Synthese (2009) 169:335–349
Knowledge, Rationality & Action 739–753
DOI 10.1007/s11229-009-9553-5

Intentions and interactive transformations of decision problems

Olivier Roy

Received: 19 November 2008 / Accepted: 6 April 2009 / Published online: 12 May 2009

Abstract In this paper I study two ways of transforming decision problems on the basis of previously adopted intentions, ruling out incompatible options and imposing a standard of relevance, with a particular focus on situations of strategic interaction. I show that in such situations problems arise which do not appear in the single-agent case, namely that transformation of decision problems can leave the agents with no option compatible with what they intend. I characterize conditions on the agents' intentions which avoid such problematic scenarios, in a way that requires each agent to take account of the intentions of others.

Keywords Intentions · Rationality · Interaction · Transformation of decision problems · Fixed points

There is a broad consensus in the philosophy of action that previously adopted intentions, along with beliefs and desires, shape or transform decision problems (Bratman 1987, 1999; Velleman 2008). Intentions impose a "standard for *relevance* for options considered in deliberation. And they constrain solutions to these problems, providing a *filter of admissibility* for options." (Bratman, 1987, p. 33, emphasis in the original) Our understanding of this role of intentions in decision making has profited greatly from research in AI and on theories of multi-agent systems. Implementation of the filtering function in various architectures has helped to refine planning systems (Bratman et al. 1991; Pollack 1992; Horty and Pollack 2001) and to develop sophisticated "belief-desire-intention" (BDI) models of autonomous agents (Cohen and Levesque 1990; Georgeff et al. 1999; Wooldridge 2000; van der Hoek et al. 2007).

O. Roy (✉)
Faculty of Philosophy, University of Groningen, Oude Boteringestraat 52,
9712 GL Groningen, The Netherlands
e-mail: o.roy@rug.nl

There is, however, a gap in this philosophical and formal literature, as it does not address the question of how agents in interactive situations transform their decision problems on the basis of what they intend. Philosophers have studied in detail the roles of intentions in coordinated and shared agency (Bratman 1999; Searle 1995; Tuomela 1995; Alonso 2008), but the philosophical work on how intentions *transform* decision problems have so far overlooked the multi-agent case. The situation is similar in AI and in BDI models: much attention has been paid to rational interaction, but the models of intention-based transformation of decision problems remain individualistic (see the references above). This is a serious shortcoming in the theory of intentions, as in recent years interaction has become one of the key issue in formal theories of rational agency (Aumann 1999; Parikh 2002; van Benthem 2006; Brandenburger 2007).

In this paper I undertake the task of filling this gap by studying how intentions transform decision problems in genuine interactive situations. I formalize two ways of transforming decision problems on the basis of the agents' intentions. For both of them it is shown that problems arise in interactive situations that do not appear in the single-agent case, namely that transformation of decision problems can leave the agents with no option compatible with what they intend. I characterize conditions on the agents' intentions that avoid such problematic scenarios, arguing that they require each agent to take account of the intentions of others.

This paper thus aims to contribute to the theory of intention by pointing out the importance of examining interactive situations, and to the theory of interactive rationality by bringing in intentions and their role in the shaping of decision problems. The paper is not aimed at application to AI architectures or multi-agent systems, and for that reason I shall not attempt to compare or even relate the present framework to those cited above. While such a comparison might be fruitful, it would stray too far from the main concern here.

The paper is organized as follows. In Sect. 1 I introduce strategic games and the representation of intentions therein. In Sects. 2 and 3 I study in turn how intentions rule out incompatible options and impose a standard of relevance. In Sect. 4 I study the interaction of these two operations, and in Sect. 5 I discuss a general issue regarding the relation between intentions and preferences. Proofs of the technical results can be found in the Appendix.

1 Framework and definitions

In this section I first introduce the formal framework in which the various results of this paper are couched. I first present games in strategic forms. I then introduce a simple model of previously-adopted intentions in such games.

1.1 Games in strategic from

I use standard strategic form games, as in e.g. (Osborne and Rubinstein 1994), except that preferences are represented qualitatively. I also restrict myself to finite sets of "pure" strategies, leaving an extension to mixed strategies for future work. A *decision problem* or *strategic game* $\mathbb{G}$ is thus a tuple $\langle I, X, \{S_i, \leq_i\}_{i \in I}, \pi \rangle$ such that :

- I is a finite set of *agents*.
- S_i is a finite set of *actions* or *strategies* for i. A *strategy profile* $\sigma \in \Pi_{i \in I} S_i$ is a vector of strategies, one for each agent in I. The strategy s_i which i plays in the profile σ is denoted $\sigma(i)$.
- X is a finite set of *outcomes*.
- $\pi : \Pi_{i \in I} S_i \rightarrow X$ is an *outcome function* that assigns to every strategy profile $\sigma \in \Pi_{i \in I} S_i$ an outcome $x \in X$. I use $\pi(s_i)$ to denote the set of outcomes that can result from the choice of s_i. Formally: $\pi(s_i) = \{x : x = \pi(s_i, \sigma_{j \neq i}) \text{ for some } \sigma_{j \neq i} \in \Pi_{j \neq i} S_j\}$.
- $\leq_i$ is a reflexive, transitive and total preference relation on X.

1.2 Previously adopted intentions and intention sets

In this paper I study the effect of *previously adopted* intentions on such games in strategic form. These are intentions that an agent adopts beforehand, i.e. some time before entering a given situation of interaction. Previously adopted intentions are thought of as relatively stable motivational states: i.e. they are "conduct controlling" in that they guide the agent toward the realization of their content, and they "persist" from their formation until their realization, except of course in the face of unexpected circumstances (Bratman 1987). Such intentions are usually distinguished from intentions "*in action*" (Searle 1983), which are seen as the "mental component" (O'Shaughnessy 1973) of intentional actions.

