Axiomathes (2017) 27:417–436
DOI 10.1007/s10516-016-9314-2

ORIGINAL PAPER

The Prisoner's Dilemma: From a Logical Point of View

Cheng-chih Tsai[1]

Received: 8 June 2016 / Accepted: 14 September 2016 / Published online: 22 September 2016

Abstract It is generally believed that, for a one-off Prisoner's Dilemma game, it is *logical* to defect. However, both players cooperating is apparently a better choice than both defecting, hence the dilemma. In this paper, by resorting to Ramsey's Test, Kripke's possible world semantics, and Stalnaker/Lewis-style account of conditionals, I show that the first horn of the Prisoner's Dilemma is an unsound argument. It originates from failing to differentiate between a possible world and a possible set of possible worlds and failing to observe that the set of accessible possible worlds associated with a possible world in general varies from conditional to conditional. This phenomenon can also be illustrated in terms of the recently developed hi-world semantics. Moreover, a meta-argument is constructed to establish the non-existence of a logical argument for defection.

Keywords The Prisoner's Dilemma · Ramsey's Test · Possible world semantics · Conditionals · Hi-world semantics · Expected utility

In Press and Dyson (2012), it was shown that in the two-player Iterated Prisoner's Dilemma (IPD), there exists a strategy for a player X to win over her evolutionary opponent Y who, without a theory of mind about X, can only accede to X's extortion. Nevertheless, Press and Dyson stressed that having a theory of mind can help Y resolve X's strategy and turn the game into an ultimatum game. Therefore, given that both players have a theory of mind about the other, there exists no definite value-maximizing strategy for IPD. In this paper, I shall argue that the same holds for the original Prisoner's Dilemma (PD): *logic* in itself does *not* suggest a value-maximizing strategy for PD—defection—as many have believed.

✉ Cheng-chih Tsai
cctsai@mmc.edu.tw; cctsai@ntu.edu.tw

[1] Center for Holistic Education, Mackay Medical College, 46, Sec. 3, Jhong-Jheng Rd, San-Jhih District, New Taipei City 252, Taiwan

Players of a PD game are no doubt in an embarrassing or 'dilemmatic' situation and decision theorists have definitely done a good job analyzing it. Yet, the modest goal of the present paper remains: to show that, even in a one-off PD game, *logic* does *not* tell us to defect.

1 The Prisoner's Dilemma

An interesting fact about the Prisoner's Dilemma is that while many researchers, such as Robert Trivers[1] and Robert Axelrod,[2] have drawn our attention to the far-reaching applicability of the IPD game and have explored all sorts of possible strategies for it, they seem to agree on one thing—the strategy for the original one-off PD is straightforward: Defect, period. For example, commenting on Press and Dyson (2012), Stewart and Plotkin (2012, p. 10134) has this to say,

> If the Prisoner's Dilemma is played only once, it always pays to defect — even though both players would benefit by both cooperating.

And in Nowak and Highfield (2012, p. 29), Nowak holds this view as well,

> In the single shot game, the one that I analyzed earlier in the discussion of the payoff matrix of the Prisoner's Dilemma, it was *logical* to defect. (italics mine)

I think both of them are wrong in thinking that it is logical to defect in the PD game, and I will illustrate that it is indeed not easy to formalize the Prisoner's Dilemma as a (logical) dilemma.

According to *Shorter Oxford English Dictionary* (2007), a dilemma is

1. In RHETORIC, a form of argument involving an opponent in choice between two (or more) alternatives, both equally unfavourable. In LOGIC, a syllogism with two conditional major premisses and a disjunctive minor premiss.

On the face of it, PD fits both of these criteria, being of the form, "I cooperate or defect; if I cooperate then my deed is inconsistent with the fact that defection is always a better choice; if I defect then my deed is inconsistent with the fact that both players cooperating is a better choice; therefore a contradiction is inevitable." But in fact it does not. Clearly, the 'argument' and 'syllogism' in the quoted entry are meant to be *sound* arguments, but, as we will see soon, the argument associated with the first horn of PD, i.e. the argument for defection, is not sound at all.

Let a_1 and a_2 be two persons involved in a PD situation, and let (R, S, T, P) be the reward for a_1 when the collective action of a_1 and a_2 lies in $(C_1C_2, C_1D_2, D_1C_2, D_1D_2)$, where C_i and D_i denote the cooperation and the defection of a_i respectively, with the assumption that $T > R > P > S$ and $2R > T + S$. Presumably, the following two horns are involved in PD:

[1] See Trivers (1971).

[2] See Axelrod (1984).

Horn 1. C_2 or D_2; If C_2 then $N(D_1)$; If D_2 then $N(D_1)$/$\therefore N(D_1)$; similarly, an argument for $N(D_2)$.

Horn 2. Yet, the values associated with C_1C_2 are higher than that associated with D_1D_2.

Here $N(D_1)$ stands for "the rational being a_1 should defect".

Now, rationality is indeed a highly praised value in itself, but to establish Nowak's claim that it was *logical* to defect, we need to formulate it in logical terms. What can "the rational being a_1 should defect" mean? The best that I can think of is that a rational agent is one who will do whatever logic suggests her to do, and $N(D_1)$ then simply means that D_1 is a logical consequence of the antecedent provided that agent a_1 seeks to maximize her value and is aware of the antecedent.[3] So, in ideal cases such as the Prisoner's Dilemma in question, we can, for simplicity, assume that the agent a_1 is rational, and reformulate the first horn of Prisoner's Dilemma as follows, with no normative mode in sight.

Horn 1 We have a sound argument for the conclusion D_1:

	P0	C_2 or D_2
	P1	If C_2 then D_1
	P2	If D_2 then D_1
[Main]	- - - - - - - - - - - - - - - - - -	
	$\therefore$	D_1

The argument [Main] of *Horn 1* is, on the face of it, an instance of Disjunction Elimination[4] and the premisses P0 P2 seem all true as well. So the Prisoner's Dilemma is *prima facie* a dilemma—it advices a_1 to defect (by *Horn 1*) and not to defect (by *Horn 2*) at the same time.

However, if we look at [Main] more closely, we will see that the premisses P1 and P2 are actually supported by the following hidden arguments:

	B1	If C_2 then $v(D_1) > v(C_1)$
	G	If $v(D_1) > v(C_1)$ then D_1
[Supportive 1]	- - - - - - - - - - - - - - - - - -	
	(P1)	If C_2 then D_1

and

[3] We shall have more to say about this assumption of awareness later.

[4] Note that if P0, P1 and P2 are understood as 'a_1 knows that C_2 or D_2', 'if a_1 knows that C_2 then she would defect', and 'if a_1 knows that D_2 then she would defect' respectively, then [Main] is no longer an instance of Disjunction Elimination.

