MARTIN DUFWENBERG and JOHAN LINDÉN

INCONSISTENCIES IN EXTENSIVE GAMES
Common Knowledge Is Not the Issue

ABSTRACT. In certain finite extensive games with perfect information, Cristina Bicchieri (1989) derives a logical contradiction from the assumptions that players are rational and that they have common knowledge of the theory of the game. She argues that this may account for play outside the Nash equilibrium. She also claims that no inconsistency arises if the players have the minimal beliefs necessary to perform backward induction. We here show that another contradiction can be derived even with minimal beliefs, so there is no paradox of common knowledge specifically. These inconsistencies do not make play outside Nash equilibrium plausible, but rather indicate that the epistemic specification must incorporate a system for belief revision. Whether rationality is common knowledge is not the issue.

1. INTRODUCTION

In many theories of economic behavior, agents are assumed to be rational, in the sense that they maximize expected utility, and to be correctly informed about the part of their environment that is relevant to their decisions. When agents engage in strategic interaction these decisions depend heavily on information regarding other players' rationality, information and beliefs. The assumptions of perfectly informed players often then entails common knowledge of certain facts, meaning that if one player knows, believes or expects a proposition, the others are aware that he does so.

Even though optimal strategic behavior is often described as utility maximization given beliefs about other players' actions, these beliefs are usually not explicitly incorporated in game theory. Without a theory of beliefs, game theory is conceptually incomplete. Introducing one would help resolve important foundational problems, such as selecting among the plethora of equilibrium concepts. It may also make game theory more applicable to situations where information exchange is important, or where players try to confuse each other by seemingly irrational actions. Cristina Bicchieri (1989) presents an extension of game theory, including a formal epistemic logic that describes players' beliefs about each other at every information set and the reasoning behind their decisions. She uses this

Erkenntnis **45:** 103–114, 1996.

theory, first to investigate the foundations of the backward induction solution, and then to model preplay communication and strategic irrationality to confuse an opponent.

Our paper concentrates on her main result, that in certain extensive finite games with perfect information, if players are rational and have common knowledge of the theory of the game, then the theory becomes inconsistent. She argues that this may account for play outside Nash equilibrium by rational players. She also claims that the theory will be consistent if players only have the minimal beliefs necessary for the reasoning behind the backward induction solution.

In this paper we refute this last claim. Though the contradiction derived by Bicchieri disappears with minimal beliefs, we show that another one may be derived even in this case. Thus, there is no paradox of common knowledge specifically.

Inconsistencies as defined by Bicchieri (1989) arise when a player at one node makes a prediction that rules out his reaching another one of his nodes later in the game tree. His theory will then become inconsistent if that particular node is reached. This happens irrespective of any assumptions of common knowledge. As soon as the players are endowed with enough information to perform the backward induction argument, they will rule out some branches of the game tree, thus making the theory inconsistent there. These inconsistencies do not indicate that deviations from the backward induction solution become plausible, but rather indicate that the epistemic specification must incorporate a system for belief revision.[1]

Though our analysis centers on Bicchieri (1989) it relates to other work too. The logic of backward induction has recently been scrutinized and criticized by among others Basu (1990, 1994), Ben-Porath (1994), Bicchieri (1989, 1992), Bicchieri and Antonelli (1993), Binmore (1987,1995), Bonanno (1991), Gul (1995), Pettit and Sugden (1989), Reny (1988, 1993), and Sugden (1991). In one way or another, the assumption of common knowledge of rationality plays an important role in this work. Because of this, opinion seems to have become widespread that common knowledge of rationality is an incoherent, troublesome, or even impossible assumption to make (see, however, Aumann (1995) for a model in which common knowledge of rationality implies backward induction in finite extensive games with perfect information). We downplay the role played by this assumption. Whether rationality is common knowledge is not a key issue.

Section 2 introduces the game analyzed by Bicchieri, along with the formal language that will be used. Section 3 reproduces Bicchieri's result, that if players are rational and have common knowledge of the theory of the game, the theory becomes inconsistent. In section 4 we show that, contrary

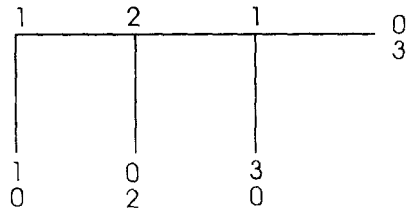


Figure 1.

to Bicchieri's conclusion, even when less knowledge is assumed, inconsistencies arise. We also comment on the extent to which these inconsistencies arise in games in general. In Section 5 we show how the inconsistencies can be jettisoned.