Given a game in strategic form $\mathbb{G}$, I model previously adopted intentions by assigning to each agent $i \in I$ an *intention set* $\iota_i \subseteq X$. Attention is thus restricted here to intentions to *realize certain outcomes* in the game, in contrast to intentions to play a certain strategy—although there is an obvious connection between the two. The intention set ι_i of agent i should be thought of as the intentions that i has formed some time before entering the game, and on the basis of which he now has to make his decision. An *intention profile* ι is a vector of intention sets, one for each agent. This representation of previously adopted intentions in games, and of their effects on decision problems, closely follows (van Hees and Roy 2008).

The intentions of an agent i are *consistent* if $\iota_i \neq \emptyset$, and inconsistent otherwise. Similarly, these intentions are *unrealizable* if for all $x \in \iota_i$ there is no profile σ such that $\pi(\sigma) = x$. Having inconsistent or unrealizable intentions, at least knowingly, is for many philosophers of action a blatant case of irrationality (Bratman 2009; Velleman 2008; Wallace 2006). I show in the next sections that the fulfilment of this standard of rationality turns out to be much more complex in interactive situations, especially when intentions filter the available options or impose a standard of relevance on them.

Throughout this paper I do not impose constraints on the relation between intention sets and preferences, and in view of this one might ask how the two stand against each other, and how the intention-based transformations of decision problems that I study next relate to known game-theoretic transformations. Could intentions be reduced to preferences? And similarly for the transformations they induce? These are legitimate questions which, however, will be easier to address when one is in possession of the results set out below, so I postpone them to the end of the paper (Sect. 5). For now I

turn to the two intentions-based transformations mentioned above: "providing a filter of admissibility" and imposing a "standard of relevance."

2 Filter of admissibility

I take "providing a filter of admissibility" to mean ruling out options that are incompatible with the agents achieving their intentions. Agents, in this sense, discard some of their strategies because they are incompatible with what the agents intend.

As a generic approach to this discarding process I define the notion of *cleaned strategy set* and *cleaned decision problem* as follows. Given a strategic game $\mathbb{G}$ with intentions, the cleaned version $cl(S_i)$ of the strategy set S_i of agent i is defined as

$$cl(S_i) = \{s_i \mid s_i \text{ is admissible for deliberation for } i\}$$

with various notions of "admissibility" to be inserted. The cleaned version of $\mathbb{G}$ with intention profile ι is the tuple $cl(\mathbb{G}) = \langle I, X^{cl}, \{cl(S_i), \leq_i^{cl}\}_{i\in I}, \pi^{cl}\rangle$ such that:

- $X^{cl} = \pi(\Pi_{i\in I} cl(S_i)) = \{x \mid x = \pi(\sigma) \text{ for some } \sigma \in \Pi_{i\in I} cl(S_i)\}$.
- $\leq_i^{cl}$ is the restriction of $\leq_i$ to X^{cl}.
- π^{cl} is π with the restricted domain $\Pi_{i\in I} cl(S_i)$.

I assume that the agents adapt their intentions to the decision problem they face after cleaning by abandoning the achievement of those outcomes that are no longer achievable. The question of intention revision is thus left for future research. I take the cleaned version ι_i^{cl} of the intention set ι_i to be $\iota_i \cap X^{cl}$, recalling the expansion operation in belief revision theory, see e.g. (Alchourron et al. 1985) and (Rott 2001).

I study only two of the many ways one could define admissibility of strategies on the basis of the agent's intentions, which I call *individualistic* and *altruistic* admissibility. I choose them because they show clearly how the interactive character of strategic games influences the process of ruling out inadmissible options.

I say that a strategy s_i is *individualistically admissible* for agent i when choosing it can yield at least one outcome he intends. Formally, a strategy s_i of agent i is individualistically admissible with respect to his intention set ι_i when $\pi(s_i) \cap \iota_i \neq \emptyset$. Conversely, a strategy is not admissible for i when choosing it would not realize any of his intentions.

Even if an agent has consistent intentions, in case his intentions are unrealizable no strategy will survive cleaning with individualistic admissibility. In such a case I say that individualistic cleaning *empties* a decision problem for the agent.

In the single-agent case, emptying a decision problem with individualistic cleaning coincides exactly with having unrealizable intentions: cleaning does indeed empty a decision problem *if and only if* ι_i is not realizable. Potential problems arising from cleaning a strategy set on the basis of one's intentions thus boil down, when there is only one agent, to meeting a rationality requirement that is well known in the philosophical theory of action.

In situations of interaction, however, more complex situations arise, since agents who clean individualistically can make the intentions of others unrealizable. Table 1

Table 1 A game with an empty cleaned version, with $X = \Pi_{i \in I} S_i$

$\mathbb{G}$	t_1	t_2
s_1	1	
s_2	2	

provides an example, with the numbers in the cells representing which outcomes are in ι_i for the corresponding agent, 1 being the row and 2 being the column player. After the first round of cleaning, s_2 is eliminated by agent 1, making the intentions of agent 2 inconsistent.

It is thus not sufficient to have realizable intentions to avoid emptying strategy sets after cleaning. To pinpoint the conditions which ensure such non-emptiness in the general case, I look at iteration, in a way that draws from (van Benthem 2007) and (Apt 2007).

Given a strategic game $\mathbb{G}$, let $cl^k(\mathbb{G}) = \langle I, X^{cl^k}, \{cl^k(S_i), \leq_i^{cl^k}\}_{i \in I}, \pi^{cl^k} \rangle$ be the strategic game that results after k iterations of the cleaning of $\mathbb{G}$. That is, $cl^0(\mathbb{G}) = \mathbb{G}$ and $cl^{k+1}(\mathbb{G}) = cl(cl^k(\mathbb{G}))$. The *smallest cleaning fixed point* $cl^{\#}(\mathbb{G})$ of $\mathbb{G}$ is defined as $cl^k(\mathbb{G})$ for the smallest k such that $cl^k(\mathbb{G}) = cl^{k+1}(\mathbb{G})$. In what follows I ignore the "smallest" and only write about the fixed point.

