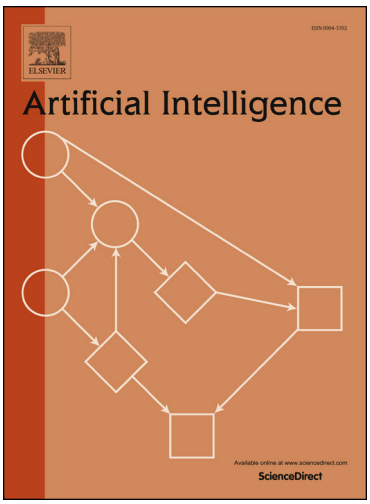
Accepted Manuscript



Belief and truth in hypothesised behaviours

Stefano V. Albrecht, Jacob W. Crandall, Subramanian Ramamoorthy

PII: S0004-3702(16)30023-6
DOI: http://dx.doi.org/10.1016/j.artint.2016.02.004
Reference: ARTINT 2931

To appear in: *Artificial Intelligence*

Received date: 27 July 2015
Revised date: 29 January 2016
Accepted date: 29 February 2016

Please cite this article in press as: S.V. Albrecht et al., Belief and truth in hypothesised behaviours, *Artif. Intell.* (2016), http://dx.doi.org/10.1016/j.artint.2016.02.004

Belief and Truth in Hypothesised Behaviours

Stefano V. Albrecht[a], Jacob W. Crandall[b], Subramanian Ramamoorthy[c]

[a]*The University of Texas at Austin, United States*
[b]*Masdar Institute of Science and Technology, United Arab Emirates*
[c]*The University of Edinburgh, United Kingdom*

Abstract

There is a long history in game theory on the topic of Bayesian or "rational" learning, in which each player maintains beliefs over a set of alternative behaviours, or types, for the other players. This idea has gained increasing interest in the artificial intelligence (AI) community, where it is used as a method to control a single agent in a system composed of multiple agents with unknown behaviours. The idea is to hypothesise a set of types, each specifying a possible behaviour for the other agents, and to plan our own actions with respect to those types which we believe are most likely, given the observed actions of the agents. The game theory literature studies this idea primarily in the context of equilibrium attainment. In contrast, many AI applications have a focus on task completion and payoff maximisation. With this perspective in mind, we identify and address a spectrum of questions pertaining to belief and truth in hypothesised types. We formulate three basic ways to incorporate evidence into posterior beliefs and show when the resulting beliefs are correct, and when they may fail to be correct. Moreover, we demonstrate that prior beliefs can have a significant impact on our ability to maximise payoffs in the long-term, and that they can be computed automatically with consistent performance effects. Furthermore, we analyse the conditions under which we are able complete our task optimally, despite inaccuracies in the hypothesised types. Finally, we show how the correctness of hypothesised types can be ascertained during the interaction via an automated statistical analysis.

Keywords: Autonomous agents, multiagent systems, game theory, type-based method

1. Introduction

There is a long history in game theory on the topic of Bayesian or "rational" learning (e.g. Nachbar, 2005; Dekel et al., 2004; Kalai and Lehrer, 1993; Jordan, 1991). Therein, players maintain beliefs about the behaviours, or "types", of other players in the form of a probability distribution over a set of alternative types. These beliefs are updated based on the observed actions, and each player chooses an action which is expected to maximise the payoffs received by the player, given the current beliefs of the player. The principal questions studied in this context are the degree to which players can learn to make correct predictions, and whether the interaction process converges to solutions such as Nash equilibrium (Nash, 1950).

This general idea, which we here refer to as the *type-based method*, has received increasing interest in the artificial intelligence (AI) community, where it is used as a method to control a single agent in a system composed of multiple agents (e.g. Albrecht and Ramamoorthy, 2013a; Barrett et al., 2011; Gmytrasiewicz and Doshi, 2005; Carmel and Markovitch, 1999). This interest