Towards an Understanding of Human Persuasion and Biases in Argumentation

Pierre Bisquert, Madalina Croitoru, Florence Dupin de Saint-Cyr, Abdelraouf Hecham

INRA & LIRMM & IRIT, France

July 6th 2016

Objectives

- Why are "good" arguments not persuasive?
- Why are "bad" arguments persuasive?
- How can we prevent these negative processes?

$\Rightarrow$ **General aim**: improve the quality of collective decision making

Persuasion in AI

- **Interactive technologies for human behavior**
  - Persuade humans in order to change behaviors [Oinas-Kukkonen, 2013]
  - $\Rightarrow$ Health-care [Lehto and Oinas-Kukkonen, 2015], environment [Burrows et al., 2014]
- **Dialogue protocols for persuasion**
  - Derived from logic and philosophy [Hamblin, 1970], [Perelman and Olbrechts-Tyteca, 1969]
  - $\Rightarrow$ Ensure rational interactions between agents [Prakken, 2006]
- **Argumentation theory**
  - Abstract and logical argumentation [Dung, 1995], [Besnard and Hunter, 2001]
  - $\Rightarrow$ Dynamics and enforcement [Baumann and Brewka, 2010], [Bisquert et al., 2013]
- etc.

Our Approach

- **Our approach**: how does it "work"?
- Link between persuasion and cognitive biases [Clements, 2013]
  - Computational analysis of cognitive biases

$\Rightarrow$ Explain why an argument has been persuasive or not

$\Rightarrow$ Understand better human persuasion processes

$\Rightarrow$ (Hopefully) Allow people to prevent manipulation attempts

Outline

Dual Process Theory

- Based on the work of **Kahneman** (and **Tversky**) [Tversky and Kahneman, 1974]
- **System 2** (S2)
  - Conscious, thorough and slow process
  - Expensive and "rational" reasoning
- **System 1** (S1)
  - Instinctive, heuristic and fast process
  - Cheap and based on associations
- Biases (generally) arise when S1 is used
  - fatigue, interest, motivation, ability, lack of knowledge

Our take on S1 & S2

- S2 is a **logical knowledge base**
  - *Beliefs*
    - ★ "Miradoux is a wheat variety", "wheat contains proteins"
  - *Opinions*
    - ★ "I like Miradoux", "I do not like spoiled wheat"
- S1 is represented by **special rules**
  - "*PastaQuality* is associated to *Italy*"
- Biases arise when S1 rules are used instead of S2 rules
  - **Cognitive availability**

But how do we build them?

- **Knowledge base**: Datalog +/- ([Arioua et al., 2015])
  - "Miradoux is a wheat variety": $\boldsymbol{wheat}(\boldsymbol{miradoux})$
  - "Wheat contains proteins": $\forall \boldsymbol{X}\ \boldsymbol{wheat}(\boldsymbol{X}) \rightarrow \boldsymbol{proteins}(\boldsymbol{X})$
  - "I like Miradoux": $\boldsymbol{like}(\boldsymbol{miradoux})$

  $\Rightarrow$ Denoted $\boldsymbol{BO}$

- **Associations**: obtained thanks to a *Game With A Purpose*
  - Allows to extract associations for different profiles
  - Associations are (manually) transformed
  - ($PastaQuality$, $Italy$): $\forall \boldsymbol{X}\ \boldsymbol{highQualityPasta}(\boldsymbol{X}) \rightarrow \boldsymbol{madeInItaly}(\boldsymbol{X})$

  $\Rightarrow$ Denoted $\boldsymbol{A}$

- Each rule has a particular cognitive effort
  - function $\boldsymbol{e}$

Example

BO

B_1 :	$wheat(miradoux)$	10
B_2 :	$spoiled\_wheat(miradoux2)$	10
B_3 :	$spoiled\_wheat(X) \rightarrow low\_protein(X)$	10
B_4 :	$low\_protein(X) \wedge has\_protein(X) \rightarrow \bot$	10
B_5 :	$wheat(X) \rightarrow has\_protein(X)$	10
B_6 :	$has\_protein(X) \rightarrow nutrient(X)$	10
O_1 :	$dislike(miradoux2)$	5
O_2 :	$like(X) \wedge dislike(X) \rightarrow \bot$	5

A

A_1 :	$nutrient(X) \rightarrow like(X)$	1
A_2 :	$has\_protein(X) \rightarrow dontcare(X)$	3

How do we reason?