B2 If D_2 then $v(D_1) > v(C_1)$
G If $v(D_1) > v(C_1)$ then D_1
[Supportive 2] - - - - - - - - - - - - - - - - - -
(P2) If D_2 then D_1

Here $v(X_1)$ stands for the reward that a_1 receives when she adopts the strategy X.[5]

How do we interpret the sentences in these arguments, especially the conditionals B1, B2 and G? The soundness of these arguments no doubt depends on how we interpret these conditionals. However, before we spell them out in detail in Sect. 3, let us briefly review, in the next section, the Kripkean possible world semantics and Ramsey's Test that shall play an essential role in our analysis of the arguments.

2 The Possible World Semantics and Ramsey's Test

Despite its great explanatory power, the Kripkean possible world semantics is often criticized for the fact that the so called 'possible worlds' are unrealistic and difficult to pin down or grasp. However, for a description of the PD game—the subject of the present paper—the possible world semantics turns out to be a perfect tool, and all possible worlds can be explicitly described. There are altogether four possible worlds, C_1C_2, C_1D_2, D_1C_2, and D_1D_2 respectively, and the binary accessibility relation R on the universal set W of possible worlds is such that every world is accessible to each other (including itself). In other words, we are now in position to analyze the PD based on an S5 setting. The fact that S5 is by far the simplest modal system will help us spot a problem more easily, should there really be one.

Another useful tool that we shall be referring to is Ramsey's Test. Ramsey famously says the following about conditionals,

> If two people are arguing 'If p will q?' and are both in doubt as to p, they are adding p hypothetically to their stock of knowledge and arguing on that basis about q; so that in a sense 'If p, q' and 'If p, -q' are contradictories.
> — Ramsey (1990), p155, footnote 1

Here are two important observations relevant to the task of this paper.

First, in terms of the possible world semantics, when someone is assessing "if p, q" at world w, the adding of p (hypothetically) to her stock of knowledge amounts to a modification—more specifically, a shrinking—of the set of possible worlds accessible to w, so that it includes only those possible worlds in which p holds. This way of assessing a conditional has two consequences, (1) in evaluating a conditional, what we actually resort to is a set of possible worlds rather than a

[5] There are some complications concerning the possible referents of $v(C_1)$ and $v(D_1)$, and we shall say more about them later.

possible world; (2) the set of possible worlds in question can vary from conditional to conditional.

Second, when a person is asserting a conditional "if p, q" for which the consequent q itself involves the self-reflexive indexical "I", things can become complicated.[6] She has two options—after adding p hypothetically to her stock of knowledge, she either supposes that the "I" in q has p in her stock of knowledge or she does not. These two options will be characterized as 'subjective' and 'objective' interpretations of such a conditional respectively.[7] The following pair of conditionals

S If there is a bomb in this room, I will leave the room in no time,
O If there is a bomb in this room, I will be blown into pieces,

can be true according to different interpretations. A subjective interpretation would make sentence *S* true, while an objective interpretation renders sentence *O* true. As a matter of fact, the subjective interpretation seems defeasible, because once challenged by "but, you *might not* know that there was a bomb!" one would normally withdraw his former assertion. Nonetheless, it is still an interpretation that we often adopt in our daily conversation.

More importantly, this phenomenon may happen to cases not involving self-reflexive indexical as well. For example, concerning P1, we may ask ourselves: After C_2 is added to our stock of knowledge, does the person a_1 knows about C_2 too, so as to make the decision to defect? For a subjective interpretation, the answer is "yes", but for an objective interpretation, the answer would be a "no".[8]

3 Is the Argument for Defection Sound?

Recall that the first horn of PD consists of three arguments [Main], [Supportive 1], and [Supportive 2]. Now, whether these arguments are sound hinges on how we interpret the conditionals.

The naive reading

To begin with, if we simply read the conditional "if p, q" as the material implication, and every sentence is supposed to be evaluated against a possible world (or, more properly, a truth assignment), then the three arguments become:

[6] See Chalmers and Hájek (2007) for a nice illustration of this problem.

[7] See Tsai (2016) for more detailed discussion of this distinction. Note that the terms used there are 'autistic' and 'realistic' instead.

[8] After all, a_1 by default has no access to the truth of C_2. She can at best reason that if C_2 then 'I *should* defect'. However, 'I *should* defect' is different from 'I *would* defect', as the truth of the former is independent of whether a_1 knows C_2 or not, while the latter is.

	P0	$C_2 \vee D_2$
	P1	$C_2 \supset D_1$
	P2	$D_2 \supset D_1$
[Main]	- - - - - - - - - - - - - - -	
	$\therefore$	D_1

	[Supportive 1]		[Supportive 2]
B1	$C_2 \supset v(D_1) > v(C_1)$	B2	$D_2 \supset v(D_1) > v(C_1)$
G	$v(D_1) > v(C_1) \supset D_1$	G	$v(D_1) > v(C_1) \supset D_1$
$\therefore$ (P1)	$C_2 \supset D_1$	$\therefore$ (P1)	$D_2 \supset D_1$

Apparently, all three arguments above are valid. However, if we have been more careful, we would have found that the formula $v(D_1) > v(C_1)$, which plays a key role in the two supportive arguments, simply makes no sense at a possible world w. Given any w, only one of $v(D_1)$ and $v(C_1)$ is applicable, because in w the agent a_1 either defects or cooperates, but not both. Therefore, the premisses P1 and P2 of [Main] are not supported by meaningful, let alone sound, arguments, hence we find no reason to accept the conclusion D_1 of the [Main] argument. Affixing the operator $\Box$ to all the material conditionals, that is, turning all material implications into strict implications, would not help either, because the mismatch of C_2 and $v(D_1) > v(C_1)$ —as statements concerning different entities—remains.

Some may disagree with my claim that $C_2 \supset v(D_1) > v(C_1)$ does not make sense at a possible world w, because, apparently, given that C_2 holds at w, in evaluating $v(D_1) > v(C_1)$, we can simply resort to the set R_w of all possible worlds accessible from w, and $v(D_1) > v(C_1)$ is true on R_w if the $v(D_1)$ of any world in R_w is greater than the $v(C_1)$ of any other world in R_w whenever $v(X)$ makes sense. However, their objection would not work. Recall that here we are interpreting the conditional B1 as a *material conditional*. So, in getting the R_w for the evaluation of $v(D_1) > v(C_1)$, we have no way to impose the constraint that the possible worlds in R_w that we are interested in are only those at which the antecedent holds. Therefore, as we have seen in the previous section, the R_w will be consisting of all four possible worlds C_1C_2, C_1D_2, D_1C_2, and D_1D_2, and clearly $v(D_1) > v(C_1)$ would not hold, even making no sense, on R_w.

In sum, the naïve trial—interpreting the conditional as either material conditional or strict conditional—fails in the following way, and we cannot reach the conclusion of defection.

