Substantive irrationality in cognitive systems

Pierre Bisquert[1] and **Madalina Croitoru**[2] and **Florence Dupin de Saint-Cyr**[3] and **Abdelraouf Hecham**[2]

Abstract. In this paper we approach both procedural and substantive irrationality of artificial agent cognitive systems and consider that when it is not possible for an agent to make a logical inference (too expensive cognitive effort or not enough knowledge) she might replace certain parts of the logical reasoning with mere associations.

1 INTRODUCTION

In artificial agents two kinds of biases have been highlighted ([8], [12], [14]). On one hand, the agent's beliefs and preferences may be incomplete and the agent may not know all the preferences or beliefs needed for complete reasoning (*e.g.* the agent's utility function is not available, or some constraints about the real world are not known). This kind of representational issued biases refers to the so called Type 1 irrationality or substantive irrationality that concerns the compliance of the results of reasoning with the agent's explicit goals and beliefs. For instance, a substantive irrational agent may eat fat while its rational goals and beliefs are in favor of healthy food. Type 2 irrationality, also known as procedural irrationality, concerns with the case when, due to the fact that computational resources (time or space available for representing and reasoning) are limited, the agent needs to make good choices in the process of deciding how to apply its efforts in reasoning. In this case what is rational for one agent is not rational for another with different limitations. Achieving procedural rationality means making rational choices about what inferences to perform, how to apply them, basically thinking about how to think. We investigate both substantive and procedural irrationality and build upon the model proposed in [4, 3]. We propose a more natural transition between two systems of reasoning: a logic based and an association based ones and propose a first cognitive model for substantive and procedural irrational agents that accounts for utterance acceptance in a logic encoding beliefs and preferences.

2 AGENT COGNITIVE MODEL

We define the cognitive model of an agent to contain beliefs, opinions, preferences and associations. The beliefs are represented using a finite set B of formulas taken from a propositional language $\mathcal{B}_\mathcal{L}$. We define an *opinion* about a belief $\varphi \in \mathcal{B}_\mathcal{L}$, denoted $\heartsuit\varphi$ (and resp. $\heartsuit\!\!\!\!\blacksquare\varphi$) as a constraint, that imposes to the situations where φ holds to be preferred (resp. strictly preferred) to the situations where φ does not hold, the opinions are gathered in a finite base $O \subset \mathcal{O}_\mathcal{L}$ where $\mathcal{O}_\mathcal{L}$ is the set of opinion formulas (that are either basic opinions or Boolean combination of them). A basic preference is a formula of the form $\alpha \trianglerighteq \beta$ (resp. $\alpha \triangleright \beta$) where $\alpha, \beta \in \mathcal{B}_\mathcal{L}$, interpreted as constraints on the preferences such that the situations where α holds should be preferred (resp. strictly preferred to) situations where β holds. Associations (elicited using [13]) encode Kahneman's System 1 [16], that is a human reasoning system dealing with quick, instinctive and heuristic thoughts. We denote by $\mathcal{A} = (\mathcal{B}_\mathcal{L} \cup \mathcal{O}_\mathcal{L} \cup \mathcal{P}_\mathcal{L}) \times (\mathcal{B}_\mathcal{L} \cup \mathcal{O}_\mathcal{L} \cup \mathcal{P}_\mathcal{L})$ the set of all possible associations between any pair of formulae. We also denote $\mathcal{B}_R$, $\mathcal{P}_R$, $\mathcal{O}_R$, $\mathcal{A}_R$ the sets of inference rules that allow us to deduce new beliefs, preferences, opinions and associations.

We define the notion of "reasoning" as the process of inferring a formula φ using a rule application sequence R from the set of logical, preference, opinion and association rules on an initial set of pieces of information K, denoted $K \vdash_R \varphi$. We call the successive application of rules R a "reasoning path". Inside this reasoning path we differentiate the use of logical inference rules from the use of an association rule. A reasoning on a formula can be achieved using different reasoning paths, each path has a cost depending on the cognitive effort needed to use the rules it contains. Intuitively it is less costly to use association rules than logical inference rules and among associations some are more or less far-fetched than others. In order to represent the cognitive effort involved by the reasoning, we are going to use the effort function e that associates an effort to the associations and the inference rules used.

A cognitive model is defined as a tuple of beliefs, opinions, preferences, associations and their subsequent effort for reasoning.