2. THE STATE OF PLAY

Bicchieri (1989) analyzes the three-stage "take-it-or-leave-it" game (TOL(3)), invented by Reny (1988) and reproduced in Figure 1. Player 1 is male, player 2 is female.

Like Bicchieri (1989) we restrict our attention to finite extensive games with perfect information. Any such game possesses at least one pure strategy sub-game perfect Nash equilibrium that can be found by backward induction (Kuhn 1953) The backward induction solution has each player choosing down at every opportunity. Note that the payoffs will be (1,0), since player 1 figures out that player 2 can figure out that 1 rationally would choose down at his second node so 2 would choose down at hers. Therefore 1 will choose down to start with, thereby ending the game.

Bicchieri (1989) explicitly models the players' beliefs and the reasoning they perform from these. She uses an epistemic modal logic with two types of modal operators, knowledge and belief, developed by Hintikka (1962) for a philosophical investigation of these notions, and first used in game theoretic analysis by Bacharach (1987). Aumann and Brandenburger (1991, p. 45) criticize the use of two epistemic modalities as unnecessarily complicated for game theoretic analysis and demonstrate that a simplified logic suffices, with only one type of modal operator, called "know", meaning that the player ascribes probability one to a proposition.[2] Here we follow their suggestion, except that we call the only modality "believe".

The logical language contains as primitive propositions $R_i (i = 1, 2, \ldots)$, meaning "player i is rational" in the sense that he maximizes expected utility. Propositions can be combined with the usual connectives of propo-

sitional logic: $\neg$, $\wedge$, $\vee$, $\Rightarrow$ and $\Leftrightarrow$ for negation, conjunction, disjunction, material implication and equivalence respectively. Finally, the epistemic modality B_i operates on any proposition p, $B_i p$ meaning "player i believes proposition p". E.g. $B_1 B_2 R_1$ is interpreted "player 1 believes that player 2 believes him to be rational". This formal language is used only to represent beliefs of the players, while our own analysis is done informally. The players are assumed to be rational and able to carry out logical deductions from their beliefs.

Now define "the theory of the game" by:

(i) the structure of the game (tree, player and node partitionings, payoffs)
(ii) the players' decision rules
(iii) the players' beliefs

In what follows: (i) the considered games will be of finite extensive form with perfect information, (ii) players will be rational and they never tremble (make mistakes), (iii) all players believe in the structure of the game and the fact that no players ever tremble. Moreover, the players will have beliefs regarding each others' rationality and beliefs. Two important cases will be analyzed: "common knowledge" of the theory of the game, and "minimal beliefs". These two concepts will be defined in Sections 3 and 4 respectively. The players optimize given their beliefs which define "the theories of the individual players". These belief sets are endowed to the players before play starts. Only true beliefs are involved and the belief sets are expanded as play through the game provides evidence regarding other players' rationality or beliefs. If a player acquires new information contradicting his previous beliefs, he can prove any choice optimal since anything follows from a contradiction.

3. INCONSISTENCIES WITH COMMON KNOWLEDGE

In general "common knowledge of a certain proposition" means that all player believe that the proposition is true, all players believe that each other player believes this, ... et cetera ad infinitum. For example with only two players, and disregarding beliefs about oneself, common knowledge of each other's rationality can be formalized as follows:

	player 1 believes:	**player 2 believes:**
	R_2	R_1
(1)	$B_2 R_1$	$B_1 R_2$
	$B_2 B_1 R_2$	$B_1 B_2 R_1$
	...	...

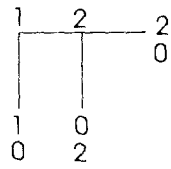


Figure 2.

Now assume that the players in the TOL(3) game have common knowledge of the theory of the considered game. Among other things this implies common knowledge of rationality and thus (1) holds.