	Validity	Truth of all premisses
[Main]	○	×
[Supportive 1]	○	×
[Supportive 2]	○	×

3.1 The Ramsey−KSL Reading

To grasp our intuition that, in evaluating B1, we only consider $v(D_1) > v(C_1)$ for the worlds at which C_2 holds, we have to incorporate Ramsey's idea into the Kripkean possible world semantics, and consider a Kripke–Stalnaker–Lewis (in short, KSL)-style reading of the conditional. As a matter of fact, the Ramsey–KSL way of understanding the conditional B1 is quite natural and we make such statements all the time. For example, the Gibbard–Harper style Causal Decision Theory can reckon "if C_2 then $v(D_1) > v(C_1)$" meaningful on the ground that, given that the other player cooperates, my utility in the nearest world where I defect and he co-operates would be greater than my utility in the nearest world where we both co-operate. Alternatively, we can say that the sentence B1 is true at w if we search the set of all those worlds accessible from w in which a_2 cooperates, and compare the $v(D_1)$ of any world with the $v(C_1)$ of any other world in it and find that the former is always higher than the latter. Either way, the antecedent sets a condition on the set of accessible possible worlds associated with a possible world, in contrast to setting a condition on the possible world itself. We can even say that the antecedent of B1 is essentially concerned with a possible set of (accessible) possible worlds rather than with a possible world. It is analogous to the following: "John's friends all know each other" is a condition on the set of friends of John, while "John is tall" is a condition on John himself.[9]

Specifically, in the case of PD, our preferred truth condition for B1 is that B1 is true at a possible set U of possible worlds if and only if so long as U is a subset of the extension $\|C_2\|$ then the $v(D_1)$ of D_1C_2 is higher than the $v(C_1)$ of C_1C_2 provided both values obtain. According to this reading, both B1 and B2 are surely true.

Now how about the premiss G, "if $v(D_1) > v(C_1)$ then D_1"? According to Ramsey's Test, we should add $v(D_1) > v(C_1)$ into our stock of knowledge and see whether we would accept D_1. The adding of the antecedent into our stock of knowledge amounts to restricting our consideration to the nearest world whose set of accessible worlds is such that $v(D_1) > v(C_1)$ holds. But, there are a couple of problems here.

First, in the Kripkean scheme, the set of accessible possible worlds for each possible world is the same, namely W. So, $v(D_1) > v(C_1)$ does not hold at any possible world. We can either regard such a counter-possible conditional as automatically true, or regard it as meaningless. However, as we would certainly be reluctant to also regard "if $v(D_1) > v(C_1)$ then C_1" as true, the second option seems more preferable. To make sense of the conditional G, we simply cannot stick to the default set of accessible worlds determined by the accessibility relation of Kripke's semantics. The natural choice would be, as we did for B1, to regard $v(D_1) > v(C_1)$

[9] Some might want to insist that the former is a contingent condition on John himself still. However, if so, then we should at least allow the set of John's friends to be a contingent set, just as John's height is a contingent fact. Analogously, for the B1 case, to regard $v(D_1) > v(C_1)$—which is clearly a condition on a set of possible worlds rather than a condition on a possible world—as a condition on w itself, we would have to allow for the possibility that w has access to a different set of accessible worlds. This, nevertheless, cannot be done in the usual Kripkean framework. In the Kripkean scheme, once a model is given, the set of accessible worlds is fixed.

as imposing a restriction on the set of possible worlds accessible from a possible world, so that G is to be evaluated against possible sets of accessible possible worlds rather than against possible worlds. Then, G is true with respect to a set of possible sets of accessible possible worlds if for every possible set U of possible worlds on which $v(D_1) > v(C_1)$ holds, C_1 holds at the world from which the possible worlds in U are accessible. This is intuitively plausible, but we should bear in mind that, strictly speaking, the usual Stalnaker/Lewis style account of conditionals does not accommodate talks of this sort. A possible world paired with a restricted/modified set of accessible possible worlds seems to be what we should have at hand in order to check whether "not $v(D_1) > v(C_1)$ or C_1" holds, and G is true with respect to a set of such pairs if and only if "not $v(D_1) > v(C_1)$ or C_1" holds for every pair. In other words, G asserts that for every possible world whose *modified* set of accessible possible worlds is such that $v(D_1) > v(C_1)$ holds, C_1 holds. Again, as in the case of B1, while the traditional Stalnaker/Lewis-style conditional restricts our attention to possible worlds for which the antecedent holds, G restricts our attention to those possible sets of accessible possible worlds for which the antecedent holds.

Granted that we can charitably interpret the conditional G so that our attention is restricted to pairs (U, w)—where w is a possible world and U is a possible set of possible worlds accessible from w—such that $v(D_1) > v(C_1)$ holds at U, would it always be the case that C_1 holds at w then? This leads to a second concern which involves the subtle distinction that we mentioned near the end of Sect. 2. We need to distinguish between an objective interpretation of G and a subjective interpretation of G. And whether a_1 has a privileged access to our knowledge of $v(D_1) > v(C_1)$ or not will make all the difference.[10] The objective minded do not assume that a_1 has the antecedent, namely $v(D_1) > v(C_1)$, in her stock of knowledge, yet the subjective minded do. We are thus divided between the following two interpretations of G.

3.1.1 Objectively Conceived a_1

Even if $v(D_1) > v(C_1)$ holds for some set U of possible worlds, there is no guarantee that a_1 knows this fact and would consequently defect. The fact that a_1 is assumed to be a completely rational being alone does not help because a_1's decision needs to be grounded on the knowledge of whether $v(D_1) > v(C_1)$ holds for U, but even if we assume that a_1 always knows whether $v(D_1) > v(C_1)$ holds for the set V of possible worlds that she has in mind, a_1's action is independent of the state of U—after all, U and V are distinct sets—unless the state of U is a sort of public knowledge/regulation that everyone, in particular a_1, is aware of. Therefore, G is in general false. As a result, while both [Supportive 1] and [Supportive 2] are valid arguments, they are unsound because they both contain a false premiss, G. In the same vein, the argument [Main] is unsound as it contains two unwarranted premisses P1 and P2. The problem of [Main] is worse than that, because it in itself is not a valid argument to start with, and we will talk about that later.

[10] Recall that D_1 stands for 'a_1 defects', and the truth of it certainly depends on who this a_1 is, in particular, whether this a_1 would have $v(D_1) > v(C_1)$ in her stock of knowledge and whether that knowledge will prompt her to defect.

In sum, according to this interpretation, we have

	Validity	Truth of all premisses
[Main]	×	×
[Supportive 1]	○	×
[Supportive 2]	○	×

3.1.2 Subjectively Conceived a_1

If we grant G the subjective interpretation, so that a_1 has an unrealistic, privileged access to the fact that $v(\mathrm{D}_1) > v(\mathrm{C}_1)$,[11] and the set V mentioned earlier becomes the same as U, then G can indeed be accepted to be true. In this case, while P0 is a truism by stipulation, P1 and P2 are both supported by sound arguments—both [Supportive 1] and [Supportive 2] are valid arguments and all the premisses B1, B2 and G are true. Therefore, the premisses of [Main] are all true now indeed.

