Asimovian Multiagents: Applying Laws of Robotics to Teams of Humans and Agents

Nathan Schurr[1], Pradeep Varakantham[1], Emma Bowring[1], Milind Tambe[1], and Barbara Grosz[2]

[1] Computer Science Department, University of Southern California
Los Angeles, California
{schurr,varakant,bowring,tambe}@usc.edu
[2] Harvard University, Maxwell-Dworkin Laboratory, Room 249
33 Oxford Street, Cambridge, MA 02138
grosz@eecs.harvard.edu

Abstract. In the March 1942 issue of "Astounding Science Fiction", Isaac Asimov for the first time enumerated his *three laws of robotics*. Decades later, researchers in agents and multiagent systems have begun to examine these laws for providing a useful set of guarantees on deployed agent systems. Motivated by unexpected failures or behavior degradations in complex mixed agent-human teams, this paper for the first time focuses on applying Asimov's first two laws to provide behavioral guarantees in such teams. However, operationalizing these laws in the context of such mixed agent-human teams raises three novel issues. First, while the laws were originally written for interaction of an individual robot and an individual human, clearly, our systems must operate in a team context. Second, key notions in these laws (e.g. causing "harm" to humans) are specified in very abstract terms and must be specified in concrete terms in implemented systems. Third, since removed from science-fiction, agents or humans may not have perfect information about the world, they must act based on these laws despite uncertainty of information. Addressing this uncertainty is a key thrust of this paper, and we illustrate that agents must detect and overcome such states of uncertainty while ensuring adherence to Asimov's laws. We illustrate the results of two different domains that each have different approaches to operationalizing Asimov's laws.

1 Introduction

Recent progress in the agents arena is bringing us closer to the reality of multiagent teams and humans working together in large-scale applications [3,4,10,11,12]. In deploying such multiagent teams and making them acceptable to human teammates, it is crucial to provide the right set of guarantees about their behavior. The unanswered question is then understanding the right set of guarantees to provide in such teams.

R.H. Bordini et al. (Eds.): ProMAS 2006, LNAI 4411, pp. 41–55, 2007.

In this paper, we focus on Asimov's three laws of robotics from his science-fiction stories that provide us a starting point for such behavior guarantees. We do not claim that these laws are the only or best collection of similar rules. However, the laws outline some of the most fundamental guarantees for agent behaviors, given their emphasis on ensuring that *no harm* comes to humans, on obeying human users, and ensuring protection of an agent. Indeed, these laws have inspired a great deal of work in agents and multiagent systems already [14,4,8]. However, in operationalizing these laws in the context of multiagent teams, three novel issues arise. First, the key notions in these laws (e.g. "no harm" to humans) are specified in very abstract terms and must be specified in concrete terms in implemented systems. Second, while the laws were originally written for interaction of an individual robot and an individual human, clearly, our systems must operate in a team context. Third, since, in many realistic domains, agents or humans may not have perfect information about the world, they must act based on these laws despite information uncertainty and must overcome their mutual information mismatch.

Indeed, as mentioned earlier, researchers have in the past advocated the use of such laws to provide guarantees in agent systems [4,8,14]. However, previous work only focused on a single law (the first law of safety) and in the process addressed two of the issues mentioned above: defining the notion of harm to humans and applying the laws to teams rather than individual agents. The key novelty of our work is going beyond previous work to consider the second of Asimov's laws, and more importantly in recognizing the fundamental role that uncertainty plays in any faithful implementation of such a law. In particular, Asimov's second law addresses situations where an agent or agent team may or may not obey human orders — it specifies that in situations where (inadvertent) harm may come to other humans, agents may disobey an order. However, in the presence of uncertainty faced either by the agents or the human user about each others' state or state of the world, either the set of agents or the human may not be completely certain of their inferences regarding potential harm to humans. The paper illustrates that in the presence of such uncertainty, agents must strive to gather additional information or provide additional information. Given that the information reduces the uncertainty, agents may only then disobey human orders to avoid harm.

To the best of our knowledge, this paper for the first time provides concrete implementations that address the three key issues outlined above in operationalizing Asimov's laws. Our implementations are focused on two diverse domains, and thus require distinct approaches in addressing these issues. The first domain is that of disaster rescue simulations. Here a human user provides inputs to a team of (semi-)autonomous fire-engines in order to extinguish maximum numbers of fires and minimize damage to property. The real-time nature of this domain precludes use of computationally expensive decision-theoretic techniques, and instead agents rely on heuristic techniques to recognize situations that may (with some probability) cause harm to humans. The second domain is that of a team of software personal assistant deployed in an office environment to assist human users to complete tasks on time. The personal