With reference to Reny, Bicchieri states that

> an obvious requirement a theory of the game has to satisfy is that it be free from contradictions at every information set. If a player were to find herself at an information set with which the theory she uses is inconsistent, she would be deprived of a theory upon which to base her decisions. This would leave the other players (and the game theorist) without a theory, too, since they would become unable to predict what she will do, and would therefore be unable to decide what to do themselves. (Bicchieri 1989, p. 71; cf. Reny 1988, p. 24)

If (1) holds the theory becomes inconsistent at 2's node. Player 2 cannot in any way explain the deviation by 1. To explain it she would have to believe that 1 is not rational ($\neg R_1$), e.g. an eternal "forward" player, or that 1 believes her not to be rational ($\neg B_1 R_2$), or that 1 believes that she believes 1 not to be rational ($\neg B_1 B_2 R_1$), i.e., 2 must believe

$$\neg R_1 \vee \neg B_1 R_2 \vee \neg B_1 B_2 R_1$$

But none of these explanations are feasible since, by (1), she believes the opposite. Bicchieri concludes: "*Allowing common knowledge of the theory of the game makes that theory inconsistent*" (Bicchieri 1989, p. 79).

Note that the inconsistency result is really not one of common knowledge, but rather of beliefs that are iterated to a sufficient degree. If player 2 believes $R_1 \wedge B_1 R_2 \wedge B_1 B_2 R_1$ that is enough to generate an inconsistency at her node.

Note also that in principle the same point could be made in the TOL(2) game in Figure 2.

4. INCONSISTENCIES WITH MINIMAL BELIEFS

Define the "minimal beliefs compatible with backward induction", the "minimal beliefs" for short, as those generated by the following algorithm:

Step one. Specify a rational local strategy to every player who owns a node that is followed only by end nodes. This requires no beliefs; the players need only compare different payoffs.

Steps two, three, ... etc. until the root is reached. Consider each node followed only by nodes considered at previous steps and, possibly, end nodes. Assign to the players the minimal subset of beliefs from (1), possibly void, that allow them to calculate optimal choices given that choices at superseding nodes will be as implied by previous steps. If beliefs can be assigned in more ways than one, endow players with a disjunction of the associated beliefs. In the end, each player's minimal beliefs will consist of the conjunction of all beliefs attributed to him in the above steps.

If we apply this algorithm to the TOL(3) game we find the minimal beliefs as:

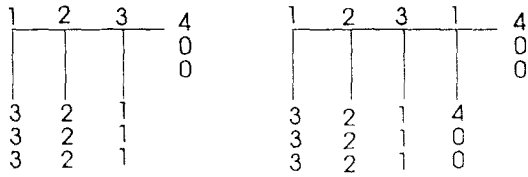
(2)

player 1 believes:	**player 2 believes:**
$R_2 \wedge B_2 R_1$	R_1

Player 2 believes R_1. She expects 1 to choose down at his second node so she rationally chooses down at hers. From $R_2 \wedge B_2 R_1$ player 1 can deduce this and, expecting 0 from choosing forward at his first node he will rather choose down.

Bicchieri purports to show that no inconsistency arises with minimal beliefs and that this case illustrates that "in order for the backward induction solution to obtain, the players must have some knowledge of the theory's assumptions, but no common knowledge of them" (Bicchieri 1989, p. 70). With only the minimal beliefs (2), compatible with backward induction, no inconsistency can be found if player 2's node is checked in isolation. Sticking to her endowment of beliefs, 2 can explain why 1 has deviated by assuming either that he believes her not to be rational ($\neg B_1 R_2$), or that 1 believes that she believes 1 not to be rational ($\neg B_1 B_2 R_1$). In contrast to the case of common knowledge discussed above, Bicchieri concludes that with "this distribution of knowledge ... a deviation is consistent with the players' beliefs" (Bicchieri 1989, p. 76).

But Bicchieri does not check for consistency at player 1's second node and her conclusion is precipitated. At this node 1 will, though he is in a position to gain the largest possible payoff, uphold inconsistent beliefs.[3] This node could only be reached if 2 chose forward and 1 can only explain this by assuming $\neg R_2 \vee \neg B_2 R_1$. But this is not allowed since by (2) he believes $R_2 \wedge B_2 R_1$. Thus there is no paradox of specifically *common* knowledge; any beliefs compatible with backward induction suffice to generate inconsistencies in the Bicchieri sense.



