The Traveler's Dilemma and its Backward Induction Argument

By

Paul Daniels

A Thesis
Submitted to the Faculty of Graduate Studies
In Partial Fulfillment of the Requirements
For the Degree of

MASTER OF ARTS

Department of Philosophy
University of Manitoba
Winnipeg, Manitoba

Abstract

This thesis is an examination of the traveler's dilemma and its backward induction argument. I begin by explaining relevant terminology, the prisoner's dilemma, and the iterated prisoner's dilemma; the discussion of which aids my examination of the traveler's dilemma and its backward induction argument.

My evaluation of the traveler's dilemma involves a dissection of the game into its different components, a presentation of the salient similarities and differences between the traveler's dilemma and the prisoner's dilemma, and the exploration of three possible solutions. The first two solutions are adapted from ones initially created to solve other backward induction argument problems. The third solution is original and its foundation rests on the unique structure of the traveler's dilemma. I focus on this third solution and consider several objections to it.

I end this thesis with some ancillary comments about the possibility of generalizing the third solution to other backward induction argument problems.

Acknowledgements

The completion of this thesis would not have been possible without the aid of many helpful parties. While I would like to thank everyone who helped me in this endeavor, for reasons of brevity I can't explicitly mention them all here. However the following few deserve special attention.

Thanks to my supervisor, Michael Stack, for agreeing to take me on as a student and for introducing me to the traveler's dilemma. I'd also like to thank Chris Tillman for being a member of my thesis committee and for offering plenty of constructive and insightful feedback at every stage. And thanks to Irwin Lipnowski for being the external member of my thesis committee.

A general thanks to the University of Manitoba philosophy department, staff and faculty, for all the invaluable support and assistance while I was a graduate student.

A special thanks to my family (especially my parents – Ron and Marion) who, throughout my education, provided me with more support and encouragement than anyone could ever ask for.

I'm also thankful to my friends – notably John Dyck and Kirk Weisgerber – for listening to my concerns, offering encouragement, and providing much needed distractions at all the right times.

I’d also like to offer thanks to everyone else that helped me complete this thesis.

Any and all mistakes or omissions are mine alone.

Table of Contents

1: **Introduction**

Consider the following case:

> Two travelers returning home from a remote island, where they bought identical antiques (or, rather, what the local tribal chief, while choking on suppressed laughter, described as "antiques"), discover that the airline has managed to smash these, as airlines generally do. The airline manager who is described by his juniors as a "corporate whiz," by which they mean a "man of low cunning," assures the passengers of adequate compensation. But since he does not know the cost of the antique, he offers the following scheme.
>
> Each of the two travelers has to write down on a piece of paper the cost of the antique. This can be any value between 2 units of money and 100 units. Denote the number chosen by traveler *i* by n_i. If both write the same number, that is, $n_1 = n_2$, then it is reasonable to assume that they are telling the truth (so argues the manager) and so each of these travelers will be paid n_1 (or n_2) units of money.
>
> If traveler *i* writes a larger number than the other (i.e., $n_i > n_j$), then it is reasonable to assume (so it seems to the manager) that *j* is being honest and *i* is lying. In that case the manager will treat the lower number, that is, n_j, as the real cost and will pay traveler *i* the sum of $n_j - 2$ and pay *j* the sum of $n_j + 2$. Traveler *i* is paid 2 units less as a penalty for lying and *j* is paid 2 units more as a reward for being so honest in relation to the other traveler.[1]

This is the traveler's dilemma game. Now consider the following: if each traveler is rational and individually interested in maximizing her own compensation, which integer, between 2 and 100, should they each choose?

Initially, the answer may seem simple enough – choose 100. If each traveler reached this conclusion, they would receive the greatest joint payoff possible. However settling for this is too hasty, since, if traveler *i* believes *j* will choose 100, it seems *i* ought to choose 99 (as it would yield her the most compensation possible - i.e.

[1] Basu, 1994: 391-392 (copyright permission granted)

101). But if *j* is equally rational the same will occur to her. When she realizes that *i* intends to choose 99, her best choice becomes 98 as it would yield her 100. But, then, if *i* reasons that that is what *j* will do, she ought to go with 97. This line of reasoning, known as a backward induction argument, compels each traveler to systematically eliminate all choices until they are only left with one the choice that is never dominated by any other: 2.

This conclusion clearly conflicts with what the travelers should intuitively do. After all, if 2 is the rationally prescribed choice, a pair of rational travelers would do worse than two irrational travelers. The fact that the rationally prescribed course of action, according to the backward induction argument, results in a suboptimal outcome for travelers intent on maximizing payoff is the core problem raised by the traveler's dilemma. As such, solutions to the traveler's dilemma must address this contradiction between intuition and the backward induction argument. The solutions I will consider here will attempt to demonstrate a flaw in the backward induction argument and explain why its conclusion ought to be rejected.

The second chapter will consist of an elucidation of the relevant concepts for this thesis. Because the traveler's dilemma is fundamentally a game theory problem, the majority of the terms discussed will be in this field. The traveler's dilemma is of philosophical importance because it highlights a contradiction between our strong intuition to do one thing and a compelling argument to do something else. In chapter two I will not critically examine or evaluate any issues with the defined concepts, but rather just present the way in which they will be used here. Chapter two will focus on such concepts as utility, dominance, rationality, and the characteristics of a game.