Every game has a unique cleaning fixed point with individualistic cleaning but, as just noted, it may be empty. Under which conditions is this avoided? To answer this question one has to isolate the condition under which a set of strategy profile survives iterated cleaning. Let me call the *cleaning core* of a strategic game $\mathbb{G}$ the set of strategy profile S^* inductively defined as follows, with $\pi^{S^n}(s_i) = \pi(s_i) \cap \{\pi(\sigma') : \sigma' \in S^n\}$.

- $S^0 = \Pi_{i \in I} S_i$.
- $S^{n+1} = S^n - \{\sigma :$ there is an i such that $\pi^{S^n}(\sigma(i)) \cap \iota_i = \emptyset\}$.
- $S^* = \bigcap_{n < \omega} S^n$.

For each strategy s_i and profile σ in the cleaning core such that $\sigma(i) = s_i$, there is at least one agent j for whom strategy $\sigma(j)$ is individualistically admissible, by looking only at what can result from the profiles in the core. This definition is close to that of a cleaning fixed point—except that it looks at strategy profiles instead of at each agent's strategy set—and indeed it exactly captures the sets of profiles which survive iterated cleaning.

Fact 1 *For any strategic game $\mathbb{G}$ and intention profile ι, S^* is empty iff $cl^{\#}(\mathbb{G})$ is empty as well, for individualistic cleaning.*

With this in hand we get as a direct corollary the answer to the question of characterizing the conditions on the intentions sets which avoid empty fixed-points for individualistic cleaning.

Corollary 1 *For any strategic game $\mathbb{G}$ and intention profile ι, $cl^{\#}(\mathbb{G})$ is empty iff there is an agent i such that $\iota_i \cap \{\pi(\sigma) : \sigma \in S^*\}$ is empty as well.*

This corollary shows that the individualistic character of admissibility must be compensated by an interlocking web of intentions and strategies if cleaning is not to

make the game empty. Intentions which yield a non-empty cleaning must contain at least *some* outcomes achievable by one of the profiles in the cleaning core. Observe that this does not mean that there must be one outcome which all agents intend, except if S^* is a singleton. Profiles in the cleaning core, in turn, are profiles which remain compatible with the intentions of *all* agents, throughout the cleaning process. This means that no agent can rule out the achievement of another agent's intentions by choosing according to a profile in the cleaning core, i.e. by choosing $\sigma(i)$ for $\sigma \in S^*$. In this sense, by intending outcomes that are realizable in the cleaning core, an agent can be thought of as acknowledging that he interacts with other agents who, like him, clean inadmissible options from their strategy set on the basis of what they intend.

The following alternative form of admissibility emphasizes this interactive character: a strategy s_i of agent i is *altruistically admissible* with respect to his intention set ι_i when there is a $j \in I$ such that $\pi(s_i) \cap \iota_j \neq \emptyset$. Following this second criterion, a strategy of agent i is admissible whenever it can yield an outcome that some agent, *not necessarily i*, intends. When agents discard options on the basis of this criterion, there is no risk of emptying the game, and the process does not need to be iterated.

Fact 2 *For $\mathbb{G}$ an arbitrary strategic game, $cl^{\#}(\mathbb{G}) = cl(\mathbb{G})$ for cleaning with altruistic admissibility.*

Fact 3 *For any strategic game $\mathbb{G}$, intention profile ι and cleaning with altruistic admissibility, there is, for all i, a realizable $x \in \iota_i$ iff $cl^{\#}(\mathbb{G})$ is not empty.*

It is thus crucial for agents to take each others' intentions into account when ruling out options in strategic games. If, on the one hand, agents rule out options without taking care of what the others intend (individualistic cleaning), they run the risk of ending up with no admissible strategy at all, unless their intentions are already attuned to those of their co-players (i.e. they intend outcomes realizable by profiles in the cleaning core). If, on the other hand, their intentions do not fit well with those of others (no further constraints on the intention sets), then they should at least take heed of what the others intend when ruling out options (altruistic cleaning).

By focusing on the single-agent case, existing studies in philosophy of action and artificial intelligence have hitherto overlooked this genuine interactive aspect of intention-based transformation of decision problems. I now turn to the other transformation I mentioned in the introduction: the imposition of a standard of relevance for options. I show that similar issues arise once we take interaction into account.

3 Standard of relevance

I take the transformation of a decision problem based on a "standard of relevance" imposed by previously adopted intentions to mean discarding options with differences that are not relevant in terms of what one intends. I say that such options are equivalent *means* to achieve what the agent intends. Formally, two strategies s_1 and s_2 in S_i are *equivalent*, denoted $s_1 \approx s_2$, whenever $\pi(s_1, \sigma_{j\neq i}) \in \iota_i$ iff $\pi(s_2, \sigma_{j\neq i}) \in \iota_i$ for all combinations of actions of other agents $\sigma_{j\neq i} \in \Pi_{j\neq i} S_j$. Strategies s_1 and s_2 in Table 2 are equivalent for the row player in that sense. The relation $\approx$ clearly induces

Table 2 A game with equivalent strategies for the row player, with $X = \Pi_{i \in I} S_i$

	t_1	t_2	t_3
s_1	1, 2	2	1
s_2	1	2	1
s_3		1	2

a partition of the set of strategies S_i into subsets $[s_i]^{\mathbb{G}}_{\approx} = \{s'_i \in S_i | s'_i \approx s_i\}$, each of which represents a distinct means for agent i to achieve what he intends.

I take here the standard of relevance imposed by intentions as inducing such a means-oriented perspective on decision problem, in the sense that the options upon which an agent i will deliberate will be the equivalence classes $[s_i]^{\mathbb{G}}_{\approx}$, or representative of them, instead of strategies in S_i. The agents will discard the differences between strategies which are equivalent means to achieve what he intends.

To make a decision from the means-oriented perspective, the agents still have to sort out these means according to some preference ordering, and there is no unique or obvious way to do so on the basis of the underlying preference ordering $\leq_i$ on the outcome set. There is no guarantee that the agents will be indifferent between the outcomes which could be yielded by the strategies in a given equivalence class $[s_i]^{\mathbb{G}}_{\approx}$, and even less that all these outcomes would be, say, at least as good as those in a different equivalence class $[s'_i]^{\mathbb{G}}_{\approx}$.

