Intentions in Equilibrium*

John Grant[1,2] **Sarit Kraus**[2,3] **Michael Wooldridge**[4]

[1]Department of Mathematics,
Towson University, Towson, D 21252, USA
jgrant@towson.edu

[2]Institute for Advanced Computer Studies
University of Maryland, College Park 20742, USA

[3]Department of Computer Science
Bar-Ilan University, Ramat-Gan, 52900 Israel
sarit@cs.biu.ac.il

[4]Department of Computer Science,
University of Liverpool, Liverpool L69 3BX, UK
mjw@liv.ac.uk

Abstract

Intentions have been widely studied in AI, both in the context of decision-making within individual agents and in multi-agent systems. Work on intentions in multi-agent systems has focused on joint intention models, which characterise the mental state of agents with a shared goal engaged in teamwork. In the absence of shared goals, however, intentions play another crucial role in multi-agent activity: they provide a basis around which agents can mutually coordinate activities. Models based on shared goals do not attempt to account for or explain this role of intentions. In this paper, we present a formal model of multi-agent systems in which belief-desire-intention agents choose their intentions taking into account the intentions of others. To understand rational mental states in such a setting, we formally define and investigate notions of multi-agent intention equilibrium, which are related to equilibrium concepts in game theory.

Introduction

Intentions have received considerable attention within the AI community, as a key component in the cognitive makeup of practical reasoning agents (Bratman 1987; Cohen and Levesque 1990; Rao and Georgeff 1992; Shoham 2009). A typical operational interpretation is that an intention corresponds to a plan that has been chosen for execution, and implies a persistent commitment on behalf of an agent to bring about some state of affairs. As well as playing a central role in the decision-making process of individual agents, intentions also play a central role in multi-agent activity. Several models of *joint intentions* have been developed, with the aim of using intentions in practical frameworks for building teams of cooperative problem solvers (Cohen and Levesque 1991; Jennings 1995; Grosz and Kraus 1996). Joint intention models typically attempt to characterise the mental state of agents involved in teamwork, and critically, they assume that a group of agents is pursuing a common goal. However, intentions also play a more "everyday" role in coordinating multi-agent activity, where intentions simply provide a basis around which to coordinate activities in the absence of common goals. For example, consider the following scenario:

*This research is based upon work supported in part by NSF grant 0705587 and under ISF grant 1357/07.

Example 1 *Bob decides to watch football on TV at home tonight; Bob's wife Alice doesn't like football, so she decides to go out and visit friends.*

Here, Alice coordinates her intentions (go out and visit friends) around Bob's (watch football on TV at home). Collective intention models such as (Cohen and Levesque 1991; Jennings 1995; Grosz and Kraus 1996) do not seem appropriate for modelling such scenarios: there is no common goal here, and no shared plan.

Our first aim in this paper is to develop a model that can capture this everyday role of intentions in coordination. We formulate a model of mental state and decision making in practical reasoning agents that enables them to choose their intentions taking into account the intentions of others. Now, since an agent *i*'s rational choice of intentions might depend on an agent *j*'s choice of intentions, and agent *j*'s choice of intentions might in turn depend on agent *i*'s choice of intentions, we are motivated to introduce notions of *multi-agent intention equilibrium*, representing possible notions of rational mental state in a multi-agent BDI setting. These equilibrium notions are related to and indeed inspired by solution concepts from game theory, in particular, the notion of Nash equilibrium (Osborne and Rubinstein 1994). The present paper is novel in two respects: it is the first to explicitly consider the issue of coordinating intentions in BDI-like systems without shared goals; and it is the first to consider equilibrium/stability notions in multi-agent BDI settings. One interesting aspect of our work is that when defining the different notions of equilibrium, the use of the BDI formalism makes explicit some assumptions that are implicit in conventional game theoretic models, such as the role of each agent's beliefs in computing an equilibrium.

The Model

The formal model of BDI agents that we use is derived from that proposed in (Grant et al. 2010). First, we assume as given a logical language L, used by all agents to represent their beliefs. We do not place any requirements on this language other than that it contains the usual classical logic connectives ($\top, \bot, \wedge, \vee, \neg, \rightarrow, \leftrightarrow$), which behave in the classical way, and that there is a proof relation $\vdash$ defined for the language. For the purposes of analysis, we will usually assume that L is classical propositional logic. We write F for the set of sentences of this language, and

write $\wedge K$ to denote the conjunction of a finite set of formulae $K \subset F$. When considering computational questions, as well as assuming that L is propositional logic, we will assume that we have an oracle for $\vdash$, i.e., we can determine in unit time whether $B \vdash \varphi$ for $B \subseteq F$ and $\varphi \in F$. (In practice, one would typically choose L to be appropriately restricted to make proof computationally tractable; but we will work with the oracle model, since it allows us to focus on the complexity of our problems.) We assume that the belief revision operation $\dot{+}$ has already been defined for L, cf. (Alchourron, Gärdenfors, and Makinson 1985; Gärdenfors 1988). We will also assume that we have a unit-time oracle for this operation, i.e., we assume we can compute $B\dot{+}\varphi$ in unit time for arbitrary $B \subseteq F$, $\varphi \in F$.