In chapter three I will present and discuss the prisoner's dilemma and the (finite-known) iterated prisoner's dilemma – both share any important qualities with the traveler's dilemma. Because the iterated prisoner's dilemma features a backward induction argument, like the traveler's dilemma, a discussion of it will facilitate my presentation of the key features of typical backward induction arguments.

In chapter four I will present and critically examine, in detail, the traveler's dilemma and the contradiction that intuition and game-theoretic reasoning generates. I will then articulate three potential solutions. The first solution will be one that was originally created to address a different backward induction argument but, as I will show, the solution cannot be successfully applied to the traveler's dilemma in virtue of its unique structure. The second solution is another adapted solution, but unlike the first solution, this solute can be successfully applied to the traveler's dilemma. I will then develop a novel solution to the traveler's dilemma and raise several objections to it. From my discussion of these three solutions will show why, despite being insusceptible to some solutions, the traveler's dilemma is amenable to new kinds of solutions.

The fifth and final chapter will review the achievements of this thesis and end with some peripheral remarks about whether or not the third solution can be generalized for other backward induction argument problems.

2: **Concepts Explicated**

Preamble

This chapter contains an overview of key game theory terminology and assumptions. Because the traveler's dilemma is fundamentally a game theory problem, the comprehension of these concepts will aid in the readers' understanding of the traveler's dilemma. Instead of explicating the relevant concepts in full generality I will formulate the key terminology and assumptions with an eye toward their application in the context of the prisoner's dilemma and the traveler's dilemma. I take this approach because this thesis is philosophical in nature and focused on the contradiction that emerges out of the traveler's dilemma rather than on the analysis of core game theory concepts.

Also note that throughout this thesis decision makers will be referred to as agents. 'Agent' should be understood as being synonymous with such terms as 'player', 'person', and 'individual'. While there may be differences between the meanings of these terms, they should nevertheless be considered interchangeable for the purposes of this thesis.

Dominance Principle

I must first establish how rational agents determine what decision to make in the situations they find themselves in. Specifically in situations where multiple agents must make a choice (or series of choices) from a specified set of choices and where the choices made by those agents each only partially influence which outcome will obtain – the combination of the choices made by all agents involved determine which particular outcome will obtain.[2] The traveler's dilemma is an example of such a situation. The dominance principle is one way agents, in such situations, can determine which choice they ought to make. The dominance principle is important for my purposes because, as we shall see later, it plays a critical role in backward induction arguments.

Before I can present the dominance principle, I must explain the idea of dominance. There are two distinct kinds of dominance. They are:[3]

> Strong Dominance: If an agent prefers all possible outcomes from one choice over any of the possible outcomes from her other choice(s), then that first choice **strongly dominates** the other(s) (for that agent).

And

> Weak Dominance: If an agent prefers some of the possible outcomes of a choice over some of the possible outcomes from her other choice(s), while also having no preference between the other possible outcomes from those choices, then that first choice **weakly dominates** the second (for that agent).

To understand the difference between strong and weak dominance, consider the following case: Two agents, *x* and *y*, must each choose between two alternatives. Assume that *x* must choose between *a* and *b*, while *y* must choose between *c* and *d*. So there are four possible outcomes: (*a, c*), (*a, d*), (*b, c*), and (*b, d*). Now assume that, in this case, *x* will receive a payoff (or reward) based on which outcome obtains. Choice *a* weakly dominates choice *b* (for *x*) if and only if the payoff that *x* will receive from

[2] Luce and Raiffa, 1989: 6-7

[3] Campbell, 1985: 16

choosing *a* is always at least as much as the payoff she would have received had she chosen *b*, and is at least some of the time better. Whereas *a* strongly dominates *b* (for *x*) if and only if the payoff *x* will receive from choosing *a* are always better than the payoff she would have received had she chosen *b*. Put differently if, for *x*, the outcomes which will occur from choosing *a* are never inferior to those of *b* for all possible outcomes and in at least one instance an outcome from *a* is superior to all those from *b*, then *a* weakly dominates *b* for *x*. Whereas if the outcomes from *a* are always superior, then it strongly dominates *b*.[4]

Having said that, I will introduce the dominance principle: [5]

> Dominance Principle: If, for an agent, one choice dominates another choice (either weakly or strongly), then that agent should choose that first choice (rather than the second choice).

But note that adhering to the dominance principle is not always appropriate. It would be imprudent in cases where the decisions made by the agents are not causally independent (i.e. if, in the aforementioned case, which choices *y* makes causally depends on what *x* does). In such cases the dominant choice is not always the rational one since the dominant choice might cause an outcome worse than the alternative.[6] So in such cases other decision making principles are often considered. But because both the traveler's dilemma and the prisoner's dilemma involve causally independent decisions, I will not discuss any other principles here.

[4] Olin, 2003: 109-110

[5] Olin, 2003: 110

[6] Campbell, 1985: 17

Utility Theory

Utility theory is a tool used to sort out what each agent wants. It abstractly represents the preferences of agents and relates their respective goals to the appropriate choice.[7] The possible outcomes[8] which could obtain are assigned different amounts of utility.

Utility function is defined as follows: [9]

> Utility function: A numerical representation of an agent's valuation of an outcome. For each agent, if an outcome is preferable compared to another, than the former is assigned a greater number than the latter. If the agent prefers some outcomes equally, they are assigned the same number.