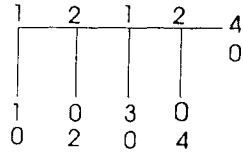
player 1 believes:	player 2 believes:	player 3 believes:
$R_2 \vee R_3$	$\varnothing$	$\varnothing$

Figure 3.

REMARK 1. In general inconsistencies arise when a player owns a node which his beliefs have led him to exclude as possible to reach. Common knowledge of rationality includes the minimal beliefs. This follows from the minimal beliefs algorithm. Hence, the class of games in which inconsistencies arise with minimal beliefs is a subset of the corresponding class with common knowledge. Furthermore, it is a strict subset. To see this consider the games in Figure 3 which also illustrate how the minimal beliefs (specified beneath) may involve a disjunction.

Minimally, the players 2 and 3 need believe nothing to have well-defined best choices, since these are dominant. Player 1 is capable of finding his optimal choice if he believes that either 2 or 3 is rational. The leftmost game in figure 3 is not inconsistent with minimal beliefs. No beliefs are contradicted as 1 never moves a second time, and the initial belief sets of 2 and 3 are null. Note that the game to the right is inconsistent, the only difference being the trivial bifurcation that assigns a decision node to 1 immediately following 2's and 3's deviations from the classical solution path. At this node player 1 would have to uphold the inconsistent conjunction of beliefs $(R_2 \vee R_3) \wedge \neg(R_2 \vee R_3)$. With common knowledge, however, both games are inconsistent in Bicchieri's sense. If player 1 starts by choosing forward then 2 cannot explain this since she must believe that 1 believes that both 2 and 3 are rational.

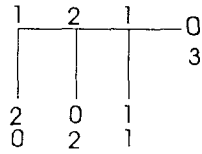
REMARK 2. The inconsistency at player 1's second node in the TOL(3) game shows that the requirement that every node be free of inconsistencies cannot be met even with minimal beliefs. On two grounds have we met with the objection that this inconsistency is in some sense "less important" than the inconsistency with common knowledge at player 2's node. First, since player 1 has a dominant choice available at his second node it may be argued that the inconsistency should be disregarded. We are sceptical



player 1 believes: $R_2 \wedge B_2 R_1 \wedge B_2 B_1 R_2$

player 2 believes: $R_1 \wedge B_1 R_2$

Figure 4.



player 1 believes: $\varnothing$

player 2 believes: R_1

Figure 5.

about this argument. Rational players ought to consult their beliefs when making decisions, even if these decisions only involve comparing payoffs, and from inconsistent beliefs anything can be derived as rational. However, granted that the objection is accepted it must be noted it is not generally true that inconsistencies arising with minimal beliefs obtain only where a dominant choice is available. In the TOL(4) game in figure 4, with minimal beliefs specified beneath, player 1 holds inconsistent beliefs at his second node where no dominant choice is available.

We give one more example with the same property. The game in Figure 5 moreover shows that it is *not* the case that inconsistencies with minimal beliefs arise only at nodes owned by a player who already has deviated from the equilibrium path. Beneath are the minimal beliefs.[4]

The backward induction solution is "down at every node". At 2's node, she must hold inconsistent beliefs. She must believe 1 is *not* rational, since he has just made a strictly dominated choice. On the other hand, the algorithm assigns player 2 the belief that 1 is rational.

Second, the inconsistency with common knowledge at player 2's node (in Figure 1) is such that player 1 holds enough beliefs to predict that

player 2 would face an inconsistency at her node (indeed it suffices that 1 believes $R_2 \wedge B_2R_1 \wedge B_2B_1R_2 \wedge B_2B_1B_2R_1$). However, one might conceivably argue that the inconsistency with minimal beliefs at player 1's second node is different in the sense that although player 1, while at his first node, might predict the inconsistency at his second node, he cannot predict that player 2 could predict it. This leaves him no strategic incentive to choose forward. This reasoning leads to the conclusion that the Nash equilibrium *outcome* obtains with minimal beliefs. Note though that the *profiles* entailed need not be subgame perfect, thus they need not have been generated by backward induction. The inconsistency at player 1's second node makes anything rational for him there, while sub-game perfection requires that he chooses down.

