Single site	D-> A
Piracy	Ship owner	Pirates	Single site	D- >A - > D
Metro	Operator	Pickpock Fare evasion	Multisite Multiattack, Cascade	D->A
Urban security	Police	Mob	Multisite spatial	D->A->D
Train	DoT, DoD	Terrorist	Multisite network	D->A->D
Reliability	Manufacturer	Customer	--	D->A
SME IS	Company	Competitor	Cyber, Integrated with RA	D->A
Oil rig cybercontrolled	Oil company	Sponsored hackers	Cyber, Multiattack	D->A->D
UAV fight	Country	Country	Multisite	D->A->D
CI	Owner	Terrorist	Multistage	General
Cybersec res allocation+cybins	IT Owner	Hacker(s)	Several decisions Random and targeted attacks	D-A, D-A-D
Social robots	Robot	User	Sequential	D->A

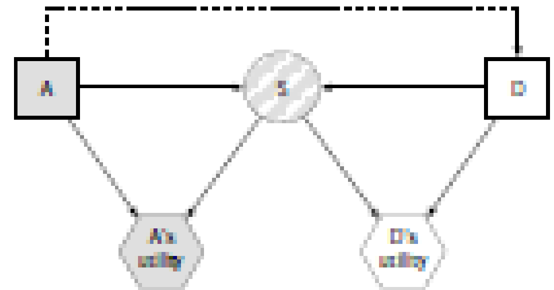
ARA templates



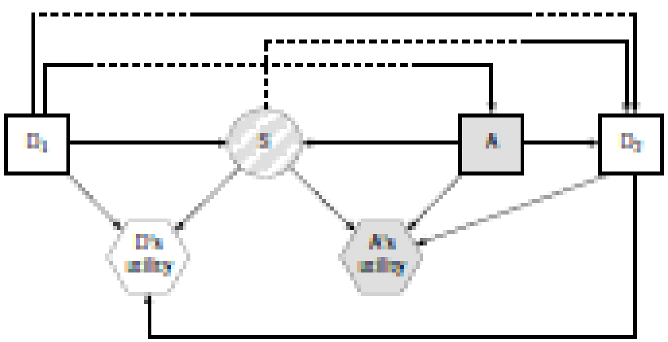
(a) Seq D-A



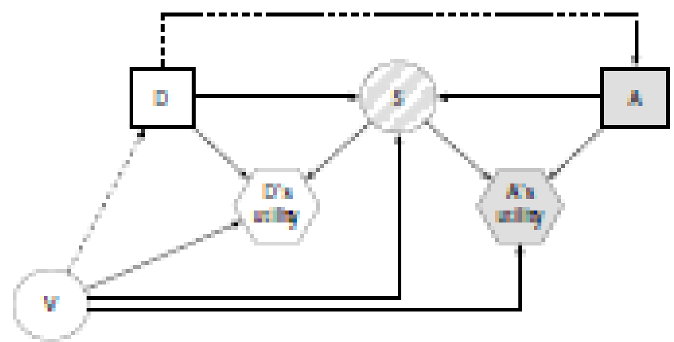
(b) Sim D-A



(c) Seq A-D

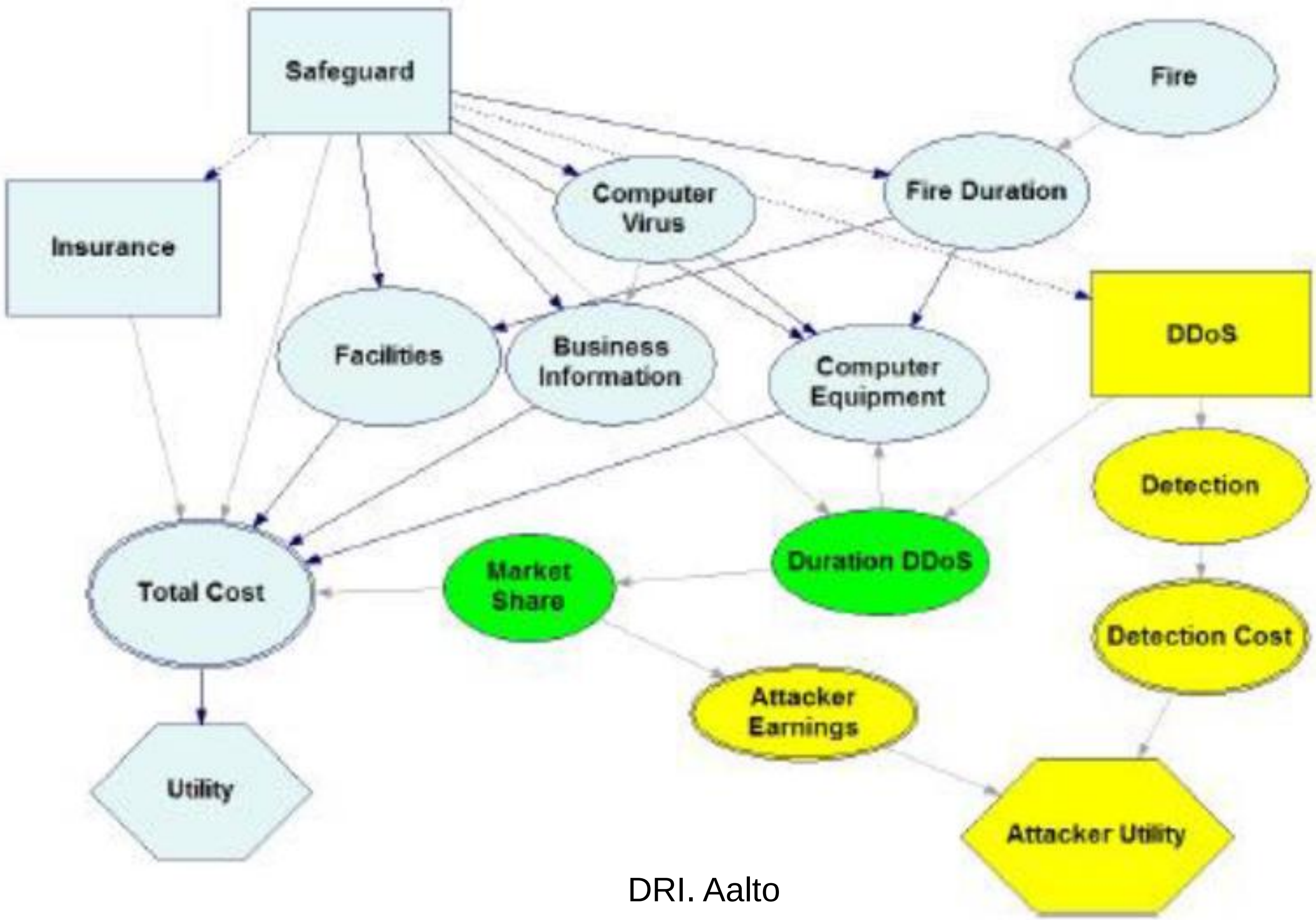


(d) Seq D-A-D

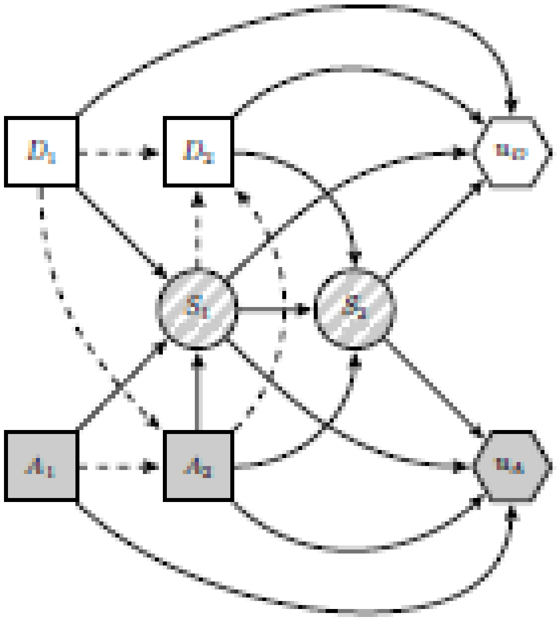


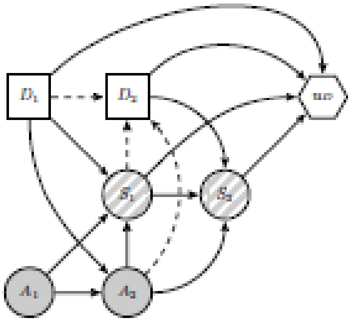
(e) Seq D-A-PI

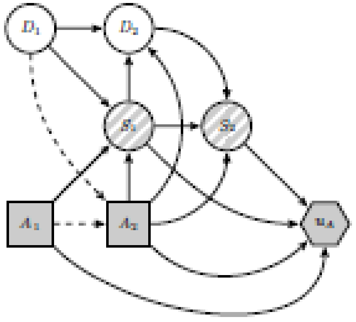
ARA templates

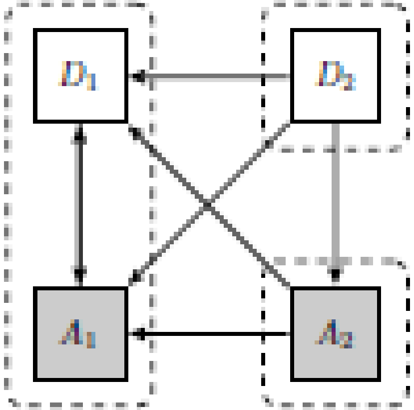


More general interactions









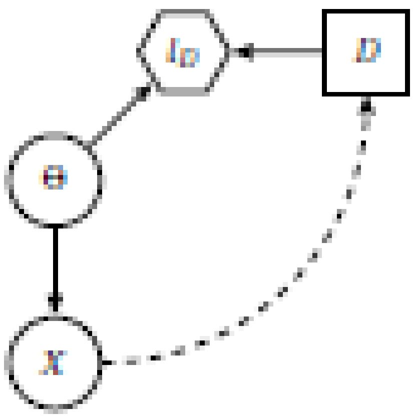
More general interactions