Utility functions provide information about how the agent ranks the different possible outcomes and should be understood as a mapping from outcomes to real numbers which reflect the preferences of a particular agent. The higher the numerical value, the more she prefers it. Also note that it is not possible to compare the utility functions of different agents because, quite simply, different people want different things. [10]

The reasons why an agent prefers one outcome over another – the reasons why she attributes a greater utility to one outcome over another – are unimportant.[11]

[7] Davis, 1983: 57-61

[8] Understand how choices are distinct from outcomes. Outcomes are the result of the combined choices made by the agents involved. Choices, on the other hand, are the alternatives available to each agent which influence which outcome will obtain. The combination of the choices made by the different agents determines which specific outcome will occur. For clarity, observe that, in the case involving *x* and *y*, *x* has to choose between *a* and *b* (i.e. those are her choices). Whereas the possible outcomes for the case are: (*a, c*), (*a, d*), (*b, c*), and (*b, d*). If *x* were to choose *a*, it would influence which outcome will occur (i.e. making it the case that only (*a, c*) or (*a, d*) could occur – which one in particular depends on the choice made by *y*).

[9] Davis, 1983: 62

[10] Luce and Raiffa, 1989: 33-34

[11] This is because the utility one finds in a particular outcome can be self-regarding (e.g. egoistic in nature) or other-regarding (e.g. altruistic in nature).

The utility of an outcome depends on how much that particular agent, for whatever reason, values it compared to her other possible outcomes.[12] Whichever outcome an agent attributes the greatest utility is the outcome she prefers over all others. If an agent has multiple choices that yield outcomes with identical utilities, then that agent prefers those outcomes equally.

After the preferences of an agent are quantified and the utilities she would receive from the possible outcomes determined, the outcomes can be ranked from most to least desired. Understand that if an agent's preferences are ordered by their desirability, then it is implied that those outcomes are also ordered by their utilities.[13]

Game Theory

The following discussion of game theory will begin with some general remarks about what makes something a game, then move on to clarify the distinctions between different types of games (e.g. zero-sum versus non-zero-sum), and end with an explanation of the game theoretic conception of rationality.

Simply put, game theory is the study of games:[14]

> Game: A situation or case wherein multiple agents must choose between different alternatives. The outcome for each agent depends on her decision(s) as well as the decision(s) made by the other agent(s) (whose interests may conflict with those of the first agent).

In games a certain number of agents interact with one another by making strategic choices which yield specific outcomes.[15] While all games involve two or

[12] Olin, 2003: 107

[13] Rapoport, 1960: 123

[14] Davis, 1983: 6

[15] Like the simple one involving *x* and *y*, described above.

more agents, all the games that I will examine in this thesis involve only two agents.[16] In any game the agents know that their choice(s), in conjunction with the choice(s) made by the other agent (whose interests may conflict with her own), determine which outcome(s) will obtain.[17]

In games the rules that govern the way agents interact with one another are well defined at its outset and the possible outcomes are detailed in the game's description. Also note that the aim of game theory is to determine what particular choice a rational agent ought to make in a given situation when her preferences, and those of the other agent, are given in utility functions and they both know some information (e.g. the rules of the game) but lack other information (e.g. the decision(s) made by the other agent).[18]

Games can be classified based on the kind of decision making involved, which depends on what information is available to the agents. The different kinds of games are: [19]

> Certainty: Each available choice leads to a specific known outcome.
>
> Risk: Each choice leads to one of a finite number of specific outcomes, each of which will occur with a known probability.
>
> Uncertainty: Each choice leads to a finite number of specific outcomes, but the probabilities of those outcomes occurring are either unknown or are not meaningful.

All the games I will discuss in this thesis are games of uncertainty.

[16] As such, all that I say will be phrased for two-person games.
[17] Sorensen, 1999: 282
[18] Rapoport, 1960: 226-227
[19] Luce and Raiffa, 1989: 13

Game Theory – Payoff matrix

Below is a payoff matrix for the simple two-choice two-player case discussed earlier.

		Agent *y*	
		c	*d*
Agent *x*	*a*	*(a, c)*	*(a, d)*
	b	*(b, c)*	*(b, d)*

Payoff matrices like this easily display the choices available to each agent and the possible outcomes that could obtain. As noted earlier, in this game each agent has two choices which are represented respectively by the rows and columns of the payoff matrix. Furthermore, the outcomes in a payoff matrix can be attributed utilities based on the preferences of each agent and, if warranted, ranked according to those preferences. To integrate this, the following modifications could be made to the above payoff matrix:

		Agent *y*	
		c	*d*
Agent *x*	*a*	*(4, 1)*	*(3, 2)*
	b	*(2, 3)*	*(1, 4)*

The numbers in the entries[20] denote the utilities each agent attributes the different outcomes; the first number in each entry indicates the utility *x* attributes that particular outcome whereas the second number is the utility attributed to it by *y*. The entries reflect the respective payoffs for both agents (i.e. the number of units of utility each agent will receive if that outcome obtains). And so, now, given these are the

[20] E.g. *(4, 1), (3, 2),* etc.

utilities, we can see that *x* ranks the possible outcomes in the following order (from most to least preferred): *(a, c), (a, d), (b, c), (b, d)*.[21] This kind of information (i.e. the information reflected in a payoff matrix), as well as other information (e.g. the rules of the game), is assumed to be common knowledge (as defined below) for all the games I discuss later on.