Next, we assume a set $\mathcal{A} = \{1, \ldots, n\}$ of *agents*. For each agent $i \in \mathcal{A}$ we assume a finite set of *actions*, $Ac_i = \{\alpha_1, \alpha_2, \ldots\}$. The set Ac_i represents the abilities of agent i. We will assume $Ac_i \cap Ac_j = \emptyset$ for all $i \neq j \in \mathcal{A}$. Agent i's knowledge about how to bring about states of affairs in its environment is captured by a set of *recipes* $R_i = \{\langle \alpha, \varphi \rangle \mid \alpha \in Ac_i \text{ and } \varphi \in F\}$. Intuitively, a recipe $\langle \alpha, \varphi \rangle \in R_i$ represents to agent $i \in \mathcal{A}$ the fact that performing action α will accomplish a state of affairs satisfying φ (see, e.g., (Pollack 1992) for discussion on the notion of "plan as recipe"). For every recipe $r = \langle \alpha, \varphi \rangle$, we assume there is a proposition $r_{\alpha,\varphi} \in L$. Intuitively, $r_{\alpha,\varphi}$ will be used to mean that: (*i*) the action α is executable, in that its precondition is currently satisfied, and (*ii*) the performance of α terminates and makes φ true. Only those recipes whose actions the agent believes can be executed are listed as beliefs using the propositions $r_{\alpha,\varphi}$. We will assume that in all recipes $\langle \alpha, \varphi \rangle$, the formula φ contains no recipe propositions.

BDI Structures

Next we define *BDI structures*, the basic model of the mental state of an agent used throughout the remainder of the paper. For the moment, however, we do not present the constraints on such structures that characterise a "rational" mental state. Formally, a BDI structure $\mathcal{S}_i$ for agent $i \in \mathcal{A}$ is a 5-tuple,

$$\mathcal{S}_i = \langle B_i, D_i, I_i, v_i, c_i \rangle$$

where:

- B_i stands for the *beliefs* of the agent; it is the logical closure of a finite set of sentences. Formally, $B_i = \{b \in F \mid B_i^0 \vdash b\}$, where $B_i^0 \subset F$ is a finite set. When we later consider computational questions, we assume beliefs B_i are represented by its basis, B_i^0.
- $D_i \subset F$ and D_i is finite. D_i stands for the *desires* of the agent. We will use $d, d_1, \ldots$ as meta-variables ranging over elements of D_i .
- $I_i \subseteq R_i$. I_i stands for the intentions of the agent.
- $v_i : \mathcal{P}(D_i) \rightarrow \mathbb{R}^{\geq}$, where $\mathbb{R}^{\geq}$ is the set of real numbers greater than or equal to 0: v_i is a valuation function that assigns a nonnegative value to each set of desires of the agent. We require that v_i satisfy the following "entailment-value" condition in case T is consistent:

  for $T \subseteq D_i$ and $T' \subseteq D_i$, if $T \vdash T'$ then $v_i(T) \geq v_i(T')$.

- $c_i : \mathcal{P}(Ac_i) \rightarrow \mathbb{R}^{\geq}$ is a *cost function* for sets of actions, which must satisfy the following conditions:

  $$\text{if } K \subseteq K' \subseteq Ac_i \text{ then } c_i(K') \geq c_i(K).$$

There are several points to make about this definition. First, note that we are explicitly representing an agent's beliefs as a set of logical formulae, closed under deduction. Under this model, agent i is said to believe φ if $\varphi \in B_i$. We use a similar representation for desires. We represent an agent's intentions as the set of recipes that it has selected, and implicit within this set of recipes, the states of affairs that it has committed to bringing about. Thus far, our model closely resembles many other models of BDI agents developed in the literature (Rao and Georgeff 1992). However, the value (v_i) and cost (c_i) functions distinguish it: as we will see, they provide the means through which an agent can choose between and commit to possible sets of intentions.

Where I is a set of intentions, we write:

$$\begin{aligned} \mathit{actions}(I) &= \{\alpha \mid \langle \alpha, \varphi \rangle \in I\} \\ \mathit{goals}(I) &= \{\varphi \mid \langle \alpha, \varphi \rangle \in I\}. \end{aligned}$$

Where B is a set of beliefs and I is a set of intentions, we say that I is *feasible* in the context of B if $\forall \langle \alpha, \varphi \rangle \in I, r_{\alpha,\varphi} \in B$, (i.e., the action part of every intention is believed to be executable). We write $\mathit{feas}(I, B)$ to denote the largest subset of I that is feasible in the context of B. Note that since B has a finite basis, $\mathit{feas}(I, B)$ must be finite.