Definition 1 (Cognitive model) *A* cognitive model *is a tuple*

$$\kappa = (B, O, P, A, e, \sqsubseteq)$$

- $B \subseteq \mathcal{B}_\mathcal{L}$ *is a set of wff representing beliefs,*
- $O \subseteq \mathcal{O}_\mathcal{L}$ *is a set of wff representing opinions,*
- $P \subseteq \mathcal{P}_\mathcal{L}$ *is a set of wff representing preferences,*
- $A \subseteq \mathcal{A}$ *is a binary relation representing the associations between formulae,*
- e *is a function* $\mathcal{B}_R \cup \mathcal{P}_R \cup \mathcal{O}_R \cup \mathcal{A}_R \rightarrow \mathbb{N} \cup \{+\infty\}$ *that represents the effort required to infer with each inference rule.*
- $\sqsubseteq \ \subseteq \mathcal{R} \times \mathcal{R}$ *is a preference relation based on* e *over reasoning paths;* $R_1 \sqsubseteq R_2$ *means* R_1 *is better than* R_2.

3 Argument Evaluation

In our work, agents reason about implicative utterances [2] and more generally about enthymemes (see [5, 9]) or arguments.

Definition 2 (Argument) *Given* $\mathcal{L} = \mathcal{B}_\mathcal{L} \cup \mathcal{P}_\mathcal{L} \cup \mathcal{O}_\mathcal{L}$*, an argument* arg *is a pair* $(\varphi \in \mathcal{L}, \alpha \in \mathcal{L})$.

An argument (φ, α) intends to state that having some beliefs and preferences described by φ leads to concluding α. In argumentation literature, some works (such as *e.g.,* [17]) propose to base the decision about whether or not an argument is acceptable on some critical

[1] GraphIK, INRA, Montpellier, France
[2] GraphIK, LIRMM, Univ. Montpellier, France
[3] IRIT, Univ. Toulouse, France

questions. For the sake of generality, we propose to base the evaluation of arguments on the classical notions that are used in argumentation in order to explain "attacks" between arguments. Classically three notions are used, called rebuttal, undermine and undercut. More precisely an argument (φ, α) can be attacked either on its conclusion (α) directly or on a part of its premises (φ) or on the link between the premises and the conclusion.

- CQ_1: $(B, O, P, A) \vdash \neg\alpha$ (is it possible to attack the conclusion?)
- CQ_2: $(B, O, P, A) \vdash \neg\varphi$ (is it possible to attack the premises?)
- CQ_3: $\varphi \vdash \alpha$ (does the premises allow to infer the conclusion?)

To define what are the answers to critical questions we will use reasoning paths. Based on the ELM model [6] we suppose here that each agent has a cognitive availability that represents the maximum cognitive effort ca she is willing to make in order to reason on an argument.

Given an argument and a finite cognitive availability ca, we can compute all the possible reasoning paths wrt ca. A positive answer to a critical question corresponds to the existence of a reasoning path that requires a cognitive effort under ca. If there is no such path, the answer to the critical question is negative.

Definition 3 (Positive/negative answers)
Given an inference $CQ : h \vdash c$ and a cognitive availability ca, given a reasoning path R, we denote:
$proof_{ca}(R, CQ) \stackrel{def}{=}$ Eff(R) $\leq$ ca and $h \vdash_R c$ where Eff(R) $= \sum_{r \in R} e(r)$. Moreover, we say that:

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

Thanks to the previous definitions, we are in position to formally define the problem of argument evaluation wrt an agent cognitive model and its cognitive availability.

Definition 4 (Potential status of arguments) *Given an agent with a cognitive model $\kappa = (B, O, P, A, e, \sqsubseteq)$, a cognitive availability ca and an argument $arg = (\varphi, \alpha)$. Let $CQ_1 = B \cup O \cup P \cup A \vdash \neg\alpha$, $CQ_2 = B \cup O \cup P \cup A \vdash \neg\varphi$, $CQ_3 = \varphi \vdash \alpha$. We say that arg is:*

- acceptable$_{ca}$ *iff $\forall c_3 \leq ca$ s.t. $positive_{c_3}(CQ_3)$ and $\forall (c_1, c_2)$ s.t. $c_1 + c_2 + c_3 = ca$, we have $negative_{c_1}(CQ_1)$ and $negative_{c_2}(CQ_2)$.*
- rejectable$_{ca}$ *iff $positive_{ca}(CQ_1)$ or $positive_{ca}(CQ_2)$ or $negative_{ca}(CQ_3)$.*
- undecidable$_{ca}$ *if it is both acceptable$_{ca}$ and rejectable$_{ca}$.*

In other words, an argument is acceptable if the link between the premises and the conclusion can be established and the agent has not enough cognitive ability to find a counter-example for either the conclusion (CQ_1) or the premises (CQ_2). In order to be able to reject an argument it is enough to find a counterexample corresponding to one of the two first critical questions or to not have a sufficient cognitive ability to infer the causal link. An undecidable argument may be found if there is a proof for CQ_3 and for CQ_1 with a total cost above ca.