Game Theory – Communication

For this thesis I am specifically interested in contests, which are defined as:[22]

> Contest: A game which does not allow the agent to communicate during, before, or after the game.

When a contest consists of a repeatedly played game, the agents can sometimes engage in a kind of tacit communication, where they can communicate certain information to each other via their choices.[23] Games that are played only once, like the traveler's dilemma, lack this method of eliciting cooperative behavior and provoking reciprocity. Furthermore, the agents in one-round games (i.e. game that are played only once) do not have the ability to engage in adaptive behavior or manipulate the beliefs or preferences of the other agent through the kind of tacit communication that is possible in some repeated games. This difference is crucial because, in one-round

[21] Also note that *y* ranks the outcomes in the reserve order.

[22] Binmore, 1990b: 43

[23] An example of such a game is the iterated prisoner's dilemma, which will be described in chapter three. In such games, this kind of tacit communication takes the form of one agent (or both) making a choice that she intends (or hopes) the other agent will interpret as a signal that she is willing make some specific choice in a subsequent stage of the game.

games, neither agent can learn from experience; the beliefs held by each agent must be formed through individual introspection before the game begins.[24]

Game Theory – Knowledge and Information

Games can also be distinguished from one another based on whether or not the agents posses certain information. As Ken Binmore notes, games of complete information are: [25]

> Complete Information Game: A game wherein information about the structure of the game is taken to be common knowledge held by both agents.

In games of complete information the structural information includes both the rules of the game (e.g. who makes what decisions), as well as the preferences and beliefs of the agents (e.g. the utilities each agent attributes the possible outcomes). Also note that, in games of complete information, the agents begin with identical informational states.[26] For this thesis the games discussed in later chapters will be interpreted as ones of complete information.[27]

Also note that, in each game discussed in subsequent chapters, some information is considered to be common knowledge. Common knowledge is defined as: [28]

[24] Goeree and Holt, 2005: 1403-1418

[25] Binmore and Brandenburger, 1990c: 122

[26] Binmore and Brandenburger, 1990c: 122

[27] Note that I construe the games discussed in this thesis as ones of complete information for simplicity, brevity, and to ensure that that the agents are able to reasonably forecast what decision(s) the other agent will make. But the reader should understand that it is not necessary to construe games as ones of complete information for the agents to be able to reasonably forecast what decision(s) the other agent will make.

[28] Binmore and Brandenburger, 1990c: 106

> Common Knowledge: Information that each agent knows, each agent knows that each agent knows, each agent knows that each agent knows that each agent knows, and so on *ad infinitum*.

Note that real people in real-life instances of games could not have the common knowledge that theoretical agents are assumed to have. Even if an agent in a real-life game correctly guesses what the other agent believes, she does not **know** that her opponent holds that belief.[29] But because I am only interested in idealized theoretical agents, I will set concerns and discussion about real people in real-life games aside.[30]

Game Theory - Zero-Sum and Non-Zero-Sum Games

One last important distinction is between zero-sum and non-zero-sum games. Every game that I will discuss in this thesis is a non-zero-sum game. However I will describe zero-sum games to illustrate how they differ from non-zero-sum games.

While zero-sum games can be useful for studying conflict, they are of little value when studying cooperation. This is because, in such games, there is never an incentive for either agent to act in a cooperative way. The definition of a zero-sum games reflects this:[31]

> Zero-Sum Game: A game in which the agents have diametrically opposed interests (i.e. the gain of one agent necessarily comes at the expense of the other).

[29] Williamson, 1992: 218

[30] It is important that the reader should not consider theoretical agents playing hypothetical games real people (nor can the decisions that those theoretical agents make be assumed to be the same decisions real people would make if playing other real people in the same game). This distinction, between what we assume theoretical agents know and real people might know, highlights why the results reached by economists who conduct experiments with actual persons imply different results or observations than work done on purely theoretical game theory models.

[31] Davis, 1983: 14

Understand that when an agents gains, she is obtaining a greater payoff. Ultimately, in these games, the agents cannot gain simultaneously.

In contrast, there are non-zero-sum games:[32]

> Non-Zero-Sum Game: A game wherein both agents can gain simultaneously (i.e. they can both receive a greater payoff at the same time) if they both make specific choices instead of other available choices. In these games the interests of the agents are not necessarily opposed.

Agents in non-zero-sum games can both gain simultaneously because the interests of the agents are not strictly opposed; there are possible outcomes where both agents could gain, were they both to make specific choices. Moreover, in non-zero-sum games each agent can prevent the other from getting more than the minimum, but it is often the case that such action would be against her own interest; this is a key difference between zero-sum and non-zero-sum games.[33]

Game Theory - Rationality

For the games discussed in this thesis, the agents will be assumed to have perfect logical abilities and perfect recollection. As a result the agents never make mistakes.[34] And while one can say that an agent is rational when she seeks maximal satisfaction of her preferences (i.e. maximizing her payoff)[35], a more robust definition of 'rationality' as it is used here is:[36]

[32] Davis, 1983: 14

[33] Davis, 1983: 84

[34] The kinds of mistakes that I am referring to are ones which would occur if the agents did not have perfect logical abilities and perfect recollection. For instance, failing to realize the consequences of their decisions or not realizing which outcomes a decision could or could not lead to. While other kinds of mistakes are not ruled out (e.g. an agent might make a choice other than that which she wanted to make in virtue of a "trembling hand"), but, for simplicity, I would like to rule out such mistakes as well.