Algorithm 2 **General computational strategy. Cyclic case**

Data: BAID $\mathcal{B}$; a topological ordering $\mathcal{N}_1, \ldots, \mathcal{N}_r$ of the component graph derived from the relevance graph for $\mathcal{B}$; the associated IDs for defender $\mathcal{D}$ and attacker $\mathcal{A}$; the decision sequences $D_1, \ldots, D_m$ and $A_1, \ldots, A_n$, respectively, relative to $\mathcal{D}$ and $\mathcal{A}$.

For $i = 1$ **to** r **do**
  While $\mathcal{N}_i \cap D^{\mathcal{A}} \neq \emptyset$ **do**
    Find $j = \max\{k \mid A_k \in \mathcal{N}_i\}$.
    While $A_j \in \mathcal{A}$ **do**
      Apply Algorithm A.1 to $\mathcal{A}$ using A-reductions.
    End While
  End While
  While $\mathcal{N}_i \neq \emptyset$ **do**
    Find $j = \max\{k \mid D_k \in \mathcal{N}_i\}$.
    While $D_j \in \mathcal{D}$ **do**
      Apply Algorithm A.1 to $\mathcal{D}$ using D-reductions.
    End While
  End While
End For

Statistical Decision Theory



$$d^*(x) = \arg\min_d \int l_D(d,\theta)\, p_D(\theta \mid x)\, \mathrm{d}\theta.$$

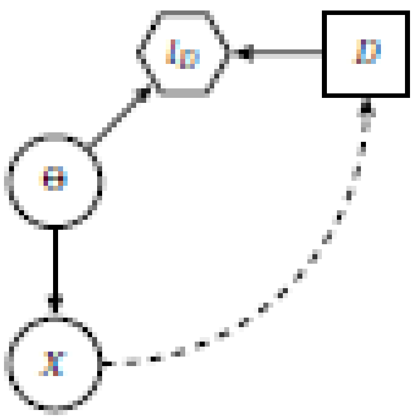
$$d^*(x) = \arg\min_d \int l_D(d,\theta)\, p_D(x \mid \theta)\, p_D(\theta)\, \mathrm{d}\theta.$$