We find it useful to extend the value function v_i to a function $\bar{v}_i$ on all subsets X of F as follows:

$$\bar{v}_i(X) = \begin{cases} v_i(\{d \in D_i \mid X \vdash d\}) & \text{if } X \nvdash \bot \\ 0 & \text{otherwise.} \end{cases}$$

Next, we define a function $\mathit{ben}_i(I, B)$, which defines the *benefit* that agent i would obtain from the set of intentions I if it were the case that beliefs B were correct:

$$\mathit{ben}_i(I, B) = \bar{v}_i(B\dot{+} \wedge \{\mathit{goals}(\mathit{feas}(I, B))\}) - c_i(\mathit{actions}(I) \cap Ac_i)$$

So, for example:

- $\mathit{ben}_i(I_i, B_i)$ is the benefit that agent i would obtain from its own intentions I_i under the assumption that its own beliefs B_i were correct;
- $\mathit{ben}_i(I_i, B_j)$ is the benefit that agent i would obtain from its own intentions I_i under the assumption that the beliefs of agent j were correct;
- $\mathit{ben}_i(I_1 \cup \cdots \cup I_n, B_i)$ is the benefit that agent i would obtain from the intentions of all agents in the system, under the assumption that its own beliefs were correct.

This definition embodies several important principles:

- The first term on the r.h.s. implicitly assumes that value is only obtained from intentions that are feasible in the context of beliefs B.
- The second term on the r.h.s. implicitly assumes that an agent only incurs costs on the basis of the actions that it

must perform; moreover, it is assumed that an agent incurs the cost of an intention irrespective of whether or not that intention is feasible (whereas an agent only obtains value from feasible intentions). This is needed because the definition allows I and B to stand for the intentions and beliefs of different agents.

Notice that we have said nothing about the *representation* of BDI structures, which is a key issue if we try to understand the complexity of various operations on them. We must first consider how the value and cost functions are represented, since naive representations of them (listing all input/output pairs that define the function) is not practical. Clearly, it would be desirable to have a representation that was polynomial in the size of the remainder of the structure, and moreover, allows the computation of the corresponding v_i and c_i functions in polynomial time. In general, one would simply want to require that v_i and c_i are polynomial time computable functions, but when considering concrete complexity questions, this is often not sufficient for establishing meaningful results. So, in this paper, we will follow (Grant et al. 2010) and assume that v_i and c_i are represented as *straight line programs*. Intuitively, a straight line program can be thought of as a sequence of assignment statements. ***Note:*** *When we state complexity results, it should be understood that these results are with respect to this representation, with respect to the oracle model for proof and belief revision for propositional logic L, and with respect to the finite representation B_i^0 of belief sets B_i.*

Weak Rationality Constraints

The following are the basic constraints we consider on our BDI structures:

A1 B_i is consistent, i.e., $B_i \nvdash \bot$.

A2 I_i is feasible in the context of B_i, i.e., $I_i \subseteq \mathit{feas}(R_i, B_i)$.

A3 $\mathit{goals}(I_i)$ is consistent, i.e., $\mathit{goals}(I_i) \nvdash \bot$.

A4 $\forall \varphi \in \mathit{goals}(I_i)$, $B_i \nvdash \varphi$.

We say a BDI structure that satisfies these properties is *weakly rational*, and hence we will talk of "WRBDI structures". In a WRBDI structure, the agent has an internally consistent model of the environment, and has a set of intentions I_i that is consistent and compatible with this model of the environment.

Intentions in Equilibrium

The weak rationality conditions presented above do not require that a set of intentions I_i is in any sense rational with respect to the cost c_i and value v_i functions. Nor do they require that intentions are coordinated in a rational way around the intentions of others. As Example 1 illustrated, our choice of intentions will be influenced by the intentions of others, and their choice by ours in turn. Alice chooses to visit friends because Bob is watching football on TV and she doesn't like football; if she had earlier announced to Bob that she was planning to watch TV that night, then Bob might well have chosen to watch the football in a bar. In this section, our aim is to consider notions of rationality that take into account the cost and value of an agent's intentions, and the intentions of others.

We first extend our model of BDI agents to a model of multi-agent systems: a *multi-agent system* $\mathcal{M}$ is simply an indexed set of WRBDI structures $\mathcal{M} = \{\mathcal{S}_1, \ldots, \mathcal{S}_n\}$. Notice that we do not in general require that agents have the same, or even collectively consistent beliefs. We will say that a multi-agent system $\mathcal{M} = \{\mathcal{S}_1, \ldots, \mathcal{S}_n\}$ has *common beliefs* if $B_1 = \cdots = B_n$. Sometimes, we find it convenient to talk about the *objective state of the system*. This is the state of the system as perceived by an omniscient external observer. When we make this assumption, we use $E_0 \subset F$ to denote the formulae characterising the objective state of the system, and we denote the closure of E_0 under $\vdash$ by E. Thus E will represent the beliefs of an omniscient external observer; an agent i with beliefs $B_i \subseteq E$ would have correct (although possibly incomplete) information about its environment.

Throughout this section, we make use of the following running example.

Example 2 *Alice owns a business, and Bob works for Alice. Alice wants to keep the office open as much as possible. She wants to take a holiday, as long as Bob staffs the office; but Bob is ill, and if he continues to be unwell, he will not be able to staff the office. Bob has two options: to take an antibiotic prescribed by his doctor, which will certainly make him well, but which is expensive, or to take homeopathic medicine which he believes will make him well. The homeopathic option is cheap, but it will not make him well. Alice does not believe that the homeopathic option will work, but Bob believes it will. What will Alice and Bob do?*

As we will see, there are several possible solutions to this example, depending on the assumptions made about what the agents know about each other's beliefs. We will formalize this example using our BDI structures.