[35] Olin, 2003: 139

[36] Rapoport, 1960: 107-108

> Rationality: An agent is rational if and only if she adopts a strategy which she believes will yield her the best possible outcome, having considered: (1) the possible outcomes of all her available choices, (2) her preference ordering among the outcomes of each available choice, (3) the utility functions she believes the other agent have regarding the possible outcomes, (4) that the other agent is equally rational, and (5) all available information.

In determining which choice ought to be made, rational agents should eliminate the choices that would never be best and recognize that the other agent will do the same.[37] And so, when each agent is construed this way, and is well-informed (e.g. about the rules of the game, available choices, possible outcomes, utility functions, etc.), they can be expected to make their respective ideal choice.[38]

The agents involved in the games I will discuss in this thesis are construed in the above way because irrational agents are unpredictable, or at least are less predictable than rational ones.[39] Failing to assume that the agents are rational in the above sense would allow them to be fallible; fallibility or unpredictability would inhibit their ability to determine which course of action is best. Eliminating the possibility of agent fallibility is advantageous because, if an agent is unlikely or unable to discern the ideal choice, the ideal choice nevertheless remains ideal. However this advantage comes at a cost: the agents can no longer be considered real

[37] Goeree and Holt, 2005: 1405

[38] Binmore, 1990a: 14

[39] Basu, 1990: 38

people playing games; "homo sapiens are replaced with homo economicus".[40] But this cost should not be considered an issue here, given my intentions in this thesis.[41]

Game Theory – Nash Equilibrium

One last concept to note is the idea of Nash equilibria. Very briefly, this can be defined as: [42]

> Nash Equilibrium: A set of strategies, one for each agent, such that neither agent could do better by deviating unilaterally.

In this context an agent deviates unilaterally from a Nash equilibrium when she selects a strategy other than one which results in the Nash equilibrium. In games with only one round, the set of strategies for each agent are simply whichever respective choice they could each make (deviating from which could in no case be better for her). Whereas, in iterated games, the set of strategies for each agent are the choices they will each make at every possible decision point in the game. An example of a Nash equilibrium is confessing in the prisoner's dilemma (explained in the next chapter). Having said that, the concept of Nash equilibrium is nuanced, and much more can be said about it than what I have offered here, but this definition will suffice for my purposes in this thesis.

Review

This chapter introduced relevant game-theoretic concepts as they will be used in this thesis.

[40] Binmore, 1990a: 14-15

[41] Which are to analyze the traveler's dilemma and determine why the course of action prescribed by its backward induction argument is not optimal – something only achievable when construing the agents as homo economicus.

[42] Basu, 2007: 92

At this point the reader should understand that each of the problems discussed in subsequent chapters are non-zero-sum two-person games of complete information, which are also contests, wherein the agents face decisions of uncertainty. The agents in each of these problems should be understood as being rational in the above defined sense, and perfect with respect to recollection and logical abilities.

This completes my overview of the key terminology that will be used throughout this thesis. I will now explain the prisoner's dilemma, the iterated prisoner's dilemma, and its backward induction argument.

3: **Related Problems**

Preamble

In this chapter I will present the prisoner's dilemma and its variant, the iterated prisoner's dilemma.[43] I raise and discuss the prisoner's dilemma because Kaushik Basu argued that the traveler's dilemma is related to the prisoner's dilemma on the basis that, if the agent's options are restricted to only two choices[44], the traveler's dilemma has the exact same structure as the prisoner's dilemma.[45] Having noted why I present the prisoner's dilemma in this thesis, I will not thoroughly discuss the similarities and difference between the two games until after I have presented the traveler's dilemma in chapter four. Presenting the material in this way best ensures that my examination of the traveler's dilemma and its backward induction argument is clear.

My presentation of the iterated prisoner's dilemma will focus on its backward induction argument. I do so to make certain that the reader understands some fundamental characteristics of backward induction arguments before I present and discuss the traveler's dilemma and its backward induction argument. A sufficient understanding of backward induction arguments is important because they play a

[43] Note that whenever I speak of "the prisoner's dilemma" I am specifically referring to a one-round prisoner's dilemma game, which is distinct from an iterated or repeatedly played prisoner's dilemma game. Whenever I speak of the latter I will refer to it as "the iterated prisoner's dilemma".
[44] Specifically restricted such that their only choices are either 2 or 3.
[45] Basu, 1994: 392 (footnote 1)

central role in how the respective contradictions are generated in both the iterated prisoner's dilemma and the traveler's dilemma.

Also note that, in order to be concise, I will only raise solutions to the prisoner's dilemma and the iterated prisoner's dilemma that suggest important revisions to them to their presentations. I do so to ensure the focus is on the core problem in these games.