5. BELIEF REVISION

Bicchieri (1989) argues that inconsistencies in the theory of the game can make play outside Nash equilibrium reasonable. She is reluctant to accept the classical solutions in games such as the finitely repeated Prisoners' Dilemma (Luce and Raiffa 1957), the Chain-Store Paradox (Selten 1978), and the Centipede game (Rosenthal 1981).[5] She declares: "Once common knowledge of the theory of the game is assumed, it is no longer necessary to change the structure of the game, or to reject the traditional definition of rationality, to allow for different solutions. Indeed, the existence of common knowledge makes deviations from the classical solution plausible, and compatible with individual rationality" (Bicchieri 1989, p. 70). The sub-title of Bicchieri's article is "A Paradox of Common Knowledge". However, inconsistencies arise with much smaller belief sets than those involving common knowledge of rationality, so conceivably one could argue that deviations from the classical solution are plausible with virtually any informational assumption. We do not, however, believe that inconsistencies account for any particular type of play, but rather indicate that the theory needs revision. Beliefs that are contradicted cannot be sustained and a reasonable description of a game must consider this. To make the inconsistencies vanish, assume the players have the following beliefs and belief revision technology. The scheme is arbitrary, any scheme that allows for all inconsistencies to be discarded would do. This scheme has the property of reinstalling the backward induction solution. It is compatible with the players initially having common knowledge of each other's rationality which we think is a natural informational assumption to make.

(i) Each player initially holds a set of beliefs regarding his and other players' rationality that are compatible with backward induction. The players also hold beliefs concerning how players revise their beliefs, how players revise beliefs about how players revise beliefs,... and so on ad infinitum. The beliefs are all true.

(ii) If a player finds his beliefs contradicted he throws out those beliefs that have been contradicted only with specific reference to those previous nodes where he realizes they were not true. For the purpose of subsequent play he retains them.

To exemplify: if, in the TOL(3) game, player 1 is called upon to move at his second node he will jettison the belief that 2 believed him to be rational, or that 2 acted rationally, *at her previous node*, nothing more. This is feasible in the logical system previously described if the belief operators are indexed with reference to the nodes as well as the players. Generally then, B_i^n would mean "player i believes at node n". If this belief revision scheme is adopted the backward induction solution always obtains.

Basu (1990) points out that traditional solution concepts require players to "turn a blind eye to another player's irrationality". The proposed solution is related in kind. We make no claims regarding its plausibility. Our object is simply to device a scheme that works.

NOTES

* We wish to thank for valuable comments Werner Güth and the other participants at a game theory workshop in Riezlern, Austria, as well as Kaushik Basu, Jonas Björnerstedt, Giacomo Bonanno, Harold Kuhn, a referee, and in particular Jörgen Weibull who called our attention to this problem. We have also benefited from comments by participants at seminars at Chicago, Princeton, and Uppsala. Financial support from The Royal Swedish Academy of Sciences (MD), Handelsbankens forskningstiftelser and Svenska Institutet (JL) is gratefully acknowledged.

[1] We note that Bicchieri (1992) derives inconsistencies with limited beliefs. However, there, and in Bicchieri and Antonelli (1993), theories are specifically designed to avoid common knowledge and belief revision assumptions. This is in contrast to our main points: if there is a paradox it is one of lack of belief revision, and initial common knowledge of rationality is a natural assumption to make.

[2] Compare also Gärdenfors (1988, p. 20) who argues that "... the concepts of truth and falsity are *irrelevant* for the analysis of belief systems. ... From the subjects point of view there is no way to tell whether she *accepts* something as knowledge, that is, has full belief in it, or whether her accepted knowledge is also *true*."

[3] Elster (1991) too analyzes the TOL(3) game. He checks player 1's second node but does not detect the inconsistency.

[4] This point was suggested to us by Giacomo Bonanno who also constructed the game. He

discusses a similar game in Bonanno (1994).

[5] Numerous experiments have verified that human players usually do not play backward induction solutions in these games. One example is McKelvey and Palfrey (1992). In a centipede game they find evidence that selfish rationals deviated from the backward induction solution in the hope that they faced altruists. Some other explanations involving bounded rationality, uncertainty regarding game structure and co-players rationality or beliefs et cetera are found in Basu (1987), Binmore (1987), Kreps, Milgrom, Roberts, Wilson (1982), Neyman (1985), Rosenthal (1981). See also the discussions in Luce and Raiffa (1957, pp. 80–81, 97–102) and Selten (1978).