Example 3 *The language contains propositions "well" (Bob is well), "open" (the office is open), and "holiday" (Alice takes a holiday). There are three actions: "anti" and "homeo" (Bob takes antibiotics and homeopathic medicine, respectively), and "go" (Alice goes on holiday). There are two recipes: $\langle \mathit{anti}, \mathit{well} \rangle$ and $\langle \mathit{homeo}, \mathit{well} \rangle$ in Bob's recipes set R_b and one for Alice, $\langle \mathit{go}, \mathit{holiday} \rangle$. Bob's belief set is $B_b = \{r_{\mathit{anti},\mathit{well}}, r_{\mathit{homeo},\mathit{well}}\}$. Alice's belief set is $B_a = \{r_{\mathit{go},\mathit{holiday}}, r_{\mathit{anti},\mathit{well}}, \mathit{well} \rightarrow \mathit{open}, \neg\mathit{holiday} \rightarrow \mathit{open}, \neg\mathit{holiday}\}$. Bob's desires are $D_b = \{\mathit{well}\}$ and Alice's $D_a = \{\mathit{open}, \mathit{holiday}\}$. Alice's valuation function is: $v_a(\{\mathit{open}\}) = 150$, $v_a(\{\mathit{holiday}\}) = 50$, $v_a(\{\mathit{open}, \mathit{holiday}\}) = 210$, and $v_a(\emptyset) = 0$. Bob's valuation function is $v_b(\{\mathit{well}\}) = 100$ and $v_b(\emptyset) = 0$. Bob's cost function $c_b(\{\mathit{homeo}\}) = 10$, $c_b(\{\mathit{anti}\}) = 20$, $c_b(\{\mathit{homeo}, \mathit{anti}\}) = 30$ and $c_b(\emptyset) = 0$ and Alice's $c_a(\{\mathit{go}\}) = 30$ and $c_a(\emptyset) = 0$.*

Locally Rational Intentions

The first possibility we explore is that agents simply ignore each other: they choose a set of intentions that would be optimal *if they were the only agent in the system*. Thus, the agent assumes that any benefit it derives will be from its own

intentions alone. Formally, we say $\mathcal{S}_i$ is *locally rational* if it satisfies the following requirement:

A6 $\nexists I' \subseteq R_i$ such that $\langle B_i, D_i, I', v_i, c_i\rangle$ is a WRBDI structure and $ben_i(I', B_i) > ben_i(I_i, B_i)$.

Example 4 *Consider Bob and Alice from Example 3. Bob's intention set of a WRBDI that satisfies A6 is* $I_b = \{\langle homeo, well\rangle\}$ *and Alice's is* $I_a = \emptyset$. *Intuitively, Alice can't go on a holiday since she can't make Bob well but rather he must take an action that does not belong to her local view.*

Proposition 1 *The problem of checking whether a WRBDI structure* $\langle B_i, D_i, I_i, v_i, c_i\rangle$ *is locally rational is co-*NP*-complete.*

Proof: We show that the complement problem is NP-complete. Membership is by a standard "guess and check" approach. For hardness, we reduce SAT. Let φ be a SAT instance in CNF over k Boolean variables $x_1, \ldots, x_k$, and with l clauses $\psi_1, \ldots, \psi_l$. Let $\mathcal{CL}_\varphi = \{\psi_1, \ldots, \psi_l\}$ denote the set of clauses in φ. We create a single agent, and so we omit subscripts. For each propositional variable x_i, we create two actions α_{x_i} and $\alpha_{\neg x_i}$, which will correspond to assignments of truth or falsity to variable x_i, respectively. For each clause $\psi_i \in \mathcal{CL}_\varphi$ we create a propositional variable z_i that represents it. We then create a set of recipes $R = \{\langle z_i, \alpha_\ell\rangle \mid z_i \text{ represents } \psi_i \in \mathcal{CL}_\varphi \text{ and } \ell \text{ is a literal in } \psi_i\}$. Define B so that all intentions are feasible, and let $D = \{z_1, \ldots, z_\ell\}$. We define $v(D) = 1$ and $v(S) = 0$ for all $S \neq D$. Then, for a set of actions T we define $c(T) = 1$ if for some x_i, $\{\alpha_{x_i}, \alpha_{\neg x_i}\} \subseteq T$, otherwise $c(T) = 0$. Finally, we define $I = \emptyset$. Now, $ben(I, B) = 0 - 0 = 0$, so I will be sub-optimal only if the agent can choose a set of intentions I' such that $ben(I', B) = 1$, i.e. $v(goals(I'), B) = 1$ and $c(actions(I')) = 0$: such a set of intentions will define a consistent satisfying valuation for the input formula φ. ∎

Nash Stability

Our next equilibrium notion assumes that each agent uses its own beliefs when choosing an appropriate set of intentions. More precisely, we are assuming that every agent can see completely and correctly the WRBDI structure of every other agent, but each agent determines not to change its beliefs (i.e., every agent assumes that its beliefs are correct). Then, when agent i tries to predict what agent j will do, it has to assume that agent j's choice of actions will be based on agent j's beliefs (and agent j has to assume that agent i's choice will be based on agent i's beliefs).