Prisoner's Dilemma

The case behind the prisoner's dilemma is typically construed as something like the following. Two criminals, call them *X* and *Y*, are arrested and are isolated from one another so that they cannot communicate. The prosecutor, interested in obtaining a confession from one or both prisoners, provides each of them with the following information:

(A) If *Y* confesses and *X* does not, *Y* will go free while *X* spends ten years in jail.
(B) If both *Y* and *X* confess, they will each spend five years in jail.
(C) If both *Y* and *X* do not confess, they will each spend one year in jail.
(D) If *Y* does not confess and *X* does confess, then *Y* will spend ten years in jail and *X* will go free.

If both agents prefer less jail time, what should a rational agent do in this situation – confess or cooperate with the other criminal and remain silent (i.e. do not confess)?

The important information from this case can be represented in the following payoff matrix, which reflects the utilities for each agent for each possible outcome:[46]

[46] Note that the numbers in this payoff matrix are negatives in order to reflect the **disutilities** each agent has for each possible outcome.

		Prisoner *Y*	
		Y Confesses	*Y Does not Confesses*
Prisoner *X*	*X Confesses*	*(-5, -5) [B]*	*(0, -10) [D]*
	X Does not Confesses	*(-10, 0) [A]*	*(-1, -1) [C]*

Before directly addressing the above question, I must first note that this version of the prisoner's dilemma is open to several uninteresting solutions or methods of dispelling the problem. For instance, one could argue that the case, as described, lacks sufficient information to make a decision. After all, should an agent be concerned about reprisal for confessing? Or should she feel compassion or loyalty for her fellow criminal? Objections like these are philosophically uninteresting because they are based on inessential aspects of the story and fail to address the presented dilemma. To insulate my discussion against these sorts of pseudo-solutions some remarks and clarifications about the case must be made.

First of all, each agent must be understood as being only interested in seeking the least amount jail time; they are each indifferent about the fate of the other prisoner.[47] And, because utilities can be self- or other-regarding, the numbers in the payoff matrix can be said to reflect their respective disdain for jail time as well as other factors (like fear of reprisal). Having said that, in the prisoner's dilemma the exact utilities attributed to the possible outcomes by the agents are ultimately unimportant;

[47] Campbell, 1985: 5

what **is** important are the preference rankings of the possible outcomes, for each agent.[48]

Furthermore, the prisoner's dilemma, like the traveler's dilemma, must be construed as being self-contained. The decisions each agent makes in the game must be taken as ones that will not impact their lives in any subsequent way. It is also important to understand that the agents make their choices at the same time and that there is no probabilistic or causal dependence between their decisions. Moreover each agent must be assumed to believe that the prosecutor (or the manager in the traveler's dilemma) is telling the truth, that the other agent faces an identical situation, and that this information is all common knowledge.[49]

Having said that, observe how, by confessing, it is possible for each agent to get her most preferred outcome (i.e. no jail time) while also avoiding her least preferred outcome (i.e. ten years). But, at the same time, confessing also eliminates her second most preferred outcome (i.e. one year in jail). So if both agents confess they will both be worse off as a result since they will both receive their third most preferred outcome (i.e. five years in jail). In contrast, if neither of them confesses, they will both receive their second most preferred outcome.

But even though both agents will realize this, confessing nevertheless remains preferable because confessing dominates not confessing. To elucidate why, consider the following: if an agent (or both for that matter) assume that the other agent will

[48] Olin, 2003: 139

[49] Campbell, 1985: 5

confess, confessing remains the rationally prescribed choice because, for each agent, the five years in jail outcome is more preferred than the ten years in jail outcome. Alternatively, if an agent assumes her opponent will not confess, then she (again) ought to confess because she prefers her no jail time outcome over her one year in jail outcome.[50] Despite the fact that both know they would better off if they both did not confess, the rationally prescribed course of action remains confessing. So a successful solution to the prisoner's dilemma must diagnose a flaw in this basic argument for confessing.

The basic argument for confessing can be explicitly stated, with reference to the payoff matrix for the prisoner's dilemma above, as:[51]

(1) *Y* will confess or *Y* will not confess.
(2) *X* will confess or *X* will not confess.
(3) If *X* will not confess and *Y* will confess, then outcome *A* will occur.
(4) If *X* will not confess and *Y* will not confess, then outcome *C* will occur.
(5) Therefore, if *X* will not confess, then outcome *A* will occur or outcome *C* will occur.
(6) *Y* prefers outcome *A* more than outcome *C*.
(7) If (5) and (6), then, if *X* will not confess, *Y* should confess.
(8) Therefore, if *X* will not confess, then *Y* should confess.
(9) If *X* will confess and *Y* will confess, then outcome *B* will occur.
(10) If *X* will confess and *Y* will not confess, then outcome *D* will occur.
(11) Therefore, if *X* will confess, then outcome *B* will occur or outcome *D* will occur.
(12) *Y* prefers outcome *B* more than outcome *D*.
(13) If (11) and (12), , if *X* will not confess, then *Y* should confess.
(14) Therefore, if *X* will confess, then *Y* should confess.
(15) Therefore, *Y* should confess.

[50] Campbell, 1985: 5-6
[51] The basic argument for confessing is applicable for any arbitrary *Y*, *X* that meet the idealization stipulation.

Observe how confessing leads both agents to an outcome less desirable than another outcome which they could achieve.

