# Human Factors in Computer Decision-Making (Extended Abstract)

Dimitrios Antos
Harvard University
33 Oxford street 217
Cambridge, MA 02138
antos@fas.harvard.edu

## ABSTRACT

This thesis investigates whether incorporating ideas from human decision-making in computer algorithms may help improve agents' decision-making performance, as either independent actors or in collaboration with humans. For independent actors, psychological cognitive appraisal theories of emotion are used to develop a lightweight algorithm that dynamically re-prioritizes their goals to direct their attention. In experiments in quickly changing and highly uncertain domains these agents are shown to perform as well as agents that compute expensive optimal solutions, and exhibit robustness with respect to the parameters of the environment. For agents interacting with humans, it is investigated whether expressing emotions has the ability to convey traits like trustworthiness and skill, and whether the appropriate emotional expression can help forge mutually beneficial relationships with the human. Finally, the theory of reasoning patterns [7] is leveraged to analyze games and make it possible to answer questions about a system's strategic behavior without having to compute an expensive, precise solution. This theory is also employed to the generate advice for human decision-makers in complex games. This advice has been experimentally shown to improve their decision-making performance.

## Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Miscellaneous

## General Terms

PhD thesis extended abstract, reasoning patterns, emotions, decision-making

## Keywords

reasoning patterns, Bayesian games, game theory, emotions, decision-making, PhD thesis abstract

## 1. INTRODUCTION

Computer systems are being extensively used for decision-making in a variety of environments. Financial investments, military operations, auctions, prediction markets, scientific research and even digital entertainment heavily leverage artificial agents that perform computations and make decisions. In such systems humans are sometimes engaged in the decision-making process. Depending on the nature of this engagement, we can distinguish two types of systems: Agents in the first type act independently and without the need to interact with a human on a regular basis, if at all. In these cases, the decision-making algorithm lays entirely "within the agent." It aims to determine a course of action for the agent based on its preferences, goals and observations. The second type of agents is required to interact (negotiate, collaborate with, or assist) humans in carrying out their tasks. In doing so, the agent may also reason about the way humans make their decisions, their preferences and the way they might react, emotionally and cognitively, to its own behavior. An agent can of course be of both types, having to both make decisions autonomously and interact with humans.

Humans have been shown to leverage a variety of cognitive techniques, computational shortcuts and psychological/emotional components to make their decisions [**?**]. On the other hand, computer decision-making techniques do not as of yet incorporate an analogue of these emotion-based or cognitive techniques; it is an open question whether adding such capabilities would improve computer decision-making. It must here be noted that these methods used by humans are not necessarily "inferior" to the game-theoretic or logical reasoning frequently used by computers [6]. In particular, in quickly changing or highly uncertain environments the costly computation of optimal solutions may be less useful than quickly adapting to changes in the environment. Furthermore, when computers need to communicate with humans, the effectiveness of such interactions may be improved by providing the agents with the appropriate emotional expression and the ability to interpret and predict the humans' emotional responses and inferences. Below the contributions, realized and expected, to both types of decision-making agents are described.

## 2. HUMAN DECISION-MAKING FOR INDEPENDENT ACTORS

Independent actors need to make decisions autonomously, often in complex environments. However, real-world environments exhibit a prohibitively large number of states and complex interactions among the various agents, rendering optimal strategies impossible to compute and necessitating the use of heuristics. However, there is no principled methodology to generate heuristics in generic domains. I

**Cite as:** Human Factors in Computer Decision-Making (extended abstract), Dimitrios Antos, *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, Yolum, Tumer, Stone and Sonenberg (eds.), May, 2–6, 2011, Taipei, Taiwan, pp. 1337-1338.

have developed such a methodology by using cognitive appraisal theories of emotion. Emotions, under these theories, are cognitive reactions to particular interpretations of how perceived stimuli (observations) might influence the agent's goals. For instance, the emotion of "fear" is a reaction to a significant goal being perceived as coming under threat; fear then motivates behaviors geared toward protecting that goal (in animals, these behaviors might involve fleeing or adopting a defensive stance). In my architecture, agents are assumed to have goals, and each goal is associated with a priority level. At every point in time, agents are performing actions geared towards achieving higher-priority goals. Agents are also equipped with the ability to interpret the information they receive, assessing whether each of their goals is assisted or obstructed by new developments seen in the world. Artificial emotions are elicited in accordance with cognitive appraisal theories and change the goals' relative priority levels. Thus, the agent is switching its "attention" to the goals that its emotions are promoting as most significant. In simulations I am showing that agents using this lightweight, emotion-based heuristic methodology perform as well as agents that compute expensive solution concepts, and even perform reasonably well in domains for which optimal solutions are impossible to compute. Among the domains examined are restless bandits (an extension of multi-armed bandits), and foraging environments. The emotion-based agents have been compared against indexing policies, MDP solutions, as well as other, non-emotion-based heuristics in terms of the utility obtained, the amount of experience required to get the agent to an acceptable performance level, and the robustness of its performance with respect to the parameter values chosen in its algorithms.