We say a system $\mathcal{M}$ with intention sets $I_1, \ldots, I_n$ is *individually stable for agent i* if the following requirement is satisfied:

A7 $\nexists I' \subseteq R_i$ such that $\langle B_i, D_i, I', v_i, c_i\rangle$ is a WRBDI structure, and $ben_i(I_1 \cup \cdots \cup I_{i-1} \cup I' \cup I_{i+1} \cup \cdots \cup I_n, B_i) > ben_i(I_1 \cup \cdots \cup I_{i-1} \cup I_i \cup I_{i+1} \cup \cdots \cup I_n, B_i)$.

Example 5 *Consider Bob and Alice from Example 3. Bob's intention set of the WRBDI that satisfies A7 is* $I_b = \{\langle homeo, well\rangle\}$ *and Alice's is* $I_a = \emptyset$. *Intuitively,* $\langle homeo, well\rangle$ *is not feasible according to Alice's beliefs, so she does not believe that Bob will be well. However, Bob believes* $\langle homeo, well\rangle$ *is feasible, and so he intends to take homeopathic medicine.*

Proposition 2 *Given a multi-agent system* $\mathcal{M} = \{\mathcal{S}_1, \ldots, \mathcal{S}_n\}$ *and agent i, the problem of checking whether the intentions of i are individually stable is co-*NP*-complete.*

Proof: We work with the complement problem. A guess-and-check approach suffices for membership, while for hardness, we reduce the problem of checking that intentions are not locally rational, which was proved NP-complete in Proposition 1. Let $\mathcal{M} = \{\mathcal{S}_1, \ldots, \mathcal{S}_n\}$ be the multi-agent system in the local rationality instance, and i be the agent that we want to check for local rationality. Create a new multi-agent system $\mathcal{M}'$ containing only the agent i, and ask whether the intentions of agent i are individually stable in $\mathcal{M}'$. ∎

We now extend individual stability to a notion of collective rationality, which for reasons that will become plain shortly, we refer to as *Nash stability*. We say $\mathcal{M}$ is Nash stable if:

A8 $\forall i \in \mathcal{A}$, $\mathcal{S}_i$ is individually stable in the context of $\mathcal{M}$.

As a corollary of Proposition 2, we have:

Proposition 3 *Given a multi-agent system* $\mathcal{M}$, *checking whether* $\mathcal{M}$ *is Nash stable is co-*NP*-complete.*

For which classes of systems $\mathcal{M}$ can we check stability in polynomial time? Consider the following definitions. When we have a system $\mathcal{M}$ in which $|I_i| \leq 1$ for all $i \in \mathcal{A}$, then we say $\mathcal{M}$ is *single intentioned*; where there is a constant $k \in \mathbb{N}$ such that $|R_i| \leq k$ for all $i \in \mathcal{A}$ then we say $\mathcal{M}$ is k-bounded.

Proposition 4 *If a multi-agent system* $\mathcal{M}$ *is either (*i*) single intentioned, or (*ii*) k-bounded, then it is possible to check in polynomial time whether* $\mathcal{M}$ *is Nash stable.*

With respect to the more general problem, of checking whether a system has a Nash stable state we have:

Proposition 5 *The following problem is* Σ_2^p*-complete: Given a multi-agent system* $\mathcal{M} = \{\mathcal{S}_1, \ldots, \mathcal{S}_n\}$, *with intention sets* $I_1, \ldots, I_n$, *do there exist intention sets* $I'_1, \ldots, I'_n$ *such that if every agent* $i \in \mathcal{A}$ *replaces* I_i *with* I'_i, *then the resulting system* $\mathcal{M}'$ *is Nash stable?*

Notice that the intentions of Alice and Bob in Examples 4 and 5 are the same, albeit for different reasons. However, individually stable intention sets (and thus Nash stable intention sets) and locally rational intention sets may be different:

Example 6 *Suppose that there are two parking spaces near Bob and Alice's office. One parking space is easier to park in. Denote Alice and Bob's attempts to park in space i,* $i = 1, 2$, *by* pa_i *and* pb_i, *respectively. The predicates for Alice parking in space i,* $i = 1, 2$, *are* a_i *and similarly for Bob:* b_i. *The recipes for Bob are* $R_b = \{\langle pb_1, b_1\rangle, \langle pb_2, b_2\rangle\}$ *and for Alice* $R_a = \{\langle pa_1, a_1\rangle, \langle pa_2, a_2\rangle\}$. *A person can't park simultaneously in two parking spaces nor can two people park in the same space. Thus,* $B_a = B_b = \{\neg(a_1 \wedge b_1), \neg(a_2 \wedge b_2), \neg(a_1 \wedge a_2)\}$. $D_a = \{a_1 \vee a_2\}$ *and* $D_b = \{b_1 \vee b_2\}$. $v_a(\{a_1 \vee a_2\}) = v_b(\{b_1 \vee b_2\}) = 100$;

$v_a(\emptyset) = v_b(\emptyset) = 0$. $c_a(\{pa_1\}) = c_b(\{pb_1\}) = 10$ *and* $c_a(\{pa_2\}) = c_b(\{pb_2\}) = 25$. *According to the axiom of local rationality (A6),* $I_a = \{\langle pa_1, a_1\rangle\}$ *and* $I_b = \{\langle pb_1, b_1\rangle\}$. *On the other hand there are two Nash stable intention sets (A7). The first is* $I_a^1 = \{\langle pa_1, a_1\rangle\}$ *and* $I_b^1 = \{\langle pb_2, b_2\rangle\}$. *The second* $I_a^2 = \{\langle pa_2, a_2\rangle\}$ *and* $I_b^1 = \{\langle pb_1, b_1\rangle\}$.