Here I take one of the many ways to construct such a preference ordering on means, by assuming that the agents "pick" a representative strategy for each means, and collect them to form their new strategy set. This allows preferences to be defined in the game that result from this transformation from those in the original game. Regarding the picking process itself, I adopt an abstract point of view and leave implicit the criterion which underlies it.

Given a strategic game $\mathbb{G}$, a function $\theta_i : \mathcal{P}(S_i) \to S_i$ such that $\theta_i(S) \in S$ for all $S \subseteq S_i$ is called i's *picking function*. A *profile* of picking functions Θ is a combination of such θ_i, one for each agent $i \in I$. These functions return, for each set of strategies—and in particular each equivalence class $[s_i]^{\mathbb{G}}_{\approx}$—the strategy that the agents picks in that set. I define them over the whole power set of strategies to facilitate the technical analysis.

The *pruned version* $pr(S_i)$ of a strategy set S_i, with respect to an intention set ι_i and a picking function θ_i is defined as:

$$pr(S_i) = \{\theta([s_i]^{\mathbb{G}}_{\approx}) : s_i \in S_i\}$$

Pruned version of a strategic game $\mathbb{G}$ are defined similarly to cleaned ones: given an intention profile ι and a profile of picking function Θ, the pruned version of $\mathbb{G}$ is the tuple $pr(\mathbb{G}) = \langle I, X^{pr}, \{pr(S_i), \leq_i^{pr}\}_{i \in I}, \pi^{pr} \rangle$ such that:

- $X^{pr} = \pi(\Pi_{i \in I} pr(S_i))$.
- $\leq_i^{pr}$ is the restriction of $\leq_i$ to X^{pr}.
- π^{pr} is π with the restricted domain $\Pi_{i \in I} pr(S_i)$.

The pruned version ι_i^{pr} of an intention set ι_i is $\iota_i \cap X^{pr}$. Agents, again, adapt their intentions in the process of pruning.

I once again adopt a general point of view and analyze iterations of pruning. Given a strategic game $\mathbb{G}$, let $pr^k(\mathbb{G})$ be the strategic game that results after k iterations of the pruning of $\mathbb{G}$. That is, $pr^0(\mathbb{G}) = \mathbb{G}$ and $pr^{k+1}(\mathbb{G}) = pr(pr^k(\mathbb{G}))$. The *pruning fixed point* $pr^{\#}(\mathbb{G})$ of $\mathbb{G}$ is defined as $pr^k(\mathbb{G})$ for the smallest k such that $pr^k(\mathbb{G}) = pr^{k+1}(\mathbb{G})$.

As for cleaning, it can occur that agents end up with inconsistent or unrealizable intentions after a few rounds of pruning, but no pruning makes a game empty.

Fact 4 *For all strategic game $\mathbb{G}$ and agent $i \in I$, $pr^{\#}(S_i)$ is not empty.*

The existence of pruning fixed points where all agents have non-empty intention sets depends on whether they intend what I call *safe* outcomes. Given a strategic game $\mathbb{G}$, an intention profile ι and a profile of picking functions Θ, the outcome $x = \pi(\sigma)$ is:

- *Safe for pruning at stage 1* iff for all agents i, $\theta_i([\sigma(i)]) = \sigma(i)$.
- *Safe for pruning at stage $n+1$* whenever it is safe for pruning at stage n and for all agents i, $\theta_i\left([\sigma(i)]_{\approx}^{pr^n(\mathbb{G})}\right) = \sigma(i)$.
- *Safe for pruning* when it is safe for pruning at all stages n.

Safe outcomes are those which the picking functions retain, whatever happens in the process of pruning. Intending safe outcomes is necessary and sufficient for an agent to keep his intention set non-empty in the process of pruning.

Fact 5 *For any strategic game $\mathbb{G}$, intention profile ι, profile Θ of picking functions and for all $i \in I$: $\iota_i^{pr^{\#}}$ is not empty iff there is a $\pi(\sigma) \in \iota_i$ safe for pruning in $\mathbb{G}$.*

Similarly to cleaning, intending outcomes safe for pruning is a way for agents to take account of the fact that they are interacting with other agents who are pruning their decision problems on the basis of what they intend. In this case this not only means taking account of the intentions of others, but also of their picking criteria. In single-agent cases pruning never makes the intention set of the agent empty, as long as the agent has realizable intentions.

This shows, once again, that new issues arise when one transforms decision problems on the basis of ones intention in situations of interaction. In the next section I look at how the pruning and cleaning interact with one another, in order to obtain a more general picture of these intention-based transformations.

4 Putting the two transformations together

I investigate here sequential applications of these cleaning and pruning operations, focusing on individualistic admissibility. Given a strategic game $\mathbb{G}$, let $t(\mathbb{G})$ be either $pr(\mathbb{G})$ or $cl(\mathbb{G})$. A *sequence of transformation of length k* is any $t^k(\mathbb{G})$ for $k \geq 0$, where $t^0(\mathbb{G}) = \mathbb{G}$ and $t^{k+1}(\mathbb{G}) = t(t^k(\mathbb{G}))$. A sequence of transformation $t^k(\mathbb{G})$ is a *transformation fixed point* whenever both $cl(t^k(\mathbb{G})) = t^k(\mathbb{G})$ and $pr(t^k(\mathbb{G})) = t^k(\mathbb{G})$.

Table 3 Counterexample to commutativity, with $X = \Pi_{i \in I} S_i$

$\mathbb{G}$	t_1	t_2
s_1		1
s_2	1, 2	1, 2

Table 4 A game with two different fixed points, with $X = \Pi_{i \in I} S_i$

$\mathbb{G}$	t_1	t_2	t_3
s_1	1	2	
s_2	1, 2		
s_3	1		1

The first notable fact about cleaning and pruning sequences, already observed by van Hees and Roy (2008), is that these operations do not in general commute. Table 3 is a counterexample, with $\theta_2([t_1]) = t_1$. The two operations do commute, however, in the single-agent case.