REFERENCES

Aumann, R.: 1995, 'Backward Induction and Common Knowledge of Rationality', *Games and Economic Behavior* **8**, 6–19.

Aumann, R. and A. Brandenburger: 1991, 'Epistemic Conditions for Nash Equilibrium,' Mimeo 91–042, Hebrew University and Harvard Business School.

Bacharach, M.: 1987, 'A Theory of Rational Decision in Games', *Erkenntnis* **27**, 17–55.

Basu, K.: 1987, 'Modeling Finitely-Repeated Games with Uncertain Termination', *Economics Letters* **23**, 147–151.

Basu, K.: 1990, 'On the Non-Existence of a Rationality Definition for Extensive Games', *International Journal of Game Theory* **19**, 33–44.

Basu, K.: 1994, 'The Travelers' Dilemma: Paradoxes of Rationality in Game Theory', *American Economic Review* **84**(2), 391–395.

Ben-Porath, E.: 1994, 'Rationality, Nash Equilibrium and Backwards Induction in Perfect Information Games', Mimeo, Northwestern University.

Bicchieri, C.: 1989, 'Self-Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge', *Erkenntnis* **30**, 69–85.

Bicchieri, C.: 1992, 'Knowledge-Dependent Games: Backward Induction', in Bicchieri and Dalla Chiara (Eds.), *Knowledge, Belief, and Strategic Interaction*, Cambridge Univ. Press.

Bicchieri C. and G. A. Antonelli: 1993, 'Game-Theoretic Axioms for Local Rationality and Bounded Knowledge', Paper presented at the Nobel Symposium on Game Theory at Björkborn Sweden, June 1993.

Binmore, K.: 1987, 'Modeling Rational Players: I', *Economics and Philosophy* **3**, 179–214.

Bimore, K.: 1995, 'Backward Induction and Rationality', Discussion Paper 95–10, University College London.

Bonanno, G.: 1991, 'The Logic of Rational Play in Games of Perfect Information', *Economics and Philosophy* **7**, 37–65.

Bonanno, G.: 1994, 'Review of Cristina Bicchieri's *Rationality and Coordination*', Forthcoming in *Economics and Philosophy*.

Elster, J.: 1991, *The Cement of Society*, Cambridge University Press.

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

Gul, F.: 1995, 'Rationality and Coherent Theories of Strategic Behavior', Forthcoming in *Journal of Economic Theory*.

Hintikka, J.: 1962, *Knowledge and Belief*, Cornell University Press.

Kreps, D., P. Milgrom, J. Roberts, and R. Wilson: 1982, 'Rational Cooperation in the Finitely-Repeated Prisoners' Dilemma', *Journal of Economic Theory* **27**, 245–252.

Kuhn, H.: 1953, 'Extensive Games and the Problem of Information', *Annals of Mathematics Studies* **28**, Princeton University Press.

Luce D. and H. Raiffa: 1957, *Games and Decisions*, Wiley, New York.

McKelvey R. and T. Palfrey: 1992, 'An Experimental Study of the Centipede Game', *Econometrica* **60**, 803–836.

Neyman, A.: 1985, 'Bounded Complexity Justifies Cooperation in the Finitely Repeated Prisoners' Dilemma', *Economics Letters* **19**, 227–29.

Pettit P. and R. Sugden: 1989, 'The Backward Induction Paradox', *The Journal of Philosophy* **4**, 169–182.

Reny, P.: 1988, 'Rationality, Common Knowledge and the Theory of Games', Ph.D. thesis, Department of Economics, Princeton University.

Reny, P.: 1993, 'Common Belief and the Theory of Games with Perfect Information', *Journal of Economic Theory* **59**, 257–274.

Rosenthal, R.: 1981, 'Games of Perfect Information, Predatory Pricing and the Chain-Store Paradox', *Journal of Economic Theory* **25**, 92–100.

Selten, R.: 1978, 'The Chain-Store Paradox', *Theory and Decision* **9**, 127–159.

Sugden, R.: 1991, 'Rational Choice: A Survey of Contributions from Economics and Philosophy', *The Economic Journal* **101**, 751–785.

Manuscript received December 12, 1994

Center for Economic Research
Tilburg University
P.O. Box 90153
5000 LE Tilburg
The Netherlands