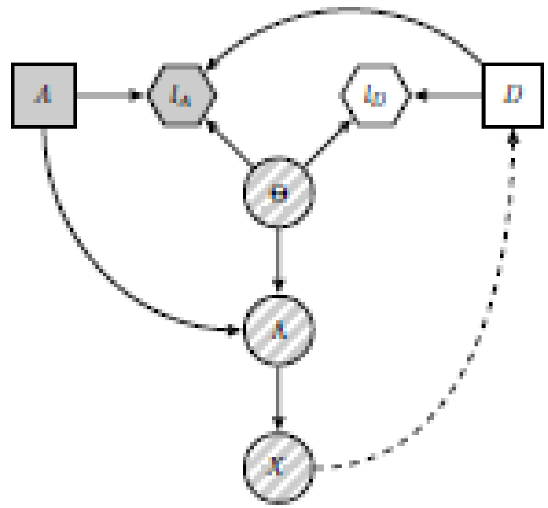
- Point estimation under quadratic loss

$$l_D(d,\theta) = (\theta - d)^2,$$

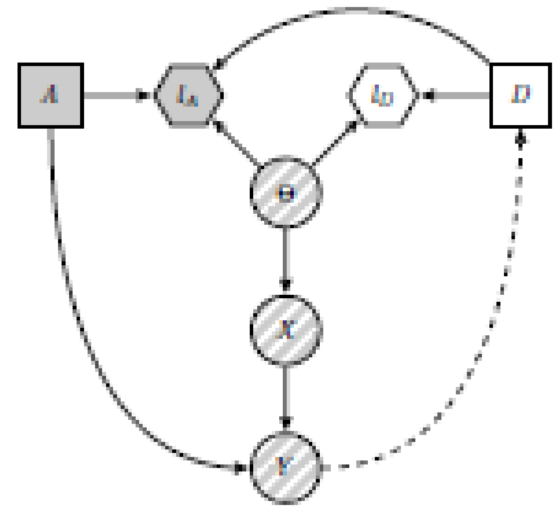
$$d^*(x) = \frac{1}{p_D(x)} \int \theta\, p_D(x \mid \theta)\, p_D(\theta)\, \mathrm{d}\theta = \int \theta\, p_D(\theta \mid x)\, \mathrm{d}\theta = E\left[\theta \mid x\right]$$

Adversarial Statistical Decision Theory

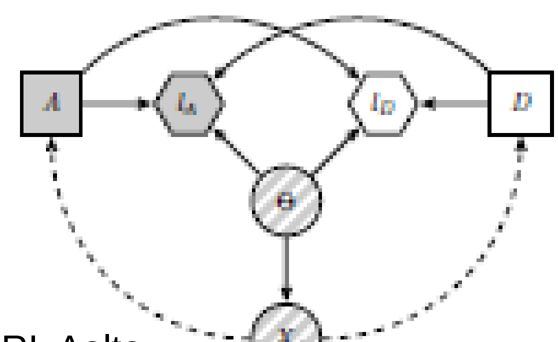




(a) Structural attacker



(b) Data-fiddler attacker

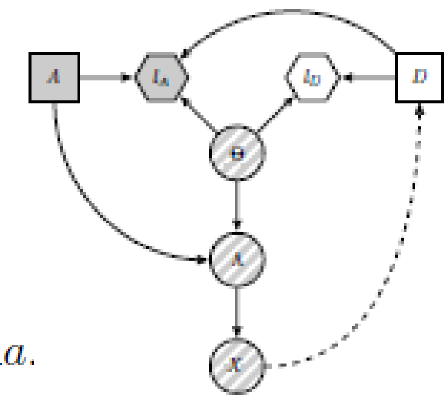


(c) Simultaneous ASDT problem

Adversarial point estimation

$$\lambda = a + \theta$$

- Quadratic loss



(a) Structural attacker

$$d^*(x) = \arg\min_d \iiint (\theta - d)^2 \, p_D(x \,|\, \lambda) \, p_D(\lambda \,|\, \theta, a) \, p_D(\theta) \, p_D(a) \, \mathrm{d}\lambda \, \mathrm{d}\theta \, \mathrm{d}a.$$

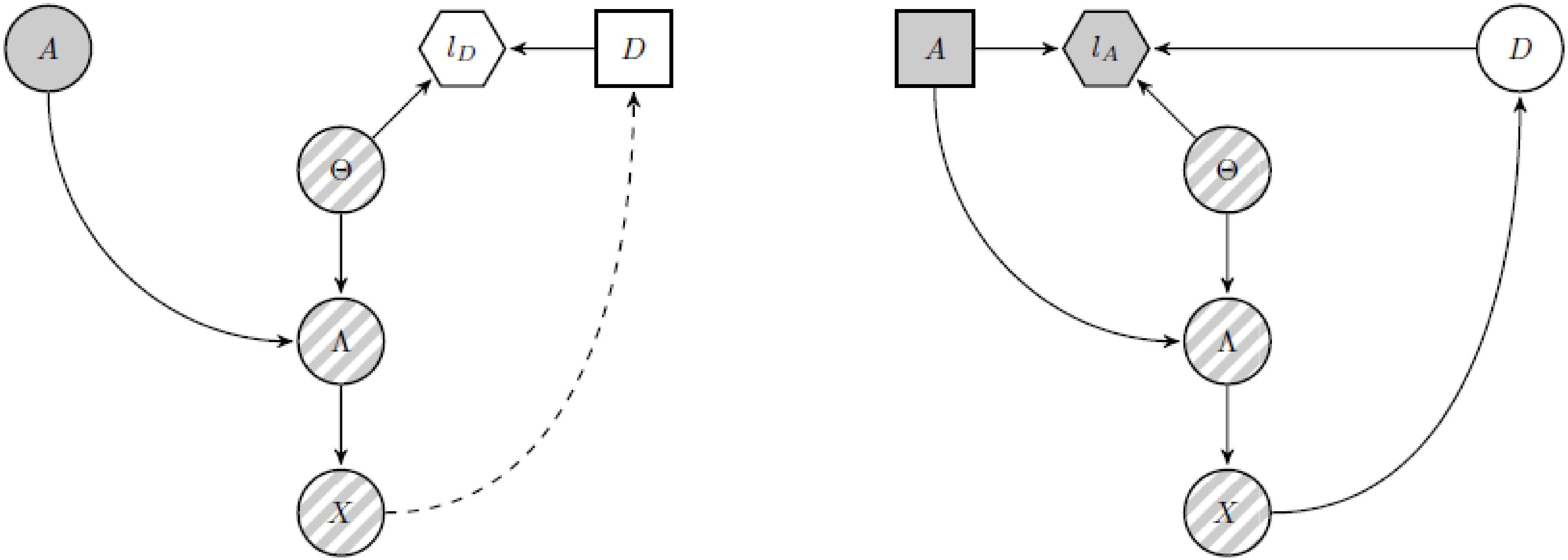
$$d^*(x) = \arg\min_d \iint (\theta - d)^2 \, p_D(x \,|\, \lambda = \theta + a) \, p_D(\theta) \, p_D(a) \, \mathrm{d}\theta \, \mathrm{d}a$$

$$d^*(x) = \frac{1}{p_D(x)} \iiint \theta \, p_D(x \,|\, \lambda) \, p_D(\lambda \,|\, \theta, a) \, p_D(\theta) \, p_D(a) \, \mathrm{d}\lambda \, \mathrm{d}\theta \, \mathrm{d}a$$

$$\begin{aligned} d^*(x) &= \frac{1}{p_D(x)} \iint \theta \, p_D(x \,|\, \lambda) \, p_D(\lambda \,|\, \theta) \, p_D(\theta) \, \mathrm{d}\lambda \, \mathrm{d}\theta \\ &= \frac{1}{p_D(x)} \int \theta \, p_D(x \,|\, \theta) \, p_D(\theta) \, \mathrm{d}\theta = \int \theta \, p_D(\theta \,|\, x) \, \mathrm{d}\theta = E\left[\theta \,|\, x\right] \end{aligned}$$