Fact 6 (van Hees and Roy 2008) *For any strategic game $\mathbb{G}$ with only one agent, intention set ι_i and picking function θ_i : $pr(cl(\mathbb{G})) = cl(pr(\mathbb{G}))$.*

Sequential cleaning and pruning create new possibilities for empty fixed points, since neither the existence of a cleaning core nor of safe outcomes, and not even a combination of the two criteria, are sufficient to ensure non-emptiness. Furthermore, there might not be a unique fixed point, as in the example of Table 4, with $\theta_1(\{s_1, s_2\}) = s_2$, $\theta_1(\{s_1, s_2, s_3\}) = s_1$ and $\theta_1(\{s_2, s_3\}) = s_2$. If the agents start by cleaning this game, then t_3 is ruled out in the first round. A round of pruning then leaves only s_1 for agent 1, after which a round of cleaning leaves him with inconsistent intentions. If, on the other hand, the agents first prune $\mathbb{G}$, then strategy s_1 of agent 1 is eliminated. After a round of cleaning and one further round of pruning only (s_2, t_1) remains, which is the obvious fixed point of this sequence.

Ignoring redundant transformations, all sequences of cleaning and pruning reach a fixed point in a finite number of steps, for every finite strategic game. Non-emptiness of this fixed point is ensured by the following strengthening of safety for pruning and cleaning the core. The outcome x of profile $\sigma \in \Pi_{i \in I} S_i$ is:

- *Safe for iterated transformations at stage 1* whenever, for all $i \in I$:
  1. $\pi(\sigma(i)) \cap \iota_i \neq \emptyset$.
  2. $\theta_i([\sigma(i)]^{\mathbb{G}}_{\approx}) = \sigma(i)$.
- *Safe for iterated transformations at stage $n + 1$* whenever it is safe for iterated transformation at stage n and for all $i \in I$:
  1. $\pi^{t^n(\mathbb{G})}(\sigma(i)) \cap \iota_i^{t^n(\mathbb{G})} \neq \emptyset$.
  2. $\theta_i([\sigma(i)]^{t^n(\mathbb{G})}_{\approx}) = \sigma(i)$.
- *Safe for iterated transformations* whenever it is safe for transformation at all n.

Fact 7 *For any strategic game $\mathbb{G}$, intention profile ι and profile of picking function Θ, if $\pi(\sigma)$ is safe for transformation in $\mathbb{G}$ then for all fixed points $t^{\#}(\mathbb{G})$, $\sigma \in \Pi_i t^{\#}(S_i)$.*

The presence of safe outcomes is thus sufficient not only to ensure that a game has no empty fixed point of iterated transformation but also that all fixed points have a non-empty intersection. It is a direct corollary that it is sufficient to avoid empty fixed-points that all agents intend outcomes which are safe for iterated transformations, not necessarily the same ones.

Corollary 2 *For any strategic game $\mathbb{G}$, intention profile ι and profile of picking function Θ, if for all $i \in I$ there is a $x = \pi(\sigma) \in \iota_i$ that is safe for iterated transformation, then for all fixed points $t^{\#}(\mathbb{G})$, $\Pi_i t^{\#}(S_i)$ is not empty.*

It is still open, however, whether intending safe outcome is necessary, in the general case. We do know, however, that it is necessary if the picking function satisfies the following constraint, recalling Sen's (2002) "property α": let a picking function θ_i be called *consistent* if $\theta_i(X) = s_i$ whenever $\theta_i(Y) = s_i$, $X \subseteq Y$ and $s_i \in X$.

Fact 8 *For any strategic game $\mathbb{G}$, intention profile ι and profile of consistent picking functions Θ, if $\sigma \in \Pi_i t^{\#}(S_i)$ for all fixed points $t^{\#}(\mathbb{G})$, then $\pi(\sigma)$ is safe for transformation in $\mathbb{G}$.*

From this and Corollary 2 we thus get that, for consistent picking functions, intending outcomes that are safe for iterated transformation is indeed necessary and sufficient to avoid empty fixed points.

Corollary 3 *For any strategic game $\mathbb{G}$, intention profile ι and profile of consistent picking function Θ, $\Pi_i t^{\#}(S_i)$ is not empty iff for all $i \in I$ there is a $x \in \iota_i$ that is safe for iterated transformation.*

By intending outcomes which are safe for iterated transformations the agents strongly acknowledge that they are interacting with agents whose intentions also transform the decision problem. The fact that the pruning and cleaning do commute when there is only one agent is in that respect illuminating: in the single agent case it is rather "simple" to transform one's decision problem on the basis of what one intends. This highlights, once again, the subtleties that interaction brings into intentions-based reasoning.

5 Discussion: intentions and preferences

In this closing section I want briefly to address the question of the relation between, on the one hand, intentions sets, cleaning and pruning and, on the other hand, preferences and known game-theoretic solutions concepts. This will put the results of the previous sections in perspective, and will position the main claims of this paper in relation to known claims in philosophy of action and game theory.

It has been claimed in philosophy of action that previously adopted intentions are not reducible to compounds of beliefs and desires—see e.g. (Bratman 1987). Intentions, like desires, are motivational states. They are, however, subject to stronger norms of coherence and consistency (Bratman 2009; Velleman 2008; Roy 2008). An agent, for instance, can rationally desire inconsistent or impossible things, but he cannot rationally intend such things. Intentions are furthermore thought of as providing a stronger

commitment to action and to practical reasoning than desires—again see (Bratman 1987) for a discussion of the "commitment to action" and the "reasoning-centred commitment" of intentions. Finally, intentions are more resistant to reconsideration than desires. While one's changes of preferences and desires can be irregular and not necessarily based on reasons, the reconsideration of rational intentions is subject to much stronger constraints, and arguably occurs less frequently.

