Thinking before acting: intentions, logic, rational choice

Roy, O.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
In this thesis I have proposed a theory of practical reasoning which drew on three contemporary paradigms: instrumental rationality from decision and game theory, epistemic reasoning from philosophical logic and computer science, and planning agency from philosophy of action. This provides a unified theory of rational planning agency, which is a theory of how agents deliberate when they take into account the demands of instrumental rationality, their background of future-directed intentions and the information they have about the rationality, intentions and information of others.

7.1 Review of the chapters

I have shown in Chapter 2 that such a broad perspective can account for personal coordination in extensive decision problems, because rational planning agents are able to break ties between equally desirable options. In Chapter 3 I brought this tie-breaking effect to the level of interactive situations, and I have shown that it provides a natural anchor for interpersonal coordination in Hi-Lo games. With the help of epistemic models for these games, I have been able to study explicitly how mutual knowledge of intentions also foster coordination in games. By the same token I was able relate the intention-based explanation of coordination to other accounts in the game theoretical literature, and to circumscribe better the differences between coordination and fully cooperative shared agency.

In Chapter 4 I have studied how rational planning agents transform their decision problem on the basis of what they intend, a phenomenon called the reasoning-centered commitment in philosophy of action. I have shown that this aspect of intention-based practical reasoning is especially sensitive to interactions. When many planning agents simultaneously transform the decision problem they face, it becomes crucial that they take each others’ intentions into account.

Chapter 5 merged the considerations of the previous chapters into a unified picture of rational planning agency. Using dynamic epistemic logic, I have
been able to relate the informational components of intentions-based deliberation with the active process of decision problem transformation. I have also shown that this framework really does capture phenomena at the intersection of planning agency and instrumental rationality, such as the relation between the game-theoretic solution concept of elimination of dominated strategies and the filtering of intention-inconsistent options. Finally, I have provided this framework with axiomatic proof systems, which give an explicit representation of practical reasoning in games with intentions.

In Chapter 6 I explored the philosophical basis of this theory of intention-based deliberation, looking at where the various norms of consistency and coherence of intentions come from. This led to hybrid pragmatism, an attempt to explain these norms in terms of similar norms which apply on acceptances in deliberation. I have argued that hybrid pragmatism is a plausible alternative to the main contemporary proposals in philosophy of action, because it does justice to both the cognitive and the pragmatic side of the norms of consistency and coherence. Furthermore, it provides a natural explanation of how future-directed intentions influence practical reasoning, and as such helps us to see better how the various pieces encountered in this thesis fit together.

7.2 Open questions

Of course, many questions were left unanswered along the way. Instead of reviewing them, I will rather present three broad research themes that this thesis opens up. The first relates more specifically to logic, the second to the theory of intentions and the third to game theory. I think, however, that from a general point of view each of them is relevant to all these fields.

I have many times mentioned that the present framework is in great need of a more elaborate theory of intention revision. This poses challenging problems from a logical point of view. Dynamic epistemic logics for belief revision have been extensively studied\(^1\). These systems have very interesting logical properties, in terms of axiomatizations and expressive power, and it is surely worth looking at how they would transfer to logic for intention revision. What it more, developing the logic of intention revision is surely a good way to establish the connection between the approach adopted in this thesis and BDI architectures. As I mentioned at the end of Chapter 5, BDI models are the main contemporary logical approach to intention-based practical reasoning. The road I have taken here did not quite allow for a systematic comparison, but at this point the issue is, to say the least, pressing.

From the philosophical point of view, hybrid pragmatism and the theory of acceptances in deliberation that I used in Chapter 6 are “new” issues that deserve

\(^1\)Girard [2007], van Benthem [2007] and Baltag and Smets [2006] are especially interesting from the current perspective.
much more detailed scrutiny. But the investigation that I carried out in Chapter 4 has also unveiled, I think, an important gap in the “core” theory of intentions. The intricacies of intention-based transformation of decision problems in interaction situations have been widely overlooked in the philosophical literature. In comparison, the notion of individual intentions with a “we content” that I used in Chapter 3 has attracted much attention, and raised important questions about conditions under which agents are justified in forming them. Similar questions obviously also apply to intention-based transformations of decision problems. Are agents always justified in cleaning or pruning their option set, even when they are uncertain about the intentions of others? If not, can one phrase the appropriate conditions of justification in terms of mutual knowledge of intentions, as it is done for intentions with a “we” content? These issues are of the greatest importance for the theory of intentions, because they concern the very “regularities which connect intentions with each others, with associated processes and activities, and with characteristic ‘inputs’ and ‘outputs’.” [Bratman, 1987, p.9] In short, they concern what intentions are.

The place of rational deliberation with intentions in interactive situations also poses very interesting problems from the point of view of game theory. As mentioned in Chapter 3, the intention-based account of coordination in Hi-Lo games occupies a point somewhere in the middle ground between purely competitive and genuinely cooperative scenarios. It makes crucial use of intentions of the form “I intend that we...” which, even though they are essentially individual states, have a clear social character. As such, intention-based practical reasoning is a bridge between competitive and cooperative game theory. It thus offers a plausible alternative to the Nash program [Serrano, 2005], a well-known attempt to translate cooperative into non-cooperative frameworks.

D. Davidson’s seminal contribution [1980] to contemporary philosophy of action was profoundly influenced by his familiarity with models of instrumental rationality from theoretical economics, and especially with decision theory [Malpas, 2005]. Since then, however, these disciplines have mostly evolved in parallel, and only recently can we see a renewal of interest in genuinely interdisciplinary work on rational decision making, witness e.g. the work of Parikh [2002], van Benthem [2006] and Bacharach [2006]. I have written this thesis with the conviction that such interdisciplinary approaches are fruitful, and my hope is that I have conveyed this conviction to the reader.

\(^2\)Harp [2008] is a notable exception.