$$p_D(a)$$

Adversarial point estimation



Adversarial point estimation

- A Bayesian adversary

$$a_B^* = \arg\min_a \iiint l_A(d, a, \theta)\, p_A(d \mid x)\, p_A(x \mid \lambda = \theta + a)\, p_A(\theta)\, \mathrm{d}d\, \mathrm{d}x\, \mathrm{d}\theta.$$

$$A_B^* = \arg\min_a \iiint L_A(d, a, \theta)\, P_A(d \mid x)\, P_A(x \mid \lambda = \theta + a)\, P_A(\theta)\, \mathrm{d}d\, \mathrm{d}x\, \mathrm{d}\theta$$

$$p_D^B(a) = P(A_B^* = a).$$

$$A_{B,k}^* = \arg\min_a \iiint L_A^k(d, a, \theta)\, P_A^k(d \mid x)\, P_A^k(x \mid \lambda = \theta + a)\, P_A^k(\theta)\, \mathrm{d}d\, \mathrm{d}x\, \mathrm{d}\theta$$

$$\hat{p}_D^B(A = a) \approx \#\{A_{B,k}^* = a\}/K$$

- Mixture, e.g.

$$\pi_B\, \hat{p}_D^B(a) + \pi_M\, \hat{p}_D^M(a)$$

Adversarial point estimation

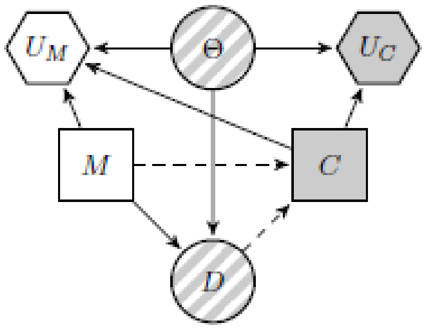
- Normal-normal model, for certain parameter choices

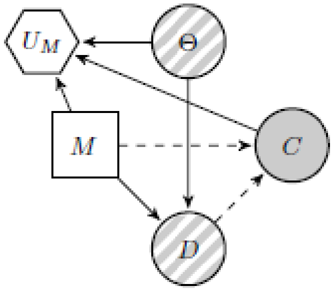
Solution Concept	Optimal Solution
Non-adversarial	$\dfrac{4\sum_{i=1}^{n} x_i}{4n+1}$
ARA: Minimax adversary	$\dfrac{4\sum_{i=1}^{n} x_i}{4n+1}$
ARA: Bayesian adversary	$\dfrac{4\left(0.318\,\xi(x,0)\sum_{i=1}^{n} x_i + 0.682\,\xi(x,1)\sum_{i=1}^{n}(x_i-1)\right)}{\left(0.318\,\xi(x,0)+0.682\,\xi(x,1)\right)(4n+1)}$
ARA: Uncertain concept	$\dfrac{4\left(0.545\,\xi(x,0)\sum_{i=1}^{n} x_i + 0.455\,\xi(x,1)\sum_{i=1}^{n}(x_i-1)\right)}{\left(0.545\,\xi(x,0)+0.455\,\xi(x,1)\right)(4n+1)}$

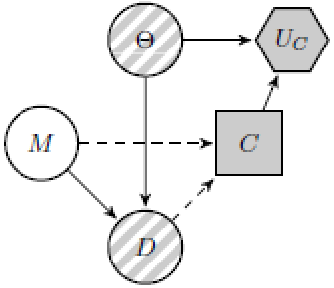
$$\xi(x,a) = \exp\left(\frac{\frac{\left(\mu_D\,\rho_D^2+\sigma_D^2\sum_{i=1}^{n}(x_i-a)\right)^2}{\rho_D^2+n\,\sigma_D^2} - \sigma_D^2\sum_{i=1}^{n}(x_i-a)^2}{2\rho_D^2\sigma_D^2}\right)$$

Adversarial reliability

- Acceptance sampling



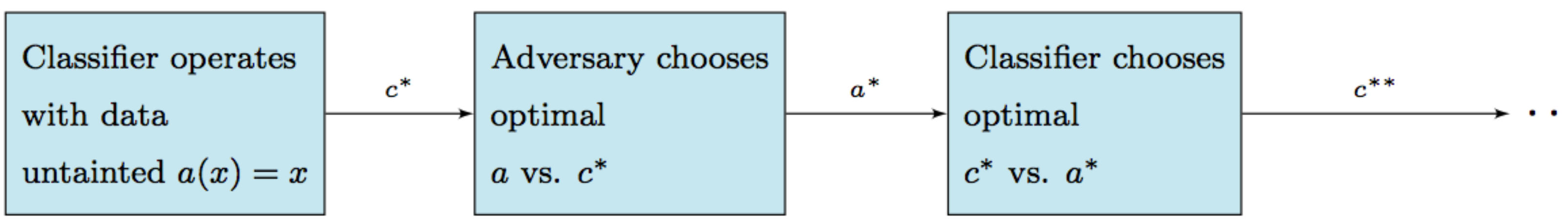




Adversarial classification as a game

- C, classifier. A, adversary
- Two classes: + malicious; - innocent.
- C and A maximise expected utility under common knowledge conditions
- Finding Nash equilibria extremely complex