## 3. HUMAN DECISION-MAKING FOR INTERACTING AGENTS

Agents interacting with humans are faced with not just the problems of effective, adaptive decision-making, but also with understanding and influencing the decision-making strategies of their human partners. For instance, agents negotiating with humans over the division of resources are able to secure better outcomes for themselves by understanding the socio-cognitive and emotional functions of their human opponents [8]. My work in this domain investigates whether emotion expressed by the agents may cause humans to perceive "traits" in the agent, such as trustworthiness, honesty, or skill. Furthermore, humans have been shown to develop "relationships" with the computers they interact with, treating them as social agents [5]. This thesis researches whether good, stable relationships can elicit better performance from both parties. This increased performance may manifest as reaching decisions quicker, making fewer mistakes, and maintaining repeated interactions even in the presence of errors due to the trust levels between the two parties. It is examined whether an agent generating "appropriate" emotional responses in its interaction with the human can assist the formation of such good relationships. If so, agents designed with the appropriate emotional expressions might enjoy a comparative advantage other agents in a market in which they compete for the humans' business.

Finally, in some domains humans are using computers to explore their options and understand the consequences of their decisions, but would prefer to retain the final call and the responsibility for their choices. In these settings, the computer needs to be able not just to compute a well-performing course of action, but also explain and justify it to the human. To this end, I have used the theory of reasoning patterns [7] to generate advice for human decision-makers. This theory exposes the reasons that make a particular strategy "good" in terms of its effects on the utility of the agents and the information flow within the game, thus offering explanations that are easy to understand by human decision-makers. To test whether this theory can be used for generating decision-making advice, I have used human subjects that played a repeated, private-information game whose size did not allow for easy computation of an optimal solution (Bayes-Nash equilibrium). Furthermore, this game had multiple equilibria, and thus it was not obvious which one should be suggested to the human. Large size, private information and the existence of multiple equilibria are all features shared by many real-world problems. To address this problem, I developed a polynomial algorithm to identify the reasoning patterns [1], and gave the human an explanation of each pattern (e.g., "by doing this action, the other player will infer that you are of this type") as well as a heuristic quantification of its effects in terms of the utility obtained. Human players who received such advice outperformed those who did not [2]. To address more complex games, such as Bayesian games without a common prior, I extended the theory of reasoning patterns [4]. Moreover, I developed a novel concise graphical representation for such games [3], which allows reasoning patterns to be identified graphically in polynomial time. The extended theory has been used to answer questions of strategic relevance (such as "would player $i$ want to lie to player $j$?") without having to solve the game. This enables the modeler of a system to predict or anticipate the behavior of agents by simply looking at the game's structure and running a lightweight analysis algorithm, without having to consider their behavior in detail, or even make restrictive assumptions about their rationality and decision-making algorithms.

## 4. REFERENCES

[1] D. Antos and A. Pfeffer. Identifying reasoning patterns in games. In *Uncertainty in Artificial Intelligence*, 2008.

[2] D. Antos and A. Pfeffer. Using reasoning patterns to help humans solve complex games. In *International Joint Conference on Artificial Intelligence*, 2009.

[3] D. Antos and A. Pfeffer. A graphical representation for bayesian games. In *AAMAS*, 2010.

[4] D. Antos and A. Pfeffer. Reasoning patterns in bayesian games. In *AAMAS*, 2011.

[5] T. W. Bickmore and R. W. Picard. Establishing and maintaining long-term human-computer relationships. In *Computer-Human Interaction*, volume 12, page 2, 2004.

[6] G. Gigerenzer and H. Brighton. Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, (1):107–143, 2009.

[7] A. Pfeffer and K. Gal. The reasoning patterns of agents in games. In *AAAI*, 2007.

[8] G. A. Van Kleef, C. De Dreu, and A. Manstead. The interpersonal effects of anger and happiness in negotiations. *Journal of Personality and Social Psychology*, (86):57–76, 2004.