Of course, there is no guarantee that a system will have a Nash stable state, as the following demonstrates.

Example 7 *Alice wants to go to the pub tonight, unless Bob will be there, in which case she would rather stay at home. Bob wants to go to the pub only if Alice is there; and if she isn't then he would rather stay at home.*

As the name suggests, Nash stability is inspired by concepts from game theory, and we can make the link between our framework and game theory precise. First, we recall some basic definitions from game theory (Osborne and Rubinstein 1994). A *finite game in strategic form* is a tuple:

$$G = \langle \mathcal{A}, \Sigma_1, \ldots, \Sigma_n, \mathcal{U}_1, \ldots, \mathcal{U}_n \rangle$$

where:

- $\mathcal{A} = \{1, \ldots, n\}$ is a set of agents;
- Σ_i is finite, non-empty set of strategies for agent i;
- $\mathcal{U}_i : \Sigma_1 \times \cdots \times \Sigma_n \to \mathbb{R}$ is the *utility function* for agent i.

A tuple of strategies $\langle \sigma_1, \ldots, \sigma_i, \ldots, \sigma_n \rangle$ where $\sigma_j \in \Sigma_j$ for each $j \in \mathcal{A}$ is said to be a *Nash equilibrium* if for all agents $i \in \mathcal{A}$ there is no other strategy $\sigma_i' \in \Sigma_i$ such that $\mathcal{U}_i(\sigma_1, \ldots, \sigma_i', \ldots, \sigma_n) > \mathcal{U}_i(\sigma_1, \ldots, \sigma_i, \ldots, \sigma_n)$.

Given a multi-agent system $\mathcal{M}$, let us define the game $G^{\mathcal{M}} = \langle \mathcal{A}^{\mathcal{M}}, \Sigma_1^{\mathcal{M}}, \ldots, \Sigma_n^{\mathcal{M}}, \mathcal{U}_1^{\mathcal{M}}, \ldots, \mathcal{U}_n^{\mathcal{M}} \rangle$ induced by $\mathcal{M}$ as follows:

- $\mathcal{A}^{\mathcal{M}} = \mathcal{A} = \{1, \ldots, n\}$
- for all $i \in \mathcal{A}$: $\Sigma_i^{\mathcal{M}} = \{feas(I, B_i) \mid I \subseteq R_i \,\&\, \langle B_i, D_i, I, v_i, c_i \rangle$ is a WRBDI structure$\}$
- for all $i \in \mathcal{A}$ and $\langle I_1, \ldots, I_n \rangle \in \Sigma_1 \times \cdots \times \Sigma_n$: $\mathcal{U}_i^{\mathcal{M}}(I_1, \ldots, I_n) = ben_i(I_1 \cup \cdots \cup I_n, B_i)$

We obtain:

Proposition 6 *Let $\mathcal{M}$ be a multi-agent system in which the intention sets of the agents are $I_1, \ldots, I_n$. Then $\mathcal{M}$ is Nash stable iff $\langle I_1, \ldots, I_n \rangle$ is a Nash equilibrium of the game $G^{\mathcal{M}}$.*

One interesting aspect of our formulation of Nash stability, as compared to the standard game-theoretic formulation of Nash equilibrium, is that the role of beliefs in defining the game structure is made explicit. As we will see in what follows, we have in fact several different stability notions, depending on whose beliefs are used to judge benefits.

Objective Nash Stability

The next possibility we consider is the equilibrium of the system from the point of view of an omniscient external observer, who can see exactly what the state of the system is, and who thus knows exactly what recipes are feasible, etc. Recall that E denotes the beliefs of an omniscient external observer, who is objectively able to see the actual state of the system. This leads to the concept of *objective Nash stability*.

As with Nash stability defined above, we define this concept with respect to an intermediate notion, in this case *objective individual stability*. An agent's set of intentions is objectively individually stable if it could not change to another set of intentions that yield greater benefit, when the benefit is measured in the context of E. Formally, $\mathcal{S}_i$ is objectively individually stable if it satisfies the following property:

A9 $\nexists I' \subseteq R_i$ such that $\langle E, D_i, I', v_i, c_i \rangle$ is a WRBDI structure, and $ben_i(I_1 \cup \cdots \cup I_{i-1} \cup I' \cup I_{i+1} \cup \cdots \cup I_n, E) > ben_i(I_1 \cup \cdots \cup I_{i-1} \cup I_i \cup I_{i+1} \cup \cdots \cup I_n, E)$.