Nevertheless, the argument [Main] is still unsound as it itself is not a valid argument in the first place. Clearly, $R_w \subseteq |\mathrm{C}_2 \vee \mathrm{D}_2|$, $R_w \cap |\mathrm{C}_2| \subseteq |\mathrm{D}_1|$, and $R_w \cap |\mathrm{D}_2| \subseteq |\mathrm{D}_1|$ together do not lead us to $R_w \subseteq |\mathrm{D}_1|$, as the property of a set is not exhausted by the properties of its constituent subsets. So, according to this interpretation, we have

	Validity	Truth of all premisses
[Main]	×	○
[Supportive 1]	○	○
[Supportive 2]	○	○

Therefore, a_1 still has not got any *logical* reason to defect.

By looking closely at [Main], an argument that has allegedly shown that defection is the logical choice for the PD, we have found that the argument is not sound after all. Apparently, there is no simple way that *Horn* 1 can be conceived as a sound argument in propositional modal logic, and it is reasonable to suspect that the PD is not a logical dilemma at all, unless someone can put *Horn 1* in another way and prove that it is indeed sound.

3.2 The Hi-World Reading

In the Ramsey–KSL reading, even if we can charitably interpret the argument [Main] and the supportive arguments in the spirit of Stalnaker/Lewis conditional and discover that [Main] is indeed not a sound argument, some of the conditionals involved in the arguments still could not be expressed in terms of traditional modal logical terms. This is due to the fact that the conditionals involve a mixture of modality of different levels, yet the Kripkean semantics simply does not provide us with a tool to analyze them, for instance, it does not allow us to consider a possible world and a possible set of possible worlds at the same time. Incidentally, the

[11] For instance, imagine that the a_1 is one of us who are pondering the three arguments in question.

recently developed hi-world semantics, which can in effect take care of any mixture of iterated modalities at one go, turns out to be able to provide us with a new way of interpreting modal formulas so that the arguments discussed in the preceding subsection can be properly formulated in terms of usual modal formulas, and receive an interpretation that capture our intuition concerning the Ramsey–KSL reading. In this subsection, we will formulate the arguments in question in terms of hi-world semantics and see from another angle why the argument [Main] is unsound.

In Becker (1952), the German logician O. Becker proposes that we should be clear about whether a sentence is to be evaluated at a case (a possible world) or a case class (a set of possible worlds, or an iterated set of possible worlds). In particular, a conditional can be concerned with a possible world w or a set U^1 of possible worlds. A material implication $\alpha \supset \beta$ concerns a possible world, while a strict implication $\alpha \rightarrow \beta \equiv \Box(\alpha \supset \beta)$ concerns a set of possible worlds. Failing to make such a distinction can lead us to wrongly accept the validity of "Obama is not here./∴ If Obama is here, he will buy everybody a drink." Clearly, the premiss here is concerned with a world, the actual world, while the conclusion is concerned with a set of possible worlds.

However, things can be more complicated than that. As we may encounter sentences such as $\Box(\Box\alpha \supset \beta)$, Becker's idea of separating w and U^1 proves to be too naïve. There indeed can be all sorts of other possibilities. One needs a more comprehensive semantic scheme for the task of analyzing it, and the hi-world semantics introduced in Tsai (2012) serves this purpose perfectly. The reader is referred to the "Appendix" for an outline of the semantics. Basically, a hi-world s takes the form $(U^0, U^1, U^2, \ldots)$ where U^0 is simply a possible world w_0, U^1 is a set of possible worlds and U^2 is a set of sets of possible worlds. A hi-world t is a sub-hi-world of s provided that $\pi_i(t) \in \pi_{i+1}(s)$ for all i, where π_i is the projection into the ith component. Every sentence is concerned with some suitable portion(s) of a hi-world. For example, the sentence $\Box\alpha \supset \beta$ is true at a hi-world $s = (w_0, U^1, U^2, \ldots)$ provided that $U^1 \cap I(\alpha)^c$ is nonempty or $w_0 \in I(\beta)$, and the truth of $\Box(\Box\alpha \supset \beta)$ at s amounts to that for every sub-hi-world t of s, $\Box\alpha \supset \beta$ holds for t.[12] Furthermore, so far as the present paper is concerned, we can impose a mild condition on our universal set of hi-worlds: every hi-world is its own sub-hi-world. This is equivalent to the self-reflexivity of the accessibility relation of a Kripkean model, and it allows us to obtain α from $\Box\alpha$.

Now let us see how we can formalize Ramsey's conditional in terms of the hi-world semantics. According to Ramsey, to accept a conditional "if α then β" is to add α into our stock of knowledge and based on that arrive at β. In terms of hi-worlds, one's stock of knowledge amounts to a set of subsets of hi-worlds, and the intersection of these subsets is the set of hi-worlds that she has in mind. The adding of α into her stock of knowledge amounts to shrinking the set of hi-worlds she has in mind by taking its intersection with the extension $\|\alpha\|_M$ of α.[13] This phenomenon can

[12] To see what the extension $I(p)$ means and to appreciate the subtle difference between $I(p)$ and $\|p\|_M$ please refer to the "Appendix".

[13] See the "Appendix" for the definition.

be illustrated in terms of the U^i's. Consider a conditional "if p then q" where both p and q are non-modal sentences. To add p into our stock of knowledge and then consider q amounts to shrinking U^1 to $U^1 \cap I(p)$ and to see whether the resulting $U^1 \cap I(p)$ lies in $I(q)$. So, "if p then q" can be translated into $\Box(p \supset q)$, as the truth condition for the latter is $U^1 \subseteq I(p)^c \cup I(q)$, which is equivalent to $U^1 \cap I(p) \subseteq I(q)$.[14] The situation for the Prisoner's Dilemma is more complicated, as we will see soon, but the basic ideas are the same.

To find a more probable interpretation of the argument(s) in question, we need to fix quite a number of problems. Let us begin with B1 and B2. Without loss of generality, I shall be concerned with B1 only here. As we have seen, the original B1, namely "if C_2 then $v(D_1) > v(C_1)$", seems outright true, yet if we translate it as "$C_2 \supset v(D_1) > v(C_1)$", then the fact that the antecedent and consequent concern different levels of worlds, namely w_0 and U^1, will immediately turn it into a false statement. What is wrong with the translation, and how can we fix it then? The following remarks will guide us to a right translation of B1.