- Dalvi et al (2004) propose a scheme



Utility sensitive Naive Bayes

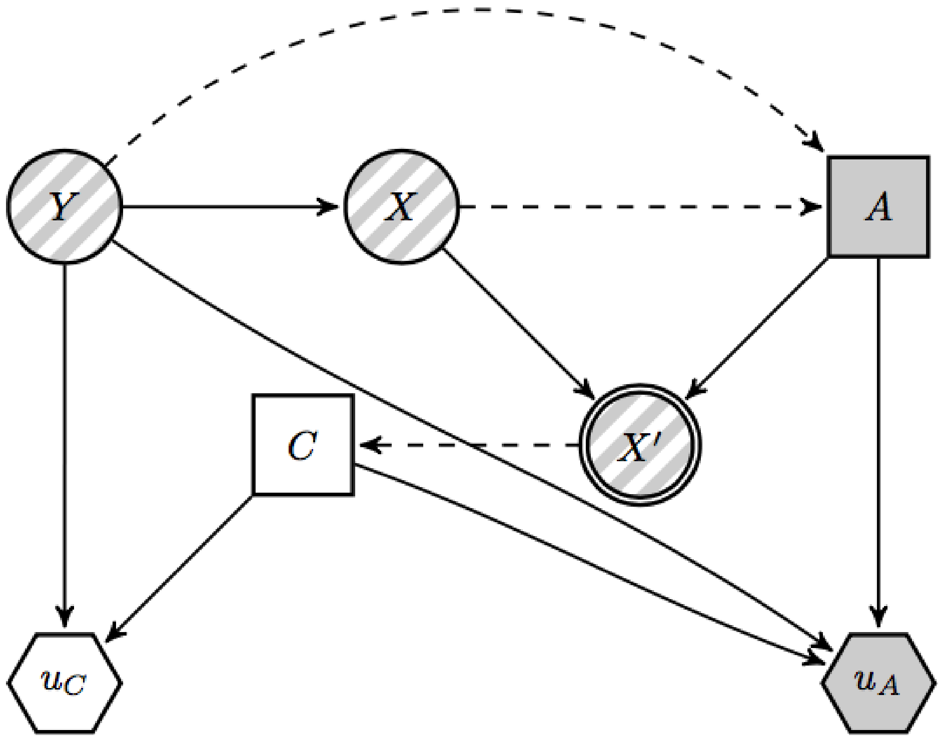
Forward myopic approach under strong common knowledge

Adversarial problems

- Adversarial classification (Dalvi et al,…)
- Adversarial signal processing (Barni et al,..)
- Adversarial learning (Lowd and Meek,..)
- Adversarial machine learning (Tygar,..)
- Adversarial SVMs (Zhou et al,…)
- …

Adversarial problems

Adversarial classification through ARA. ACRA



Dalvi et al's pioneer AC model from ARA perspective

malicious (+) or innocent (−)

ACRA. Classifier problem

$$\begin{aligned} c(x') &= \arg\max_{y_C} \sum_{y\in\{+,-\}} u_C(y_C, y) p_C(y|x') = \arg\max_{y_C} \sum_{y\in\{+,-\}} u_C(y_C, y) p_C(y) p_C(x'|y) = \\ &= \arg\max_{y_C} \sum_{y\in\{+,-\}} u_C(y_C, y) p_C(y) \sum_{x\in\mathcal{X}'} \sum_{a\in\mathcal{A}(x)} p_C(x', x, a|y). \end{aligned}$$

....

$$= \arg\max_{y_C} \left[u_C(y_C, +) p_C(+) \sum_{x\in\mathcal{X}'} p_C(a_{x\to x'}|x, +) p_C(x|+) + u_C(y_C, -) p_C(x'|-) p_C(-) \right]$$

ACRA. Adversary problem

$$a^*(x,y) = \arg\max_a \int \left[u_A(c(a(x)) = +, y, a) \cdot p + u_A(c(a(x)) = -, y, a) \cdot (1-p) \right] f_A(p|a(x))dp.$$

$$\int \left[u_A(+,+,a)\ p + u_A(-,+,a)\ (1-p) \right] f_A(p|a(x))dp =$$
$$= [u_A(+,+,a) - u_A(-,+,a)] p^A_{a(x)} + u_A(-,+,a).$$

$$A^*(x,+) = \arg\max_a \left([U_A(+,+,a) - U_A(-,+,a)]\, P^A_{a(x)} + U_A(-,+,a) \right)$$

random version
of

$$p_C(a|x,+) = Pr(A^*(x,+) = a)$$

$$p^A_{a(x)} = \int p f_A(p|a(x))dp$$

$$P_A(c|x') \sim \beta e(\delta_1, \delta_2)$$

$$\frac{\delta_1}{\delta_1 + \delta_2} = Pr_A(c(x') = +)$$

ACRA. Spam detection approach

1. PREPROCESSING

   Train a probabilistic classifier to estimate $p_C(y)$ and $p_C(x|y)$, assuming that the training set has not been tainted.

2. OPERATION

   Read x'.

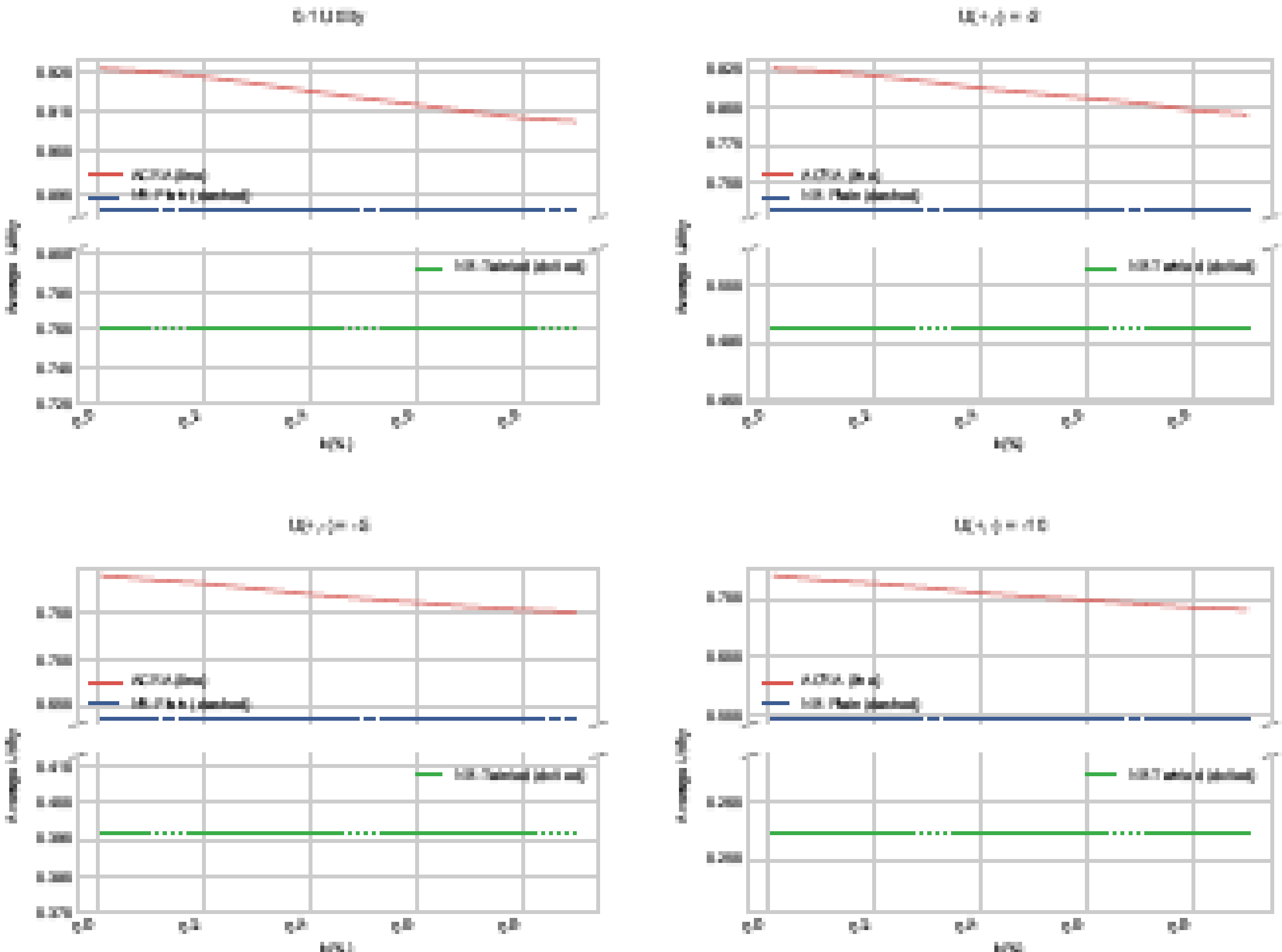
   ESTIMATE $p_C(a_{x \to x'}|x, +)$.