4 DISCUSSION AND RELATED WORK

The highly influential cognitive psychology work in dual systems ([16, 7, 11, 1, 10, 15]) associate such biases with two reasoning systems: one system that is slow but logically precise and another system that is fast but logically sloppy. The distinction does not make clear the interaction between biases due to logically flawed reasoning and biases due to sub optimal reasoning choices done because of cognitive limitations. This distinction is interesting when addressing the evaluation of biased argument.

In this paper we consider the problem of argument evaluation by agents that are both logically biased (*i.e.* may either reason exclusively logically or by combining logical reasoning with associations) and that have a limited cognitive availability. Following the highly influential cognitive psychology work in dual systems ([16, 7, 11, 1, 10, 15]) proposal considers that, when it is not possible for an agent to make a logical inference (too expensive cognitive effort or not enough knowledge), she might replace certain parts of the logical reasoning with mere associations. Using associations may lower the reasoning effort needed for argument evaluation and subsequently affect the argument acceptance.

ACKNOWLEDGEMENTS

The authors acknowledge the support of ANS1 1208 IATE INCOM INRA grant and ANR-12-0012 grant QUALINCA. We particularly thank Ulrike Hahn for her clear-sighted comments and the organizers of Dagstuhl Seminar 15221 on "Multi-disciplinary Approaches to Reasoning with Imperfect Information and Knowledge".

REFERENCES

[1] C. Beevers, 'Cognitive vulnerability to depression: A dual process model', *Clinical Psychology Review*, **25**(7), 975 – 1002, (2005).
[2] P. Besnard and A. Hunter, 'A logic-based theory of deductive arguments', *Artificial Intelligence*, **128**(1-2), 203 – 235, (2001).
[3] P. Bisquert, M. Croitoru, and F. Dupin de Saint-Cyr, 'Four ways to evaluate arguments according to agent engagement', in *Brain Informatics and Health*, 445–456, Springer, (2015).
[4] P. Bisquert, M. Croitoru, and F. Dupin de Saint-Cyr, 'Towards a dual process cognitive model for argument evaluation', in *Scalable Uncertainty Management*, 298–313, Springer, (2015).
[5] E. Black and A. Hunter, 'Using enthymemes in an inquiry dialogue system', in *Proc of the 7th Int. Conf. on Autonomous Agents and Multiag. Syst. (AAMAS 2008)*, pp. 437–444, (2008).
[6] J. Cacioppo and R. Petty, 'The elaboration likelihood model of persuasion', *Advances in Consumer Research*, **11**(1), 673–675, (1984).
[7] P. Croskerry, G. Singhal, and S. Mamede, 'Cognitive debiasing 1: origins of bias and theory of debiasing', *BMJ Quality & Safety*, **22**(Suppl 2), 58–64, (2013).
[8] J. Doyle, 'Rationality and its roles in reasoning', *Computational Intelligence*, **8**(2), 376–409, (1992).
[9] F. Dupin de Saint-Cyr, 'Handling enthymemes in time-limited persuasion dialogs', in *Int. Conf. on Scalable Uncertainty Management (SUM)*, number 6929 in LNAI, pp. 149–162. Springer-Verlag, (2011).
[10] S. Epstein, 'Integration of the cognitive and the psychodynamic unconscious.', *American Psychologist*, **49**(8), 709, (1994).
[11] J. Evans and J. Curtis-Holmes, 'Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning', *Thinking & Reasoning*, **11**(4), 382–389, (2005).
[12] I. Good, 'The probabilistic explanation of information, evidence, surprise, causality, explanation, and utility', *Foundations of statistical inference*, 108–141, (1971).
[13] A. Hecham, M. Croitoru, P. Bisquert, and P. Buche, 'Extending gwaps for building profile aware associative networks', in *Proceedings of ICCS 2016*, pp. 43–58, (2016).
[14] H. Simon, 'From substantive to procedural rationality', in *25 Years of Economic Theory*, 65–86, Springer, (1976).
[15] S. Sloman, 'The empirical case for two systems of reasoning.', *Psychological Bulletin*, **119**(1), 3, (1996).
[16] A. Tversky and D. Kahneman, 'Judgment under uncertainty: Heuristics and biases', *Science*, **185**(4157), 1124–1131, (1974).
[17] D. Walton, C. Reed, and F. Macagno, *Argumentation Schemes*, Cambridge University Press, Cambridge, 2008.