First, as we have explained earlier, when we say "if C_2 then $v(D_1) > v(C_1)$", we are not asserting a specific connection between w_0 and U^1. Rather, we are claiming that given that a set U^1 is such that the agent a_2 cooperates, $v(D_1) > v(C_1)$ holds for that U^1. In other words, both the antecedent and the consequent of B1 are concerning the same level of a hi-world, namely the U^1. As a consequence, we should arrive at some modified translation of B1 which contains $\Box C_2 \supset v(D_1) > v(C_1)$ as a proper part. In everyday language, B1 can be put in the following way: "if a set of possible worlds is such that agent a_2 always cooperates then $v(D_1) > v(C_1)$ holds for that set".

Second, the remark in the last section concerning the formalization of a Ramsey conditional applies to B1 as well. In other words, in evaluating B1, we first add $\Box C_2$ into our stock of knowledge, which amounts to taking the intersection of the set $\hat{s}$ of all sub-hi-worlds of s with $\|\Box C_2\|_M$, and then decide whether the resulting set is a subset of $\|v(D_1) > v(C_1)\|_M$. So, the correct translation of B1 should be $\Box(\Box C_2 \supset v(D_1) > v(C_1))$ instead.[15]

Strictly speaking, when we are unsure of whether C_2 or D_2 holds necessarily, the expression $v(D_1) > v(C_1)$ makes no sense at all, because each of $v(D_1)$ and $v(C_1)$ may have two distinct values. However, charitably speaking, by $v(D_1) > v(C_1)$ we could mean that for any world w of U^1 for which $v(D_1)$ is applicable, and for any world w' of U^1 for which $v(C_1)$ is applicable, the value $v(D_1)$ is greater than the value $v(C_1)$. The present remark is important in the sense that without such an interpretation, the antecedent of G would be meaningless for most U^1's.

[14] There is a further complication concerning whether we should impose the Existential Import for our conditionals, that is, whether we should impose $\Diamond p$ into a Ramsey conditional and translate 'if p then q' into $\Diamond p \wedge \Box(p \supset q)$ instead. But, I will ignore the problem here and refer interested readers to Tsai (2016) for more details.

[15] A worth-mentioning fact is that in hi-world semantics, the most unrestrictive mode is that every hi-world takes the form $(w_0, D, P(D), P^2(D), \ldots)$. In this case, every hi-world is a sub-hi-world of each other, and then $\Box(p \supset q)$ and $\Box(\Box p \supset \Box q)$ can be shown to entail each other. In comparison, in the Kripkean semantics, these two formulas are not logically equivalent even in S_5.

Now, concerning G, we should note the following. First, in contrast to B1 and B2, for which the antecedents C_2 and D_2 are elevated into $\Box C_2$ and $\Box D_2$ in hi-world semantics so that the antecedents and consequents are about the same level of worlds, the consequent D_1 of G is really about the plain world w_0 and the conditional presumes that agent a_1 is rational.

Second, according to Ramsey's Test, to decide whether to accept G, we need to add the antecedent $v(D_1) > v(C_1)$ into our stock of knowledge and, based on that, decide whether a_1 defects always. Recall that $v(D_1) > v(C_1)$ is true with respect to U^1 provided that the value for a_1 in those possible worlds of U^1 in which she defects is always higher than that associated with those possible worlds of U^1 in which she cooperates. However, adding the antecedent $v(D_1) > v(C_1)$ into our stock of knowledge amounts to shrinking U^2 so that for all the elements V^1's of U^2, $v(D_1) > v(C_1)$ holds, and to see if a_1 defects always amounts to seeing if for all elements w's of U^1, D_1 holds at w. So, in terms of hi-world semantics, G can be translated into $\Box(v(D_1) > v(C_1) \supset D_1)$ and it is true at s if and only if $\hat{s} \subset \|v(D_1) > v(C_1) \supset D_1\|_M$, where again, $\hat{s}$ stands for the set of all sub-hi-worlds of s.

Finally, given that B1, B2 and G are translated as above, P1 and P2 can, unsurprisingly, be pinned down as $\Box(\Box C_2 \supset D_1)$ and $\Box(\Box D_2 \supset D_1)$ respectively, and P0 surely is primarily concerned with U^1 rather than w_0, and thus should be translated into $\Box(C_2 \vee D_2)$. So the three arguments that we are concerned with are expressed as follows. Again, $\alpha \rightarrow \beta$ stands for the strict conditional $\Box(\alpha \supset \beta)$.

	P0	$\Box(C_2 \vee D_2)$
	P1	$\Box C_2 \rightarrow D_1$
	P2	$\Box D_2 \rightarrow D_1$
[Main]	- - - - - - - - - - -	
	$\therefore$	D_1

	[Supportive 1]		[Supportive 2]
B1	$\Box C_2 \rightarrow v(D_1) > v(C_1)$	B2	$\Box D_2 \rightarrow v(D_1) > v(C_1)$
G	$v(D_1) > v(C_1) \rightarrow D_1$	G	$v(D_1) > v(C_1) \rightarrow D_1$
$\therefore$ (P1)	$\Box C_2 \rightarrow D_1$	$\therefore$ (P2)	$\Box D_2 \rightarrow D_1$

In terms of hi-world semantics, an argument is valid provided that for all possible hi-worlds $s = (w_0, U^1, U^2, \ldots)$, if the premisses are all true at s then the conclusion is true at s, and the argument is sound provided that the premisses are all true with respect to the actual hi-world as well. Undoubtedly, [Supportive 1] and [Supportive 2] are both valid arguments. However, granted that a_1 is a rational being, do we want to accept that G is true? It depends on how we read into a_1. If she is rational but not omniscient then even if $v(D_1) > v(C_1)$ holds for U^1 she may not know it, so G is not true, so [Supportive 1] and [Supportive 2] are unsound. If, on the other hand, we grant a_1 the mental power of knowing that $v(D_1) > v(C_1)$ holds for U^1 whenever it holds, then G is true and both of the supportive argument are sound and,

as a consequence, the argument [Main] have three true premisses. Do we then finally arrive at a sound argument in support of the first horn of PD? By no means. Clearly, $\Box(C_2 \vee D_2)$, $\Box C_2 \rightarrow D_1$, and $\Box D_2 \rightarrow D_1$ cannot lead us to D_1—given that a hi-world $s = (w_0, U^1, U^2, \ldots)$ is such that $U^1 \subseteq I(C_2 \vee D_2)$, "for any $w' \in U^1$ and $V^1 \in U^2$, $V^1 \not\subset I(C_2)$ or $w' \in I(D_1)$", and "for any $w' \in U^1$ and $V^1 \in U^2$, $V^1 \not\subset I(D_2)$ or $w' \in I(D_1)$" all hold, we still *cannot* conclude that $w_0 \in I(D_1)$, even if we presuppose that every hi-world is its own sub-hi-world, in particular, $w_0 \in U^1$ and $U^1 \in U^2$.