When every agent's intentions are objectively individually stable, then the system is *objectively Nash stable*:

A10 $\forall i \in \mathcal{A}$, $\mathcal{S}_i$ is objectively individually stable.

Example 8 *Considering Bob and Alice from Example 3 since taking homeopathic medicine will not make Bob well,* $E = \{r_{go,holiday}, r_{anti,well}, well \to open, \neg holiday \to open, \neg holiday\}$.

The sets of intentions that are objectively individually stable are $I_b = \{\langle anti, well\rangle\}$ *and* $I_a = \langle go, holiday\rangle$.

Given a multi-agent system $\mathcal{M}$ characterised by E, let us define the *objective game*

$$G^{\mathcal{M},E} = \langle \mathcal{A}^{\mathcal{M},E}, \Sigma_1^{\mathcal{M},E}, \ldots, \Sigma_n^{\mathcal{M},E}, \mathcal{U}_1^{\mathcal{M},E}, \ldots, \mathcal{U}_n^{\mathcal{M},E} \rangle$$

induced by $\mathcal{M}$ as follows ($\mathcal{A}^{\mathcal{M},E}$ is defined as expected):

- for all $i \in \mathcal{A}$: $\Sigma_i^{\mathcal{M},E} = \{feas(I, E) \mid I \subseteq R_i \,\&\, \langle E, D_i, I, v_i, c_i \rangle$ is a WRBDI structure$\}$
- for all $i \in \mathcal{A}$ and $\langle I_1, \ldots, I_n \rangle \in \Sigma_1 \times \cdots \times \Sigma_n$: $\mathcal{U}_i^{\mathcal{M},E}(I_1, \ldots, I_n) = ben_i(I_1 \cup \cdots \cup I_n, E)$.

We immediately obtain:

Proposition 7 *Let $\mathcal{M}$ be a multi-agent system in state E, in which the intention sets of the agents are $I_1, \ldots, I_n$. Then $\mathcal{M}$ is objectively Nash stable iff $\langle I_1, \ldots, I_n \rangle$ is a Nash equilibrium of the game $G^{\mathcal{M},E}$.*

Subjective Nash Stability

There is a difficulty in accepting either Nash stability or objective Nash stability as a rational state of a multi-agent system, for the following reason. In order for an agent i to be able to see that a system is Nash stable, it must be able to compute $ben_j(I_1 \cup \cdots \cup I_n, B_j)$ for each $j \in \mathcal{A}$, which implies that agent i has access to the beliefs B_j of each other player $j \in \mathcal{A}$. In other words, it is only possible to see that a system is Nash stable if one knows the entire state of the system. In order to see that a system is objectively Nash stable, each agent i must be able to compute $ben_j(I_1 \cup \cdots \cup I_n, E)$ for each $j \in \mathcal{A}$, which implies that agent i has access to the objective state of the system E. This motivates us to consider *Nash stability from the point of view of individual agents*: we call this *subjective* Nash stability. In the same fashion as above, we define this with respect to subjective individual stability. We say agent i is subjectively individually stable from the perspective of agent j if:

A11 $\nexists I' \subseteq R_i$ such that $\langle B_j, D_i, I', v_i, c_i \rangle$ is a WRBDI structure, and $ben_i(I_1 \cup \cdots \cup I_{i-1} \cup I' \cup I_{i+1} \cup \cdots \cup I_n, B_j) > ben_i(I_1 \cup \cdots \cup I_{i-1} \cup I_i \cup I_{i+1} \cup \cdots \cup I_n, B_j)$.

We then say that $\mathcal{M}$ is subjectively Nash stable from the point of view of j if:

A12 $\forall i \in \mathcal{A}$, $\mathcal{S}_i$ is subjectively individually stable with respect to agent j in the context of $\mathcal{M}$.

Example 9 *Consider Bob and Alice from Example 3. The subjectively individually stable with respect to Bob is* $I_b = \{\langle homeo, well\rangle\}$ *and* $I_a = \{\langle go, holiday\rangle\}$*: Bob thinks it is rational for him to take homeopathic medicine while Alice goes on holiday. However, the subjectively individually stable with respect to Alice is* $I_b = \{\langle anti, well\rangle\}$ *and* $I_a = \{\langle go, holiday\rangle\}$.

Given a multi-agent system $\mathcal{M}$ and agent i in $\mathcal{M}$, define the *j-subjective game* to be the game induced by the beliefs of agent j: intuitively, the game that j thinks it (and every other agent) is playing. Formally, the j-subjective game $G^{\mathcal{M},j} = \langle \mathcal{A}^{\mathcal{M},j}, \Sigma_1^{\mathcal{M},j}, \ldots, \Sigma_n^{\mathcal{M},j}, \mathcal{U}_1^{\mathcal{M},j}, \ldots, \mathcal{U}_n^{\mathcal{M},j}\rangle$ induced by $\mathcal{M}$ and j is as follows ($\mathcal{A}^{\mathcal{M},j}$ is defined as expected):

- for all $i \in \mathcal{A}$: $\Sigma_i^{\mathcal{M},j} = \{feas(I, B_j) \mid I \subseteq R_i \ \& \ \langle B_j, D_i, I, v_i, c_i\rangle$ is a WRBDI structure$\}$
- for all $i \in \mathcal{A}$ and $\langle I_1, \ldots, I_n\rangle \in \Sigma_1 \times \cdots \times \Sigma_n$: $\mathcal{U}_i^{\mathcal{M},j}(I_1, \ldots, I_n) = ben_i(I_1 \cup \cdots \cup I_n, B_j)$

Now, it will not typically be the case that $G^{\mathcal{M}} = G^{\mathcal{M},j}$, and so Nash stable systems $\mathcal{M}$ will not typically correspond to the Nash equilibria of $G^{\mathcal{M},j}$ as in proposition 7. However, we can easily state a sufficient condition for subjective games to coincide.