Reasoning

- **Reasoning**: $K \vdash_R \varphi$, with R a sequence from $\boldsymbol{BO} \cup \boldsymbol{A}$
- Successive application of rules R: **reasoning path**

- $wheat(miradoux) \vdash_{R_1} like(miradoux)$, with $R_1 = \langle B_5, B_6, A_1 \rangle$:
  - B_5 : $wheat(X) \rightarrow has\_protein(X)$,
  - B_6 : $has\_protein(X) \rightarrow nutrient(X)$
  - A_1 : $nutrient(X) \rightarrow like(X)$,

  $\Rightarrow$ Total effort of R_1: 21
- $wheat(miradoux) \vdash_{R_2} dontcare(miradoux)$, with $R_2 = \langle B_5, A_2 \rangle$:
  - A_2 : $has\_protein(X) \rightarrow dontcare(X)$

  $\Rightarrow$ Total effort of R_2: 13

Cognitive Model

Definition

A **cognitive model** is a tuple $\kappa = (BO, A, e)$

- ***BO***: beliefs and opinions,
- ***A***: associations,
- ***e*** is a function $BO \cup A \to \mathbb{N} \cup \{+\infty\}$: effort required for each rule,

- Cognitive availability outside of the model

Outline

What is an argument?

Definition

An argument is a pair (φ, α) stating that having some beliefs and opinions described by φ leads to concluding α.

- *"Miradoux is a very good wheat variety since it contains proteins"*

$\Rightarrow$ $(\mathit{has\_protein}(\mathit{miradoux}), \mathit{like}(\mathit{miradoux}))$

How do we evaluate this argument?

Critical Questions

- CQ_1: $BO \cup A \cup \{\alpha\} \vdash \bot$? (is it possible to attack the conclusion?)
- CQ_2: $BO \cup A \cup \{\varphi\} \vdash \bot$? (is it possible to attack the premises?)
- CQ_3: $\varphi \vdash \alpha$? (does the premises allow to infer the conclusion?)

With argument (*has_protein*(*miradoux*), *like*(*miradoux*)):

- CQ_1: $BO \cup A \cup \{like(miradoux)\} \vdash \bot$
- CQ_2: $BO \cup A \cup \{has\_protein(miradoux)\} \vdash \bot$
- CQ_3: $has\_protein(miradoux) \vdash like(miradoux)$

Positive/Negative Answers

Proofs

Given a $CQ : h \vdash c$, a cognitive value cv and a reasoning path R:

$$proof_{ca}(R, CQ) \stackrel{\text{def}}{=} (eff(R) \leq cv \text{ and } h \vdash_R c)$$

where $eff(R) = \sum_{r \in R} e(r)$.

Positive/Negative Answers

Moreover, we say that:

- CQ is **answered positively** wrt to cv iff $\exists R$ s.t. $proof_{cv}(R, CQ)$, denoted $positive_{cv}(CQ)$,
- CQ is **answered negatively** wrt to cv iff $\nexists R$ s.t. $proof_{cv}(R, CQ)$, denoted $negative_{cv}(CQ)$.

Positive/Negative Answers – Example

BO	B_1 :	$wheat(miradoux)$	10
	B_2 :	$spoiled\_wheat(miradoux2)$	10
	B_3 :	$spoiled\_wheat(X) \rightarrow low\_protein(X)$	10
	B_4 :	$low\_protein(X) \wedge has\_protein(X) \rightarrow \bot$	10
	B_5 :	$wheat(X) \rightarrow has\_protein(X)$	10
	B_6 :	$has\_protein(X) \rightarrow nutrient(X)$	10
	O_1 :	$dislike(miradoux2)$	5
	O_2 :	$like(X) \wedge dislike(X) \rightarrow \bot$	5
A	A_1 :	$nutrient(X) \rightarrow like(X)$	1
	A_2 :	$has\_protein(X) \rightarrow dontcare(X)$	3

Argument $(has\_protein(miradoux), like(miradoux))$:

- CQ_1 is **answered negatively**:
  $\nexists R$ s.t. $BO \cup A \cup \{like(miradoux)\} \vdash_R \bot$
- CQ_3 is **answered positively** (with $cv \geq 21$):
  $has\_protein(miradoux) \vdash_{R_1} like(miradoux)$ with $R_1 = \langle B_5, B_6, A_1 \rangle$

Potential Status

Potential Status of Arguments

Given ca, we say that an argument is:

- ***acceptable*$_{ca}$** iff there is an allocation $c_1 + c_2 + c_3 = ca$ s.t. $negative_{c_1}(CQ_1)$, $negative_{c_2}(CQ_2)$, $positive_{c_3}(CQ_3)$
  - The agent may potentially accept the argument
- ***rejectable*$_{ca}$** iff $positive_{ca}(CQ_1)$ or $positive_{ca}(CQ_2)$ or $negative_{ca}(CQ_3)$.
  - The agent may potentially reject the argument

- An argument can be both $acceptable_{ca}$ and $rejectable_{ca}$
- How can we be more precise about the status?