   Solve

$$c(x') = \arg\max_{y_C} \left[u(y_C, +)\widehat{p}_C(+) \sum_{x \in \mathcal{X}'} \widehat{p}_C(a_{x \to x'}|x, +)\widehat{p}_C(x|+) + u(y_C, -)\widehat{p}_C(x'|-)\widehat{p}_C(-) \right].$$

   Output $c(x')$.

ACRA. Spam detection approach



ACRA. Computational enhancements

$$= \arg\max_{y_C} \left[u_C(y_C,+) p_C(+) \sum_{x \in \mathcal{X}'} p_C(a_{x \to x'} | x, +) p_C(x|+) + u_C(y_C,-) p_C(x'|-) p_C(-) \right]$$

Note first that the optimization problem (1) may be reformulated as setting $c(x') = +$ if and only if $\sum_{x \in \mathcal{X}'} p_C(a_{x \to x'} | x, +) p_C(x|+) > t$, where

$$t = \frac{\left[u_C(-,-) - u_C(+,-) \right] p_C(x'|-) p_C(-)}{\left[u_C(+,+) - u_C(-,+) \right] p_C(+)}.$$

$$I = \frac{1}{N} \sum_n p_C(a_{x_n \to x'} | x_n, +) I(x_n \in \mathcal{X}') > t.$$

Importance sampling. Sequentially decide

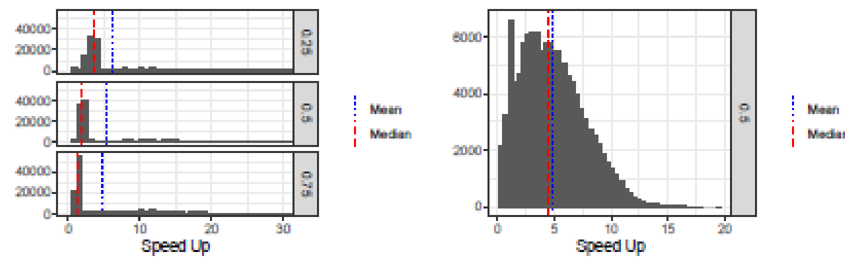
estimation of $p_C(a_{x \to x'} | x, +)$ Small Monte Carlo sample size

$$\widehat{p}_C(a_{x \to x'} \,|\, x, +) \simeq \frac{\#\{a_k^* = a_{x \to x'}\} + 1}{K + |(\mathcal{A}(x))|}.$$

Regression Metamodel
Parallel processing

ACRA computational enhancements

	Size	Accuracy	FPR	FNR
ACRA	1.00	0.919	$1.87 \cdot 10^{-2}$	$1.77 \cdot 10^{-1}$
MC ACRA	0.75	0.912	$3.20 \cdot 10^{-2}$	$1.74 \cdot 10^{-1}$
MC ACRA	0.50	0.905	$2.70 \cdot 10^{-2}$	$1.99 \cdot 10^{-1}$
MC ACRA	0.25	0.885	$2.09 \cdot 10^{-2}$	$2.60 \cdot 10^{-1}$
NB-Plain	-	0.886	$6.77 \cdot 10^{-2}$	$1.85 \cdot 10^{-1}$
NB-Tainted	-	0.761	$6.77 \cdot 10^{-2}$	$5.00 \cdot 10^{-1}$



(a) (b)

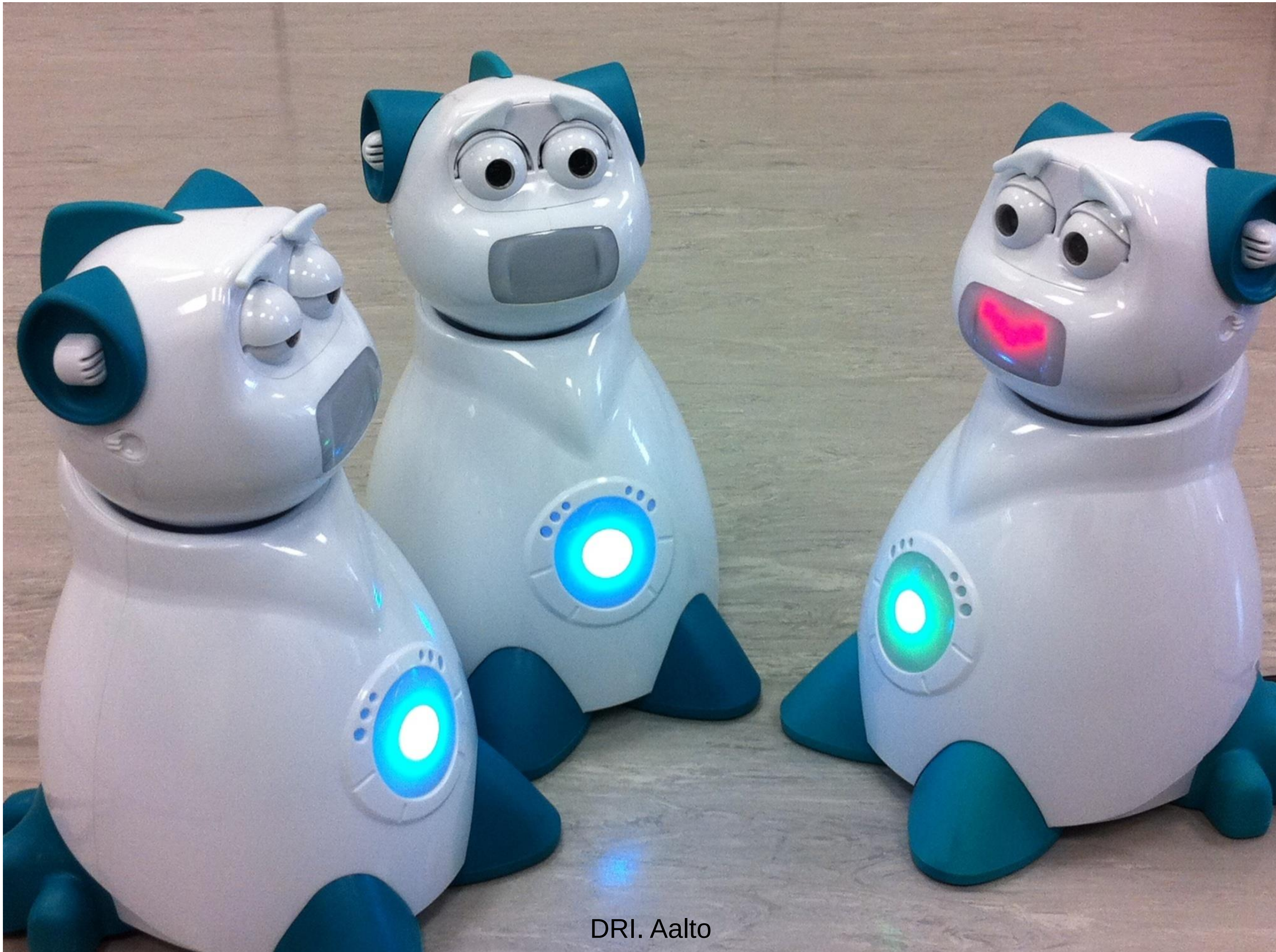
Size	Mean	Median
0.25	6.20	3.69
0.50	5.30	2.00
0.75	4.86	1.31