In sum, the arguments that seem to support the first horn of PD involve a mixture of modality of different levels. These arguments can either be dealt with in terms of (1) the usual KSL-style account of conditionals, or be reformulated in terms of (2) the hi-world semantics, which can in effect take care of any mixture of iterated modalities at one go. Furthermore, the rational agent a_1 can be assumed to either (i) *have* or (ii) *does not have* a privileged access to the truth of the antecedent. However, for each of the four possible interpretations, 1(i), 1(ii), 2(i), and 2(ii), we find that the argument [Main] is never a sound argument. So the first horn of PD remains in want of a sound argument that would support it.[16]

4 A Meta-argument for the Non-existence of a Logical Argument for Defection

A reviewer for an earlier version of this paper has objected that I have not exhausted all possible formulations of the arguments in question, in particular, I have not considered formulating PD in terms of quantified modal logic, so I have not ruled out the possibility that there is indeed a more complicated argument which proves that defection is the logical consequence of a PD game. A short answer to this objection would be that it is the responsibility of those who claim that it is logical to defect in a PD game to come up with an explicit sound argument for their claim. However, it would be much better if I can find a meta-argument that shows that, in general, no such sound argument exists, and this is what I would attempt to do in this section.

An anonymous reviewer for this journal reminds me that the notion of a game involves not only players and available strategies, but also a complex system of knowledge, preferences and beliefs, and without capturing the interaction between the players, my treatment of the PD game would not be complete. Indeed, to model the interaction between players in a game is very important. However, given that in this paper what we are interested is the one-off PD game rather than the iterated PD game, the "interaction" between the two players can at best amount to envisaging the opponent's thoughts and trying to outwit them. Yet, as your opponent can be any kind of players (rational, emotional, religious, criminal, your twin etc.), without knowing in advance who you are playing with, it is indeed difficult, if not impossible, to formulate a uniform argument for defection that captures what is

[16] The author would like to thank an anonymous reviewer for this journal for pointing out that the analysis concerning conditionals in this paper could be developed into other domains as well, in particular, its relation to constructivism could be explored. However, the treatment of this more general subject will have to await another paper.

going on in one's thoughts about the opponent's thoughts in a PD game. Luckily, the modest goal of this paper is merely to show that logic alone does not tell us to defect in a PD game, and it turns out that a *reductio ad absurdum* meta-argument concerning two ideally rational players suffices to achieve this goal. Moreover, in the process of presenting the meta-argument, we will have, in effect, taken into account the "interaction" between the two players—represented as a repeated reflection on each player's logical actions.

To be more specific, in this section, instead of trying to imagine what the underlying argument(s) are when one claims that it is logical to defect (LTD) and spill much ink on it, I will, for the sake of argument, simply assume that there indeed exists, as many authors have believed, a sound argument that allows them to get to the LTD conclusion. And if there exists such an argument for someone to always defect, it would work for the special case where the opponent is a perfectly rational being as well. Then I show that this will lead to the paradoxical result that it is logical for the two rational players in a PD game to cooperate as well. By *reductio ad absurdum*, we then have proved that *there cannot be* such an argument for defection, contrary to what the other authors have believed.

Let me assume, without explicitly spelling out the argument, that we have an argument [*] in support of D_1,

$$\begin{array}{ll} & \mathscr{K} \\ [*] & \text{- - - - - - - - - -} \\ & \therefore D_1 \end{array}$$

while the $\mathscr{K}$ here denotes the set consisting of all the premisses known to both players. These premisses are either rules of the PD game or are logically deducible from these rules and/or other known facts. For example, we can imagine that among the premisses in $\mathscr{K}$ there is one that says that both agents are perfectly rational beings who abide by logical rules.

Next, I would show that [*] would lead to a "paradoxical" result, namely that agent a_1 would cooperate as well.

By stipulation, agent a_2 is a rational being and the public information contained in $\mathscr{K}$ is available to agent a_2 as well, so we would have the following sound argument [*′] too.

$$\begin{array}{ll} & \mathscr{K} \\ [*'] & \text{- - - - - - - - - - - -} \\ & \therefore D_2 \end{array}$$

In general, if there exists a pure logical argument for an agent in a PD game to take a particular action *X*, then by symmetry, the other agent would be forced by logic to take the same action.

Now, as a consequence of the soundness of [*] and [*′], we have the truth of $D_1 \wedge D_2$. But then consider the following argument,

P1 $D_1 \wedge D_2$
P2 If $D_1 \wedge D_2$ then $((C_1 \wedge C_2) \vee (D_1 \wedge D_2))$
P3 If $((C_1 \wedge C_2) \vee (D_1 \wedge D_2))$ then $v(C_1) > v(D_1)$
P4 If $v(C_1) > v(D_1)$ then C_1
[Paradox, *Naïve*] - - - - - - - - - - - - - - - - - - -
$\therefore C_1$

P1 is true by the hypothetical soundness of [*] (and [*′]). P2 is true for most, if not all, interpretations of conditionals. P3 is true by the specification of the game. Finally, P4 seems true by the fact that agent a_1 is a rational being and that he would maximize his value whenever possible. An analogous argument [Paradox, *Naïve′*] would then give us C_2 as well. So, on the face of it, we have obtained a paradoxical result, namely that if there is sound argument[17] in support of $D_1 \wedge D_2$ then there exists a sound argument in support of $C_1 \wedge C_2$ as well.

Recall, however, that we have a similar argument earlier,

$C_2 \vee D_2$
If C_2 then $v(D_1) > v(C_1)$
If D_2 then $v(D_1) > v(C_1)$
If $v(D_1) > v(C_1)$ then D_1
- - - - - - - - - - - - - - - - - - - -
$\therefore D_1$

We have shown that this is an unsound argument in Sect. 3. And one of the reasons that it fails to be sound is that even if $v(D_1) > v(C_1)$ holds, agent a_1 may not know it, so he may not defect accordingly to maximize his value. If I am correct in maintaining that this is indeed a problem, then the P4 of [Paradox, *Naïve*] may not be true as well. In other words, [Paradox, *Naïve*] is in need of modification. The actions of both agents are guided not only by their rationality but also by their knowledge. So, it is necessary to distinguish between a proposition A and corresponding proposition $K(A)$, where the latter stands for that both agents know A.[18]

[17] Here, Modus Ponens and Hypothetical Syllogism (in the order of *if A then B*, *if B then C*/ ∴ *if A then C*) are assumed to be valid argument forms.

[18] As the behaviors of both agents are guided by logic and shaped by public knowledge, if an agent knows A then the other knows it as well. So here we do not need to introduce a subscript to indicate who the knower is.