Potential Status

Potential Status of Arguments

Given ca, we say that an argument is:

- ***acceptable*$_{ca}$** iff there is an allocation $c_1 + c_2 + c_3 = ca$ s.t. $negative_{c_1}(CQ_1)$, $negative_{c_2}(CQ_2)$, $positive_{c_3}(CQ_3)$
  - The agent may potentially accept the argument
- ***rejectable*$_{ca}$** iff $positive_{ca}(CQ_1)$ or $positive_{ca}(CQ_2)$ or $negative_{ca}(CQ_3)$.
  - The agent may potentially reject the argument

- An argument can be both $acceptable_{ca}$ and $rejectable_{ca}$
- How can we be more precise about the status?
  - Work in progress...
  - Reasoning tendency: preference relation over reasoning path

Outline

Summary

- Preliminary formalization of dual process theory and its link with human persuasion
- Proposition of a cognitive model acknowledging biases during argument evaluation
- Application on a real use case (Durum wheat knowledge base, implementation of a “GWAP”)

Perspectives

- Evaluation strategies
- Rationality properties
- Cognitive model update
- More elaborate logic of “beliefs and preferences”
- Empirical study

References I

Arioua, A., Buche, P., Croitoru, M., and Thomopoulos, R. (2015).
Using explanation dialogue for durum wheat knowledge base acquisition.
Technical report, UMR IATE, LIRMM, GraphIK, University of Montpellier.

Baumann, R. and Brewka, G. (2010).
Expanding argumentation frameworks: Enforcing and monotonicity results.
In *Proceeding of the 2010 conference on Computational Models of Argument: Proceedings of COMMA 2010*, pages 75–86, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Besnard, P. and Hunter, A. (2001).
A logic-based theory of deductive arguments.
Artificial Intelligence, 128(1-2):203–235.

Bisquert, P., Cayrol, C., Dupin de Saint Cyr Bannay, F., and Lagasquie-Schiex, M.-C. (2013).
Characterizing change in abstract argumentation systems.
In Simari, G. and Fermé, E., editors, *Trends in Belief Revision and Argumentation Dynamics*, pages 1–30. College Publications, http://www.collegepublications.co.uk/.

Burrows, R., Johnson, H., and Johnson, P. (2014).
Developing an online social media system to influence pro-environmental behaviour based on user values.
In *9th International Conference on Persuasive Technology, Extended Abstract*.

References II


Clements, C. S. (2013).
Perception and Persuasion in Legal Argumentation: Using Informal Fallacies and Cognitive Biases to Win the War of Words.
BYU Law Review, 319.


Dung, P. M. (1995).
On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games.
Artificial Intelligence, 77(2):321–358.


Hamblin, C. (1970).
Fallacies.
University paperback. Methuen.


Lehto, T. and Oinas-Kukkonen, H. (2015).
Explaining and predicting perceived effectiveness and use continuance intention of a behaviour change support system for weight loss.
Behaviour & Information Technology, 34(2):176–189.


Oinas-Kukkonen, H. (2013).
A foundation for the study of behavior change support systems.
Personal and Ubiquitous Computing, 17(6):1223–1235.

References III


Perelman, C. and Olbrechts-Tyteca, L. (1969).
The New Rhetoric: A Treatise on Argumentation.
University of Notre Dame Press.


Prakken, H. (2006).
Formal systems for persuasion dialogue.
Knowledge Engineering Review, 21(2):163–188.


Tversky, A. and Kahneman, D. (1974).
Judgment under uncertainty: Heuristics and biases.
Science, 185(4157):1124–1131.

GWAP

All Participants		Experts		Non-Experts	
Italy	⊕	Yellowness	⊕	Italy	⊕
Cooking time	⊙	Color	⊙	Cooking time	⊙
Taste	⊙	Protein Content	⊕	Price	⊙
Protein Content	⊕	Texture	⊕	Taste	⊙
Yellowness	⊕	Stickiness	⊕	Brand	⊙
Price	⊙	Cooking loss	⊖	Slow Sugar	⊕
Gluten	⊕	Drying Temperature	⊕	Tomato Sauce	⊕
Brand	⊙	Hydration	⊕	Panzanni	⊕