Despite the fact that both agents rank mutual non-confession as a more preferable outcome than mutual confession, their rationality has doomed them to the outcome (between mutual confession and mutual non-confession) that they each prefer least. The salient point of this is summarized in an observation by Anatol Rapoport: the prisoner's dilemma "is a powerful example of a social situation in which the "sum" of the two individual interests adds up to a disadvantage to both".[52] Because the basic argument for confessing leads both agents to an outcome less desirable than another outcome which they could achieve, there appears to be reason to believe that the agents should pursue that other more desirable outcome – that they should not confess.[53]

And so, for some, the conclusion reached by the basic argument for confessing intuitively conflicts with what rational agents should be able to accomplish in the prisoner's dilemma. However, because my primary aim in this thesis is the traveler's dilemma and its backward induction argument, I will not examine solutions that seek to diagnose a flaw in the basic argument for confessing for the prisoner's dilemma. For my purposes here, confessing will be considered the rational course of action in the prisoner's dilemma.

[52] Rapoport, 1960: 177

[53] One could infer the contradictory conclusion that *Y* should confess and *Y* should not confess.

Iterated Prisoner's Dilemma

I will now explicate the iterated prisoner's dilemma. After which I will elucidate what our intuition tells us rational agents should do, the different course of action dictated by the backward induction argument, and why the conclusions from those two arguments generate a contradiction. I will then explain the essential and common aspects of backward induction arguments.

As the name suggests, the iterated prisoner's dilemma is a game in which two agents engage repeatedly in prisoner's dilemma games, for a specific number of rounds. Iterated prisoner's dilemmas can be distinguished as one of two versions: finite-known iterated prisoner's dilemmas, and finite-unknown iterated prisoner's dilemmas.[54] The former refers to instances where the game is played a finite number of rounds, which is at least twice. In these games both agents know, at the outset, the duration of the game (i.e. how many rounds they will play). In contrast, in finite-unknown iterated prisoner's dilemmas neither agent knows the duration of the game, so at no point during the game does either agent know when she is playing the last round. Also, at the outset of the game, the agents know which kind of iterated prisoner's dilemma is being played.

While there are interesting issues worth exploring in finite-unknown iterated prisoner's dilemma games, in this thesis I am only concerned with finite-known iterated prisoner's dilemma games. As such, when I refer to iterated prisoner's

[54] For completeness I must note that here are also infinitely-known and infinitely-unknown versions of the game, but these versions are vastly differently than the finite-known version and, because my interested is solely with that version of the game, these other two versions are not mentioned or discussed in this thesis in any greater capacity than this.

dilemma games I should be understood as only referring to finite-known iterated prisoner's dilemmas.

In the iterated prisoner's dilemma, in all rounds after the first, the payoffs that the agents received from previous rounds are considered common knowledge.[55] Furthermore, as in the prisoner's dilemma, no formal or explicit communication is allowed between the agents.[56] However in the iterated prisoner's dilemma each agent has the ability to relay specific information to the other agent tacitly through the choices they make. As such, in every round but the last, each agent must consider the possibility that her move might influence the choices the other agent will make in all subsequent rounds.[57] This indirect communication is the only way in which the agents can communicate any information to one another.

Iterated Prisoner's Dilemma – Arguments

Even if confessing is accepted as the rational thing to do in the prisoner's dilemma, in iterated prisoner's dilemmas, it seems intuitively clear that confessing in every round should not be the rationally prescribed course of action. This is because both agents have the opportunity to each receive a greater total payoff by mutually not confessing at least some of the time. And because of the tacit communication mechanism available in the iterated prisoner's dilemma, rational agents, who are

[55] I.e. the payoffs distributed from the previously played round(s) are considered information that they both know before they make their choices for the next round. (I.e. if, in the first round of the game, prisoner *X* does not confess while prisoner *Y* confessed, both have that information before they make their respective choices in the second round. They also both know that they both know that information, and *ad infinitum*.)

[56] I.e. as stated in chapter two, both these games (as well as the traveler's dilemma) are construed as contests.

[57] Pettit and Sugden, 1989: 170

primarily interested in maximizing their own respective total payoffs, would consider the decisions made by the other agent in earlier rounds to determine if the mutual non-confession outcome is attainable in future rounds.

Since a rational agent would recognize this ability to gain more (i.e. more than were each agent to confess in every round), a rational agent would utilize the tacit communication mechanism to influence the decisions of the other agent; a rational agent would attempt to elicit a mutually advantageous strategy (i.e. induce the other agent to pursue the mutual non-confession outcome).[58] In virtue of the game structure, the only way an agent could signal a willingness to not confess in future rounds is to not confess in earlier rounds. So it seems that both agents would not confess at least some of the time (in an attempt to elicit a mutually advantageous strategy), while also looking for signs that the other agent is willing to engage is such a strategy.

Having said that, there is also an incentive to not confess early in the game. Agents who confess in the first round, or repeatedly confess, could face punishment in later rounds (e.g. the other agent could decide to never not confess in any later rounds, thereby ensuring that the agent who failed to initially not confess would never obtain either of her two most preferred outcomes in later round). If this were to occur, that agent who always confessed would be guaranteed her third most preferred outcome for all of those later rounds (provided she kept confessing) – which

[58] Olin, 2003: 154

is clearly to her detriment than if she had been able to receive her second most preferred outcome at least some of the time.

In short, by not confessing in at least some rounds each agent can get a greater payoff than if they were to both confess in every round. The agents can achieve this by taking steps to influence the decisions that the other agent will make. Intuitively, this is what rational agents should do.