As a result, we obtain the following modified argument

P1 $K(D_1 \wedge D_2)$
P2 If $K(D_1 \wedge D_2)$ then $K((C_1 \wedge C_2) \vee (D_1 \wedge D_2))$
P3 If $K((C_1 \wedge C_2) \vee (D_1 \wedge D_2))$ then $K(v(C_1) > v(D_1))$
P4 If $K(v(C_1) > v(D_1))$ then C_1
[Paradox] -
$\therefore C_1$

This becomes a sound argument which, together with its counterpart argument [Paradox′] for C_2, would entail $C_1 \wedge C_2$. Surely, this is an unwelcoming, "paradoxical" result: given that it is logical to defect in a PD game, it is logical to cooperate in a PD game as well. What is the problem after all? The answer is simple, there is simply *no* sound argument in support of defection to begin with! Insofar as we do not hypothesize the existence of such an argument, we would not have come to this paradoxical result in the first place.

Given that it is not *logical* to defect in a PD game, one might suspect that perhaps it is logical to cooperate in a PD game instead. But is it so? Recall that if two twins play a PD game and by definition their final action of cooperation or defection would agree with each other, then it is logical for them to cooperate. Now, for a pair of perfectly rational beings who are not twins, playing a PD game with each other, would it be logical for them to cooperate as well? Adopting the strategy we employed earlier, we can assume that there is a sound argument in support of cooperation. By symmetry we would obtain the truth of $C_1 \wedge C_2$. Then the following argument [Paradox]

P1 $K(C_1 \wedge C_2)$
P2 If $K(C_1 \wedge C_2)$ then $K((C_1 \wedge C_2) \vee (D_1 \wedge C_2))$
P3 If $K((C_1 \wedge C_2) \vee (D_1 \wedge C_2))$ then $K(v(D_1) > v(C_1))$
P4 If $K(v(D_1) > v(C_1))$ then D_1
[Paradox] - - - - - - - - - - - - - - - - - -
$\therefore D_1$

and its counterpart argument [Paradox′] for D_2 would lead us to $D_1 \wedge D_2$. So, again we obtain a "paradoxical" result: given that it is logical to cooperate in a PD game, it is logical to defect in a PD game as well. Again, the "paradox" can be easily avoided by dropping the assumption that it is logical to cooperate.

So, the conclusion here is that for a pair of perfectly rational beings playing a one-off PD game with each other, logic itself *does not* tell them to defect, nor does it tell them to cooperate. Their action can at best be influenced by other factors or concerns. This is not a strange result at all. To illustrate this point, let us consider the following simple example. Two perfectly rational beings are playing a $\oplus$–$\otimes$ game with each other. Each player can either play $\oplus$ or play $\otimes$. If the two players produce

the same sign then they both get one point. If the signs they produce are different then they both get zero point. Does logic tell them to play $\oplus$? Or, does logic tell them to play $\otimes$? Of course not. Evidently, each player knows that the best result would come when they produce the same sign, but they simply have no means to know beforehand what sign the other player would produce. So, despite that both players are perfectly rational beings who abide by logical laws, there is no logical argument to support the statement that $\oplus_1$ *if and only if* $\oplus_2$, nor is there a logical argument for the statement that $\otimes_1$ *if and only if* $\otimes_2$. As a result, logic can offer them no help at all.[19]

Now, an interesting question to ask concerning the PD game is this: would perfectly rational twins knowingly playing with their twin accept the soundness of [Paradox]? After all, the P1 would be true for them. I think the answer is no, and the problematic premiss is P3, because $v(\mathrm{D}_1)$ would be without reference to begin with. As a result, we still would not be bothered by a paradox.

In sum, for players who are ideal twins, there are indeed extra-logical factors such as gene compositions or mystical connections that would help them to come to the logical decision of cooperation, but for non-twins, logic itself neither instructs them to defect nor instructs them to cooperate. If one mistakenly thought that logic does instruct the players to opt for one option, he will be forced by logic to admit the paradoxical result that the players would opt for the other option as well. But, as we have stressed repeatedly, we should not have that false impression in the first place, in particular, logic does not tell us to defect at all.

5 Some Final Remarks

A reviewer for an earlier version of the paper has suggested that by resorting to expected utility, one finds a perfect formulation of the argument underlying the horn of the Prisoner's Dilemma that we have been discussing in Sect. 1. The basic idea is that the expected utilities $EU(\mathrm{C}_1)$ and $EU(\mathrm{D}_1)$ associated with the agent's cooperation and defection can be given as

$$\begin{aligned} EU(\mathrm{C}_1) &= P(\mathrm{C}_2)V(\mathrm{C}_1|\mathrm{C}_2) + P(\mathrm{D}_2)V(\mathrm{C}_1|\mathrm{D}_2), \\ EU(\mathrm{D}_1) &= P(\mathrm{C}_2)V(\mathrm{D}_1|\mathrm{C}_2) + P(\mathrm{D}_2)V(\mathrm{D}_1|\mathrm{D}_2) \end{aligned}$$

respectively, where $P(\mathrm{X})$ is the probability of X, and $V(\mathrm{X}|\mathrm{Y})$ is the expected value of the agent when the action X of the agent is paired with the action Y of the opponent, and evidently $EU(\mathrm{D}_1)$ is greater than $EU(\mathrm{C}_1)$, independently of the probabilities, so it is rational for the agent to defect.

Once again, this is precisely what the present paper sets to argue against. Logic itself does not tell us to defect, impaired rationality does. To see that the expected utility argument above does not work for the case that we are concerned with here, namely, the one-off Prisoner's Dilemma, observe the following:

[19] Note, however, in contrast, if the players happen to be twins that would always produce the same sign, then they can indeed optimize their outcome without resorting to any logical reasoning.

1. It is not a spelt-out, sound argument to start with. In particular, why does $EU(D_1) > EU(C_1)$ imply D_1? In my presentation in the preceding sections, so long as we know D_2 or know C_2, $v(D_1) > v(C_1)$ makes sense and, since both $v(D_1)$ and $v(C_1)$ are definite real values that the agent cares, $v(D_1) > v(C_1) \rightarrow D_1$ is readily acceptable, However, in case we are unsure about whether D_2 or C_2, $v(D_1)$ and $v(C_1)$ become meaningless. In contrast, the $EU(D_1)$ and $EU(C_1)$ here make sense even if we are unsure about whether D_2 or C_2. But, the problem here becomes: why should the agent make decisions based on the relative value of these two expected utilities? The proponents of the expected utility account are responsible for providing the missing link that explains an agent's concern for *expected utility*. In particular, why should the agent care only about the expected utility for a particular move rather than about some (weighted) sum of expected utilities for all possible subsequent moves that are about to come?[20] After all, we are concerned with a *one-off* PD game, and shouldn't we take into consideration, in advance, all possible effects before we make the one and only move?
2. Note that the formulas for the expected utilities $EU(D_1)$ and $EU(C_1)$ resemble that of the expected fitness for two strategies C (always cooperates) and D (always defects) in a social evolutionary context.[21] Specifically,

$$W(C) = Pr(C|C)V(C|C) + Pr(D|C)V(C|D),$$
$$W(D) = Pr(C|D)V(D|C) + Pr(D|D)V(D|D),$$

where $Pr(Y|X)$ is the conditional probability of an X interacting with a Y. Assuming that the chance of meeting a co-operator or a defector is independent of the strategy that one adopts, and that the frequency of individuals that adopt the strategy C is p, then we have

$$W(C) = p\,V(C|C) + (1-p)V(C|D),$$
$$W(D) = p\,V(D|C) + (1-p)V(D|D).$$

Now, clearly, $W(D) > W(C)$, independently of p, as $V(D|C) > V(C|C)$ and $V(D|D) > V(C|D)$, so the selection favours strategy D, and D is an evolutionary stable strategy, which actually makes the value $W(D)$ to decrease from generation to generation. In other words, the system would reach an evolutionary dead end in the end. However, again, why should the *one-off* PD player a_1 care about the evolutionary group fitness in the first place? The evolutionary dead end of all defections evidently suggests that we should have a second thought about it.