	Dataset	Size	Accuracy	FPR	FNR
MC ACRA	UCI	0.5	0.904	$3.69 \cdot 10^{-2}$	$1.87 \cdot 10^{-1}$
NB-Plain	UCI	-	0.887	$6.56 \cdot 10^{-2}$	$1.87 \cdot 10^{-1}$
NB-Tainted	UCI	-	0.724	$6.56 \cdot 10^{-2}$	$6.01 \cdot 10^{-1}$
MC ACRA	Enron-Spam	0.5	0.824	$1.32 \cdot 10^{-1}$	$3.05 \cdot 10^{-1}$
NB-Plain	Enron-Spam	-	0.721	$2.83 \cdot 10^{-1}$	$2.68 \cdot 10^{-1}$
NB-Tainted	Enron-Spam	-	0.534	$2.83 \cdot 10^{-1}$	1.00
MC ACRA	Ling-Spam	0.5	0.958	$3.90 \cdot 10^{-2}$	$5.68 \cdot 10^{-2}$
NB-Plain	Ling-Spam	-	0.957	$4.00 \cdot 10^{-2}$	$5.75 \cdot 10^{-2}$
NB-Tainted	Ling-Spam	-	0.800	$4.00 \cdot 10^{-2}$	1.00

Table 3: Comparison between MC ACRA and NB under 2-GWI attacks.

ARA vs GT

- Provide different solutions
- Dominance and ARA
- ‘Iterated dominance’ and ARA
- Ficticious play and ARA
- Level-k and ARA
- GT, Sensitivity analysis, If sensitive, ARA.
- Different types of adversaries