This thesis is of great importance for the theory of intentions, but the analysis of cleaning and clustering does not need to, and in fact does not take a stance on it. I have shown that if one accepts that previously adopted intentions are states that carry a reasoning-centred commitment and that are subject to consistency requirements, then significant complications arise in situations of interaction.

It is clear that the analysis of iterated transformation presupposes many of the purportedly distinguishing features of intentions. The filtering effect and the standard of relevance, which cleaning and pruning are intended to reflect, are parts of what it means for intentions to carry a reasoning-centred commitment (Bratman 1987). Furthermore, intention sets satisfy some of the consistency constraints mentioned above: agents cannot intend impossible things ($\iota_i \neq \emptyset$) and much of the discussion in Sect. 3 revolved around avoiding cases of unrealizable intentions.

This analysis of the reasoning-centred commitment, however, does not rest on whether these features really do distinguish intentions from beliefs and preferences. If it can be shown that they really do so, then cleaning and pruning constitute a genuine extension of game-theoretic models. But it is not the purpose of the analysis I propose in this paper to provide such an argument. Rather, the goal was to show that *if* one accepts that previously adopted intentions are states that carry a reasoning-centred commitment and that are subject to consistency requirements, *then* important complications arise in situations of interaction. Indeed, this thesis remains of interest even if one denies that the reasoning-centred commitment and the consistency constraints suffice to sustain the non-reductionist thesis. For in this case one has available the tools to investigate the conditions under which cleaning and pruning can be recaptured by standard game-theoretical solution concepts, as the current model is sufficiently general to support this investigation.

This raises another point, namely that the philosophical thesis concerning the (non-) reducibility of intentions to beliefs and desires does not preclude that intentions could or should in some way be aligned to these cognitive and conative states, and the model presented here is neutral regarding this issue as well. Indeed, in this paper no constraints have been imposed on the relation between intentions and preferences, which means that the model is sufficiently general to carry investigations along the line proposed by Sen (2005), who argues that intentions and commitments are of interest especially when they diverge from preferences, or to explore cases where intentions and preferences do converge, as e.g. (van Hees and Roy 2008) and (Roy 2008). Observe that the latter case could be used to further the reductionist perspective, by investigating the possibility of re-capturing cleaning in terms of a dynamic elimination of strictly dominated strategies (van Benthem 2007), via an alignment of intentions to "Bayesian rationality" (Aumann 1987).

The analysis proposed here is thus neither committed to the reductionist thesis, nor to a particular view of the relation between intentions and preferences, and this

should be seen as one of its assets. The aim of this paper was to bring the analysis of intention-based transformation of decision problems to situations of interaction, and to show important issues that arise at this level. The lesson is valuable, I have just argued, for both side of the reductionist/non-reductionist debate.

6 Conclusion

I have taken steps towards filling an important gap in the literature on philosophy of action, artificial intelligence and multi-agent systems concerning the role of intentions in the transformation of a decision problem, by studying this role in genuine interactive situations. I have shown that in such situations new issues readily arise as agents can make each others' intentions unrealizable or even inconsistent. I have characterized conditions under which such cases are avoided and in so doing have revealed an important interactive characteristic of intention-based transformation of decision problems, one which has hitherto passed unnoticed in the philosophical and formal literature.

At this point two future directions of research seem most pressing. First is the issue mentioned in the introduction: comparing and relating the current approach to the existing formal work in AI or multi-agent systems. This would not only gear the present approach towards application, but would also surely unveil even more new issues concerning intentions in interaction. Second, it would be very worthwhile to make explicit the recurring idea of taking others intentions into account by enriching the game theoretical models I use with the information that players might have about each others' choices, preferences and intentions. This kind of epistemic perspective has proved fruitful in game theory, see e.g. (Brandenburger 2007), and would move us even further towards a unified account of rational interaction.

Acknowledgements I am very grateful to Martin van Hees, Johan van Benthem, Richard Bradley and the two anonymous referees of *Synthese* for detailed comments on the various versions of this paper. The paper also profited greatly from discussions with David Israel, and with the participants at LOII'08 in Hamburg and LOFT'08 in Amsterdam. Financial support from the Conseil de Recherches en Sciences Humaines du Canada, Scholarship scholarship # 752-2006-0345, is gratefully acknowledged.

Appendix

Proof of Fact 1 *For any strategic game* $\mathbb{G}$ *and intention profile* ι, S^* *is empty iff* $cl^{\#}(\mathbb{G})$ *is empty as well, for individualistic cleaning.*

By Definition, $S^* \neq \emptyset$ is the same as saying that we can find a $\sigma \in S^*$ such that for all i, $\pi^{S^*}(\sigma(i)) \cap \iota_i \neq \emptyset$. We show by induction that $\pi(S^k) = X^{cl^k}$, for all k. This is enough to show the equivalence, for then we know that $X^{cl^{\#}} \cap \iota_i \neq \emptyset$, which we know is the same as $cl^{\#}(\mathbb{G})$ being non-empty. The basic case of the induction, $k = 0$, is trivial. For the induction step, assume the claim is proved for k. We have

that $x \in \pi(S^{k+1})$ iff there is a $\sigma \in S^{k+1}$ such that $\pi(\sigma) = x$. This in turns happens iff $\pi^{S^k}(\sigma(i)) \cap \iota_i \neq \emptyset$, for all i. But by the inductive hypothesis this just says that $\pi(\sigma(i)) \cap X^{cl^k} \cap \iota_i \neq \emptyset$, which is just the definition of x being in X^{x+1}.