3. Recall that in social evolution, the fitness of a strategy at the present generation will affect the frequency of the strategy at the next generation, which in turns

[20] We should take into consideration the possible long term effect that an action can have on $P(C_2)$ and $P(D_2)$, rather than take them as held constants.

[21] See McElreath and Boyd (2007).

affects the fitness of the strategy in the next generation.[22] I claim that if our agent a_1 is rational enough to compare $W'_{\mathrm{D}}(\mathrm{D})$ and $W'_{\mathrm{C}}(\mathrm{C})$—regarding them as more relevant to her benefit—rather than comparing $W(\mathrm{D})$ and $W(\mathrm{C})$ as a blind evolutionary system does, than she would not reach the conclusion that defection is the rational choice. Here $W'_{\mathrm{X}}(\mathrm{Y})$ stands for the fitness of strategy Y after a_1, as a particular individual, chooses to perform X previously. An example suffices to illustrate this point.

Let the playoff matrix be

	C	D
C	10000	0
D	10001	1

And, for simplicity, assume that the action of a_1 would affect the frequency p of co-operators by 1/100—the action (either cooperation or defection) of agent a_1, being an individual in the population himself, would either increase or decrease the population's overall chance of meeting a co-operator. Then we have $w(\mathrm{D}) - w(\mathrm{C}) = 1$, regardless of p, but

$$W'_{\mathrm{C}}(\mathrm{C}) = (p + 1/100)V(\mathrm{C}|\mathrm{C}) + (1 - p - 1/100)V(\mathrm{C}|\mathrm{D}) = w(\mathrm{C}) + 100$$
$$W'_{\mathrm{D}}(\mathrm{D}) = (p - 1/100)V(\mathrm{D}|\mathrm{C}) + (1 - p + 1/100)V(\mathrm{D}|\mathrm{D}) = w(\mathrm{D}) - 100$$

Therefore, $W'_{\mathrm{C}}(\mathrm{C}) - W'_{\mathrm{D}}(\mathrm{D}) = (w(\mathrm{C}) - w(\mathrm{D})) + 200 = 199 > 0$. In other words, even if the agent is concerned primarily with expected utility, the revised expected utility would tell her that there is no rational ground for defection.

Acknowledgments This work was supported in part by the National Science Council, Taiwan (Grant No. NSC 101-2410-H-715-001-) and the Ministry of Science and Technology, Taiwan (Grant No. MOST 102-2410-H-715-001-MY3).

Appendix

The language L is defined in the usual way. A model M for L consists of a non-empty domain set D together with an interpretation generated by an interpretation function I to be defined below.

1. **The atomic truth sets**
   For each atomic formula p_i, $I(p_i) \subseteq D$.
2. **The interpretation $\|\alpha\|_{\mathrm{M}}$ of an expression α with respect to M**
   1. A *hi-world s* is an element of $\Pi_{i=0}^{\infty}(P^*)^i(D)$, where P is the power set operator and P^* is defined by $P^*(A) = P(A)\backslash\{\varnothing\}$ where $\varnothing$ is the empty set.

[22] Here we adopt the notion of social evolution only as a means to help explain how one would have calculated in advance, *in her mind*, what the prospect of her action would be, before she makes her decision. The PD game we are concerned with remains the one-off PD game.

2. A hi-world t is a *sub-hi-world* of s provided that $\pi_i(t) \in \pi_{i+1}(s)$ for all $i \geq 0$, where π_i is the projection into the ith component.
3. If α is an atomic formula, then $\|\alpha\|_M = \Pi_{i=1}^{\infty} U^i$, where $U^1 = I(\alpha)$ and $U^i = (P^*)^i(D)$ for $i > 1$.
4. If α is a formula, then
   $\|\Box\alpha\|_M = \{s \in \Pi_{i=0}^{\infty}(P^*)^i(D) | \ t \in \|\alpha\|_M$ for all sub-hi-worlds t of $s\}$
   $\|\Diamond\alpha\|_M = \{s \in \Pi_{i=0}^{\infty}(P^*)^i(D) |$ there is a sub-hi-worlds t of s such that $t \in \|\alpha\|_M\}$
5. |If α and β are formulas, then
   $\|\neg\alpha\|_M = \|\alpha\|_M^c = \Pi_{i=0}^{\infty}(P^*)^i(D) \setminus \|\alpha\|_M$
   $\|\alpha \vee \beta\|_M = \|\alpha\|_M \cup \|\beta\|_M$
   $\|\alpha \supset \beta\|_M = \|\neg\alpha \vee \beta\|_M$

References

Axelrod R (1984) The evolution of cooperation. Penguin, London

Becker O (1952) Untersuchungen úbef den Modalkalkúl. Westkulturverlag Anton Hain, Meisenheim/Glan

Chalmers D, Hájek A (2007) Ramsey + Moore = God. Analysis 67:170–172

McElreath R, Boyd R (2007) Mathematical models of social evolution, a guide for the perplexed. University of Chicago Press, Chicago

Nowak MA, Highfield R (2012) SuperCooperators: altruism, evolution, and why we need each other to succeed. Free Press, New York

Press WH, Dyson FJ (2012) Iterated Prisoner's Dilemma contains strategies that dominate any evolutionary opponent. Proc Natl Acad Sci USA 109:10409–10413

Ramsey FP (1990) General propositions and causality. In: Mellor DH (ed) Philosophical papers. Cambridge University Press, Cambridge, pp 145–163

Stewart AJ, Plotkin JB (2012) Extortion and cooperation in the Prisoner's Dilemma. Proc Natl Acad Sci USA 109:10134–10135

Trivers RL (1971) The evolution of reciprocal altruism. Q Rev Biol 46:35–57

Tsai C-C (2012) The genesis of hi-worlds: towards a principle-based possible world semantics. Erkenntnis 76(1):101–114

Tsai C-C (2016) Becker, Ramsey, and hi-world semantics. Toward a unified account of conditionals. Croat J Philos 16(1):69–89