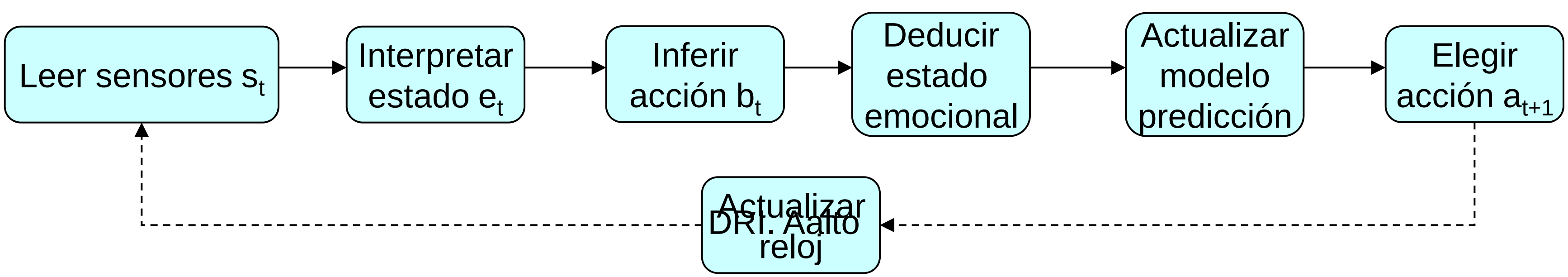
DRI. Aalto

Problem

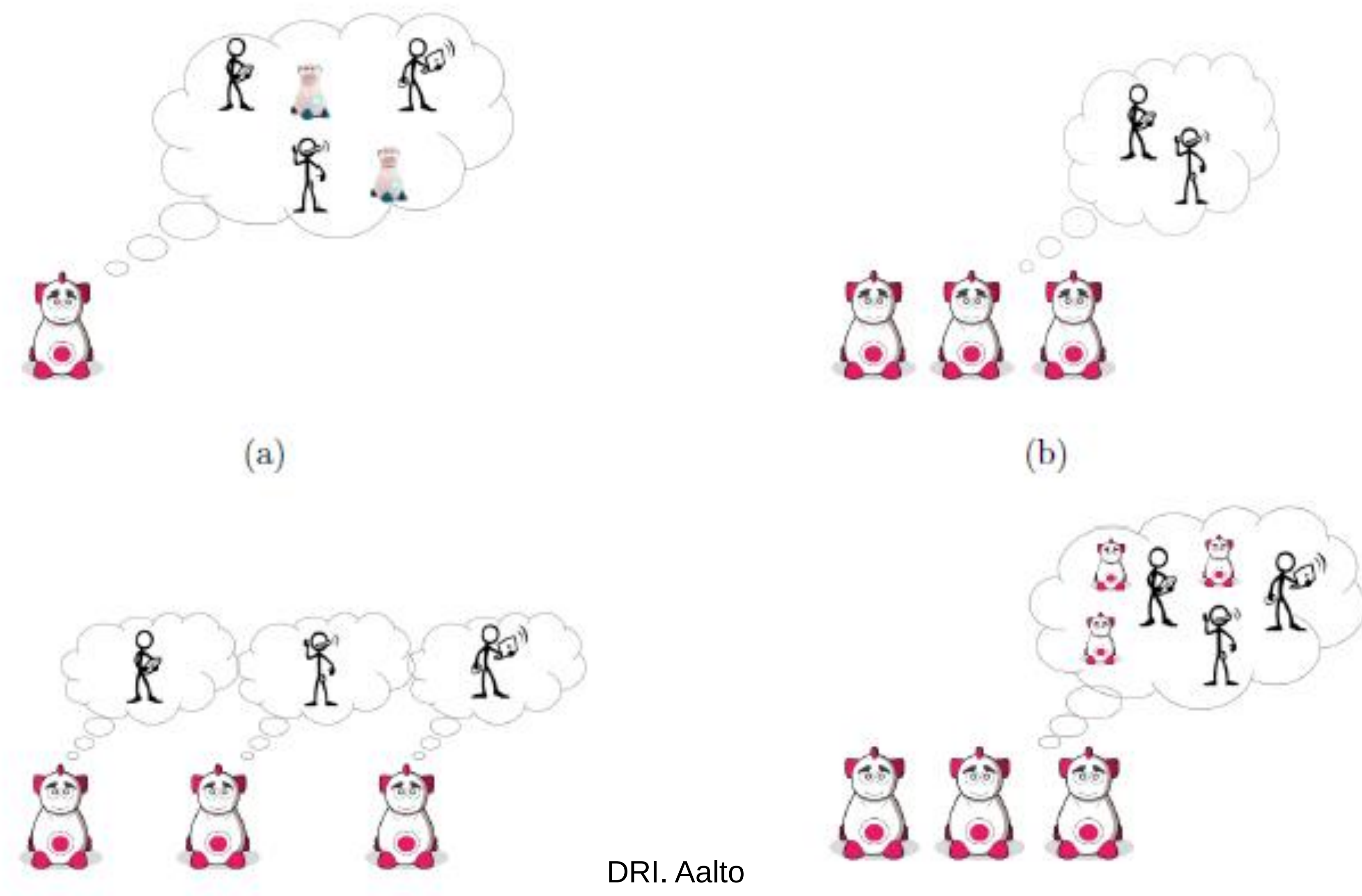
- An agent makes decisions in a finite set
- Has sensors providing information around it
- It relates with a user which makes decisions
- They’re both within an environment which evolves (under the control of the user)

Basic framework

$$\max_{a_t \in \mathcal{A}} \psi(a_t) = \sum_{b_t, e_t} u(a_t, b_t, e_t) \times p(b_t, e_t \mid a_t, (a_{t-1}, b_{t-1}, e_{t-1}), (a_{t-2}, b_{t-2}, e_{t-2}))$$



Basic framework



Single-stage computational schemes

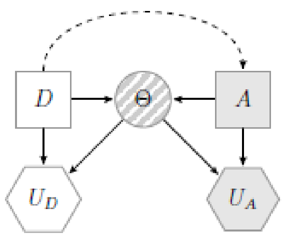
- Augmented probability simulation

$$\max_{x \in \mathcal{X}} \int u(x,\theta) p(\theta|x)\, \mathrm{d}\theta,$$

$$\pi(x,\theta) \propto u(x,\theta) p(\theta|x).$$

$$\pi(x) \propto \int u(x,\theta) p(\theta|x)\, \mathrm{d}\theta.$$

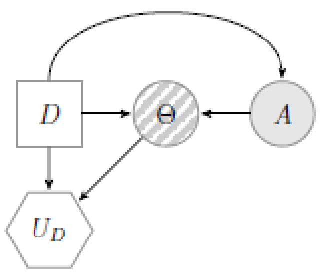
Single-stage computational schemes



$$\psi_D(d) = \int \psi_D(a,d) p_D(a|d)\,\mathrm{d}a = \int \left[\int u_D(d,\theta) p_D(\theta|d,a)\,\mathrm{d}\theta \right] p_D(a|d)\,\mathrm{d}a.$$

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

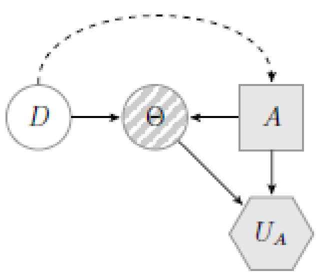
$$p_D(a|d) = \mathbb{P}_F\left[a = \arg\max_{x \in \mathcal{A}} \Psi_A(x,d) \right]$$



$$\Pi_A(a, \theta_A|d) \propto U_A(a,\theta_A) P_A(\theta_A|a,d).$$

$$A^*(d) = \text{mode}\,(\Pi_A(a|d))$$

$$\pi_D(d, a, \theta_D) \propto u_D(d, \theta_D)\ p_D(\theta_D|a,d)\ p_D(a)$$



$$d^* = \text{mode}\,(\pi_D(d))$$

input: N, M, J

for $d \in \mathcal{D}$ **do**

- **for** $j = 1$ **to** J **do**
  - Sample U_A^j, P_A^j and define Π_A^j
  - Initialize θ_A^0
  - **for** $i = 1$ **to** M **do**
    - Sample $a^{(i)}$ from $\Pi_A^j(a|\theta_A^{(i-1)}, d)$
    - Sample $\theta^{(i)}$ from $\Pi_A^j(\theta_A|a^{(i)}, d)$
  - Estimate a_j^* as mode of $\{a^{(i)}\}$
- Estimate $p_D(a|d)$ from $\{a_j^*\}$

Initialize $(d^{(0)}, \theta_D^{(0)})$

for $i = 1$ **to** N **do**

- Draw $d^{(i)}$ from $\pi_D(d|a^{(i-1)}, \theta_D^{(i-1)})$
- Draw $\theta_D^{(i)}$ from $\pi_D(\theta_D|a^{(i-1)}, d^{(i-1)})$
- Draw $a^{(i)}$ from $\pi_D(a|d^{(i)}, \theta_D^{(i)})$

Estimate d^* as mode of $\{d^{(i)}\}$

Discussion

- Traditional statistical/ML/risk analysis problems perturbated by presence of adversaries
- Traditionally treated from a game theoretic perspective (common knowledge)
- An ARA approach to mitigate common knowledge
- Different opponent models, beyond SEU
- Concept uncertainty, Mixtures
- Robustness and ARA (GT, ARA, Robust ARA)

Other themes

- Differential games
- Multiagent reinforcement learning
- Competition and cooperation
- Cybersecurity and cyberinsurance: CYBECO
- Efficient computational schemes
- Computational environment
- Fake news
- Malware detection
- Attacker models
- Generative adversarial networks
- Generic approach: point estimation, interval estimation,…
- Multiple attackers, Multiple defenders

Thanks!!!

Collabs welcome

david.rios@icmat.es

SPOR DataLab https://www.icmat.es/spor/

Aisoy Robotics https://www.aisoy.com

It’s a risky life @YouTube

CYBECO https://www.cybeco.eu/