Proof of Fact 2 *For $\mathbb{G}$ an arbitrary strategic game, $cl^{\#}(\mathbb{G}) = cl(\mathbb{G})$ for cleaning with altruistic admissibility.*

We show that $cl(cl(\mathbb{G})) = cl(\mathbb{G})$. Given the definition of the cleaning operation, it is enough to show that $cl(cl(S_i)) = cl(S_i)$ for all i. It should be clear that $cl(cl(S_i)) \subseteq cl(S_i)$. It remains to show the converse. Assume that $s_i \in cl(S_i)$. Since cleaning is done with altruistic admissibility, this means that there is a σ such that $\sigma(i) = s_i$ and a $j \in I$ such that $\pi(\sigma) \in \iota_j$. But then $\sigma(i') \in cl(S_{i'})$ for all $i' \in I$, and so $\sigma \in \Pi_{i \in I} cl(S_i)$. This means that $\pi(\sigma) \in X^{cl}$, which in turn implies that $\pi^{cl}(\sigma) \in \iota_j^{cl}$. We thus know that there is a $\sigma \in \Pi_{i \in I} cl(S_i)$ such that $\sigma(i) = s_i$ and a j such that $\pi^{cl}(\sigma) \in \iota_j^{cl}$, which means that $s_i \in cl(cl(S_i))$.

Proof of Fact 3 *For any strategic game $\mathbb{G}$, intention profile ι and cleaning with altruistic admissibility, there is, for all i, a realizable $x \in \iota_i$ iff $cl^{\#}(\mathbb{G})$ is not empty.*

There is a realizable $x \in \iota_i$ for all i iff for all i there is a σ such that $\pi(\sigma) \in \iota_i$. But this is this same as saying that for all j there is a strategy s_j such that $\sigma(j) = s_j$ and an i such that $\pi(\sigma) \in \iota_i$ which, by Fact 2, means that $cl^{\#}(\mathbb{G})$ is not empty.

Proof of Fact 4 *For all strategic game $\mathbb{G}$ and agent $i \in I$, $pr^{\#}(S_i) \neq \emptyset$.*

This is shown by induction on $pr^k(\mathbb{G})$. The basic case is trivial. For the induction step, observe that the picking function θ_i is defined for the whole power set of S_i. This means, given the inductive hypothesis, that $\theta_i([s_i]_{\approx}^{pr^k(\mathbb{G})})$ is well-defined and in $[s_i]_{\approx}^{pr^k(\mathbb{G})}$ for any $s_i \in pr^k(S_i)$, which is enough to show that $pr^{k+1}(S_i)$ is also not empty.

Proof of Fact 5 *For any strategic game $\mathbb{G}$, intention profile ι, profile of picking function Θ and for all $i \in I$, $\iota_i^{pr^{\#}} \neq \emptyset$ iff there is a $\pi(\sigma) \in \iota_i$ safe for pruning in $\mathbb{G}$.*

From right to left. Take any $x \in \iota_i^{pr^{\#}}$. By definition we know that there is a $\sigma \in \Pi_{i \in I} pr^{\#}(S_i)$ such that $\pi(\sigma) = x$. But this happens iff $\sigma \in \Pi_{i \in I} pr^k(S_i)$ for all k, and so that $\theta_i([\sigma(i)]_{\approx}^{pr^k(\mathbb{G})}) = \sigma(i)$ also for all k, which in turns means that x is safe for pruning in $\mathbb{G}$. Left to right, take any such $\pi(\sigma) \in \iota_i$. We show that $\pi(\sigma) \in X^{pr^k}$ for all k. The basic case is trivial, so assume that $\pi(\sigma) \in X^{pr^k}$. We know by definition that $\pi(\sigma)$ is safe for pruning at k, which gives automatically that $\pi(\sigma) \in X^{pr^{k+1}}$.

Proof of Fact 7 *For any strategic game $\mathbb{G}$, intention profile ι and profile of consistent picking function Θ, if $\pi(\sigma)$ is safe for transformation in $\mathbb{G}$ then for all fixed points $t^{\#}(\mathbb{G})$, $\sigma \in \Pi_i t^{\#}(S_i)$.*

This is shown by induction on k for an arbitrary fixed point $t^k(S_i)$. The proof is a direct application of the definition of safety for transformation.

Proof of Fact 8 *For any strategic game $\mathbb{G}$, intention profile ι and profile of consistent picking function Θ, if $\sigma \in \Pi_i t^{\#}(S_i)$ for all fixed points $t^{\#}(\mathbb{G})$, then $\pi(\sigma)$ is safe for transformation in $\mathbb{G}$.*

We show by "backward" induction that $\pi(\sigma)$ is safe for transformation at any k for all sequences $t^k(\mathbb{G})$. For the basic case, take k to be the length of the longest, non-redundant fixed point of $\mathbb{G}$. We show that $\pi(\sigma)$ is safe for transformation at stage k for all sequences of that length. Observe that by the choice of k all $t^k(\mathbb{G})$ are fixed points. We thus know by assumption that $\sigma \in \Pi_{i\in I} t^k(S_i)$. But then it must be safe for transformation at stage k. If clause (1) were violated at one of these, say $t'^k(\mathbb{G})$, then we would have $cl(t'^k(\mathbb{G})) \neq t'^k(\mathbb{G})$, against the fact that $t'^k(\mathbb{G})$ is a fixed point. By the same reasoning we know that clause (2) cannot be violated either. Furthermore, by the fact that $t'^{k+1}(\mathbb{G}) = t'^k(\mathbb{G})$, we know that it is safe for transformation at all stages $l > k$.

For the induction step, take any $0 \leq n < k$ and assume that for all sequences $t^{n+1}(\mathbb{G})$ of length $n+1$, $\pi(\sigma)$ is safe for transformation at stage $n+1$. Take any $t^n(\mathbb{G})$. By our induction hypothesis, that $\pi(\sigma)$ is safe for transformation at both $cl(t^n(\mathbb{G}))$ and $pr(t^n(\mathbb{G}))$. This secures clause (2) of the definition of safety for transformation, and also gives us that $\sigma \in \Pi_{i\in I} t^n(S_i)$. Now, because it is safe for transformation in $cl(t^n(\mathbb{G}))$, we know that $\pi^{cl(t^n(\mathbb{G}))}(\sigma(i)) \cap \iota_i^{cl(t^n(\mathbb{G}))} \neq \emptyset$ for all i. But since $\pi^{cl(t^n(\mathbb{G}))}(\sigma(i)) \subseteq \pi^{t^n(\mathbb{G})}(\sigma(i))$, and the same for the intention set, we know that $\pi^{t^n(\mathbb{G})}(\sigma(i)) \cap \iota_i^{t^n(\mathbb{G})} \neq \emptyset$ for all i. For condition (2), we also know that $\theta_i[\sigma(i)]_{\approx}^{cl(t^n(\mathbb{G}))} = \sigma(i)$ for all i from the fact that $\pi(\sigma)$ is safe for transformation at stage $n+1$. By Lemma 1 (below) and the assumption that θ_i is consistent for all i, we can conclude that $\theta_i([\sigma(i)]_{\approx}^{t^n(\mathbb{G})}) = \sigma(i)$, which completes the proof because we took an arbitrary $t^n(\mathbb{G})$.