Proposition 8 *Let* $\mathcal{M}$ *be a multi-agent system with common beliefs. Then* $\forall j \in \mathcal{A}$, $G^{\mathcal{M}} = G^{\mathcal{M},j}$. *If the agents have common beliefs E, then* $\forall j \in \mathcal{A}$, $G^{\mathcal{M}} = G^{\mathcal{M},E} = G^{\mathcal{M},j}$.

It is of course possible that a system may be subjectively Nash stable for *every* agent, even though beliefs are different. In this case every agent has a correct model of what other agents will do, but for the "wrong reasons" (cf. discussion in (Kalai and Lerhrer 1993)).

Related Work & Discussion

Although intentions in BDI-like systems have been widely studied in AI over the past two decades (Bratman 1987; Cohen and Levesque 1990; Rao and Georgeff 1992), and a variety of joint intention models have been proposed and evaluated (Cohen and Levesque 1991; Jennings 1995; Grosz and Kraus 1996), our work is novel in two respects: it is the first to explicitly consider the issue of coordinating intentions without shared goals; and it is the first to consider equilibrium/stability notions in BDI settings. (Nair 2004; Nair and Tambe 2005) considers related questions, in a hybrid BDI-POMDP setting. (Zuckerman et al. 2007) presented a BDI model for adversarial environments. They gave behavioural axioms specifying the intentions an agent should consider to adopt in such settings, focusing on one agent's point of view. However, again, no stability notions were considered. Finally, (Larbi, Konieczny, and Marquis 2007) uses game-theoretic notions (including Nash equilibrium) applied to multiagent planning but without common goals. Plans are evaluated in terms of combinations that may occur when mixed with the other agents' plans.

An interesting aspect of our work is that it shows how notions of rational mental state in multi-agent BDI-based systems can usefully be understood through analogues of game theoretic concepts. Of particular interest is the fact that the role of belief in computing solution concepts is made explicit, with different perspectives on whose beliefs are or should be used leading to different notions of equilibrium. Future work might consider both conceptual questions (such as other equilibrium notions), as well as computational questions (further consideration of tractable instances).

References

Alchourron, C. E.; Gärdenfors, P.; and Makinson, D. 1985. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic* 50:510–530.

Bratman, M. E. 1987. *Intention, Plans, and Practical Reason*. Harvard University Press: Cambridge, MA.

Cohen, P. R., and Levesque, H. J. 1990. Intention is choice with commitment. *Artificial Intelligence* 42:213–261.

Cohen, P. R., and Levesque, H. J. 1991. Teamwork. *Nous* 25(4):487–512.

Gärdenfors, P. 1988. *Knowledge in Flux*. The MIT Press: Cambridge, MA.

Grant, J.; Kraus, S.; Perlis, D.; and Wooldridge, M. 2010. Postulates for revising BDI structures. *Synthese*.

Grosz, B. J., and Kraus, S. 1996. Collaborative plans for complex group action. *AIJ* 86(2):269–357.

Jennings, N. R. 1995. Controlling cooperative problem solving in industrial multi-agent systems using joint intentions. *Artificial Intelligence* 75(2):195–240.

Kalai, E., and Lerhrer, E. 1993. Subjective equilibria in repeated games. *Econometrica* 61(5):1231–1240.

Larbi, R. B.; Konieczny, S.; and Marquis, P. 2007. Extending classical planning to the multi-agent case: A game-theoretic approach. In *ECSQARU*, 731–742.

Nair, R., and Tambe, M. 2005. Hybrid BDI-POMDP framework for multiagent teaming. *JAIR* 23:367–420.

Nair, R. 2004. *Coordinating multiagent teams in uncertain domains using distributed POMDPs*. Ph.D. Dissertation, University of Southern California.

Osborne, M. J., and Rubinstein, A. 1994. *A Course in Game Theory*. The MIT Press: Cambridge, MA.

Pollack, M. E. 1992. The uses of plans. *AIJ* 57(1):43–68.

Rao, A. S., and Georgeff, M. P. 1992. An abstract architecture for rational agents. In *Proc. of KR92*, 439–449.

Shoham, Y. 2009. Logical theories of intention and the database perspective. *Journal of Philosophical Logic* 38(6):633–648.

Zuckerman, I.; Kraus, S.; Rosenschein, J. S.; and Kaminka, G. A. 2007. An adversarial environment model for bounded rational agents in zero-sum interactions. In *Proc. of AAMAS07*.