Lemma 1 *For any strategic game $\mathbb{G}$ and intention set ι_i and strategy $s_i \in cl(S_i)$, $[s_i]_{\approx}^{\mathbb{G}} \subseteq [s_i]_{\approx}^{cl(\mathbb{G})}$.*

Proof Take any $s_i' \in [s_i]_{\approx}^{\mathbb{G}}$. Since $s_i \in cl(S_i)$, we know that there is a $\sigma_{j\neq i}$ such that $\pi(s_i, \sigma_{j\neq i}) \in \iota_i$. But because $s_i' \approx s_i$, it must also be that $\pi(s_i', \sigma_{j\neq i}) \in \iota_i$, and so that $s_i' \in cl(S_i)$. Now, observe that $\{\sigma \in \Pi_{i\in I} cl(S_i) : \sigma(i) = s_i\} \subseteq \{\sigma \in S_i : \sigma(i) = s_i\}$, and the same for s_i'. But then, because $s_i' \approx s_i$, it must also be that $s_i' \in [s_i]_{\approx}^{cl(\mathbb{G})}$.

References

Alchourron, C. E., Gardenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic, 50*(2), 510–530.

Alonso, F. (2008). Shared intention, reliance, and interpersonal obligations. Ph.D. thesis, Stanford University.

Apt, K. (2007). The many faces of rationalizability. *The B.E. Journal of Theoretical Economics, 7*(1). Article 18.

Aumann, R. (1987). Correlated equilibrium as an expression of Bayesian rationality. *Econometrica, 55*, 1–18.

Aumann, R. (1999). Interactive epistemology I: Knowledge. *International Journal of Game Theory, 28*, 263–300.

Brandenburger, A. (2007). The power of paradox: Some recent developments in interactive epistemology. *International Journal of Game Theory, 35*, 465–492.
Bratman, M. (1987). *Intention, plans and practical reason*. London: Harvard University Press.
Bratman, M. (1999). *Faces of intention; selected essays on intention and agency*. Cambridge University Press.
Bratman, M. (2009). Intention, belief, practical, theoretical. In J. Timmerman, J. Skorupski, & S. Robertson (Eds.), *Spheres of reason*. Oxford: Oxford University Press.
Bratman, M., Israel, D., & Pollack, M. (1991). Plans and resource-bounded practical reasoning. In J. Pollock & R. Cummins (Eds.), *Philosophy and AI: Essays at the interface* (pp. 7–22).
Cohen, P., & Levesque, H. (1990). Intention is choice with commitment. *Artificial Intelligence, 42*(2–3), 213–261.
Georgeff, M., Pell, B., Pollack, M., Tambe, M., & Wooldridge, M. (1999). The belief-desire-intention model of agency. In J. Muller, M. Singh, & A. Rao (Eds.), *Intelligent Agents V*, Vol. 1365 of *Springer-Verlag Lecture Notes in AI*.
Horty, J., & Pollack, M. (2001). Evaluating new options in the context of existing plans. *Artificial Intelligence, 127*(2), 199–220.
Osborne, M., & Rubinstein, A. (1994). *A course in game theory*. MIT Press.
O'Shaughnessy, B. (1973). Trying (As the mental "Pineal gland"). *The Journal of Philosophy, 70*(13, On Trying and Intending), 365–386.
Parikh, R. (2002). Social software. *Synthese, 132*(3), 187–211.
Pollack, M. (1992). The uses of plan. *Artificial Intelligence, 57*(1), 43–69.
Rott, H. (2001). *Change, choice and inference: A study of belief revision and nonmonotonic reasoning*, Oxford Logic Guides. Oxford: Oxford University Press.
Roy, O. (2008). Thinking before acting: Intentions, logic, rational choice. Ph.D. thesis, Institute for Logic, Language and Computation, University of Amsterdam.
Searle, J. (1983). *Intentionality*. Cambridge University Press.
Searle, J. (1995). *The construction of social reality*. London: Allen Lane.
Sen, A. (2002). *Rationality and freedom*. Cambridge, MA: Harvard University Press.
Sen, A. (2005). Why exactly is commitment important for rationality? *Economics and Philosophy, 21*(01), 5–14.
Tuomela, R. (1995). *The importance of Us: A philosophical study of basic social notions*. Stanford: Stanford University Press.
van Benthem, J. (2006). Epistemic logic and epistemology, the state of their affairs. *Philosophical Studies, 128*, 49–76.
van Benthem, J. (2007). Rational dynamic and epistemic logic in games. *International Game Theory Review, 9*(1), 13–45.
van der Hoek, W., Jamroga, W., & Wooldrige, M. (2007). Towards a theory of intention revision. *Synthese, 155*, 265–290.
van Hees, M., & Roy, O. (2008). Intentions and plans in decision and game theory. In B. Verbeek (Ed.), *Reasons and intentions* (pp. 207–226). Ashgate Publishers.
Velleman, J. (2008). What good is a will?. In A. Leist & H. Baumann (Eds.), *Action in context* (pp. 193–215). Berlin/New York.
Wallace, R. (2006). *Normativity and the will*. Oxford University Press.
Wooldridge, M. (2000). *Reasoning about rational agents*, *Intelligent robotics and autonomous agents series*. Cambirdge: MIT Press.