But there is another convincing argument about what rational agents should do in the iterated prisoner's dilemma. Suppose the two agents, prisoner *X* and prisoner *Y*, are engaged in an iterated prisoner's dilemma which will last for three rounds. Moreover, the payoff matrix for each round is identical to the payoff matrix described above for the prisoner's dilemma. While both agents will recognize the intuitive argument for not confessing above, even if not confessing initially seems like the rationally prescribed course of action for some rounds, no matter what the agents do in the first two rounds, what should be their respective choices in the last round? Clearly, two rational agents ought to both confess.[59]

Even though the argument from intuition suggests that there are other factors which influence what choices rational agents should make, such factors cannot influence their decisions in the last round. That is, no matter what has transpired thus far in the game, after the last round, there are no longer any opportunities to influence, or be influenced by, the decisions of the other agent. Neither agent will have an opportunity to utilize the information they will receive after the last round

[59] Provided the basic argument for confessing in the prisoner's dilemma is sound.

has been played (i.e. the decisions made in that last round). Since there is no opportunity for further influence after the final round of the game, that final round is equivalent to a one-round prisoner's dilemma. As a result the basic argument for confessing can be applied in the last round; both agents ought to confess in the last round.

Because both agents will realize that they ought to confess in the last round, they will both realize that it is not the case that the decisions made in the second last round will affect or influence their decisions in the final round. In virtue of this, the second last round is also equivalent to a one-round prisoner's dilemma. So the basic argument for confessing also applies in the second round as well.

Having said that, the same can be said for the third last round (i.e. for this three round game, the first round). Had we considered a game with more than three rounds, this pattern would continue no matter how many rounds there are in the game being played. Ultimately, then, mutual confession will occur at every round. This is the backward induction argument for the iterated prisoner's dilemma.

Before I continue any further, for clarity, I must articulate some important assumptions behind the backward induction argument. First, the agents are rational (as defined in chapter two) for the duration of the game; and at the outset of the game, both agents believe that the other agent is rational.[60] Second, throughout the game the

[60] While both agents believe the other agent is rational at the outset of the game, both agents could refine their beliefs during the game. For instance, if *X* received information that *Y* is making irrational choices, *X* could refine her beliefs about whether or not *Y* is rational. *X* would then play the game with those refined beliefs. Observe how this refinement can only occur in repeated games, like the iterated prisoner's dilemma, and not games like the prisoner's dilemma or the traveler's dilemma. This is

agents have perfect recollection and perfect logical abilities, thereby ensuring that they will not make mistakes. Third, the agents are aware that they are engaged in an *m*-stage[61] iterated prisoner's dilemma.[62]

I can now explicitly state the intuitive argument to not confess in every round, the backward induction argument, and the contradiction generated when their conclusions are joined together. Note that the presentations of these arguments are formulated for a three-round version of the iterated prisoner's dilemma, where, at each stage, the payoff matrix is:[63]

		Prisoner *Y*	
		Y Confesses	*Y Does not Confesses*
Prisoner *X*	*X Confesses*	*(-5, -5) [B]*	*(0, -10) [D]*
	X Does not Confesses	*(-10, 0) [A]*	*(-1, -1) [C]*

Note first the argument from intuition:

(1) If *Y* will confess in every round, then it is not the case that outcome *C* will ever occur.
(2) *Y* prefers outcome *C* over outcome *B* or outcome *D*.
(3) If (2) and (3), then *Y* should not confess in at least some rounds.
(4) Therefore, *Y* should not confess in at least some rounds.
(5) If *X* will confess in every round, then it is not the case that outcome *C* will ever occur.
(6) *X* prefers outcome *C* over outcome *B* or outcome *A*.
(7) If (7) and (8), then *X* should not confess in at least some rounds.
(8) Therefore, *X* should not confess in at least some rounds.
(9) If *X* should not confess in at least some rounds and *Y* should not confess in at least some rounds, then both agents should not confess in at least some rounds.
(10) Therefore, both agents should not confess in at least some rounds.

because, in non-repeated games, neither agent can receive any information that might warrant a revision of her beliefs before she decides on a course of action.

[61] Where *m* represents the number of rounds being played.

[62] Olin, 2003: 155-156

[63] Also note that this payoff matrix is identical to payoff matrix for the prisoner's dilemma considered earlier in this chapter.

And the backward induction argument:[64]

(11) The basic argument for confessing applies in round *n.*

(12) If the basic argument for confessing applies in round *n*, then both agents should confess in round *n.*

(13) Therefore, both agents should confess in round *n.*

(14) If both agents should confess in round *n*, then the basic argument for confessing applies in round *n-1*.

(15) If basic argument for confessing applies in round *n-1*, then both agents should confess in round *n-1*.

(16) Therefore, both agents should confess in round *n-1*.

(17) If both agents should confess in round *n-1*, then the basic argument for confessing applies in round *n-2*.

(18) If basic argument for confessing applies in round *n-2*, then both agents should confess in round *n-2.*

(19) Therefore, both agents should confess in round *n-2*.

(20) If both agents should confess in round *n* and both agents should confess in round *n-1* and both agents should confess in round *n-2*, then both agents should confess in every round.

(21) Therefore, both agents should confess in every round.

And when the conclusions from these two arguments are brought together:

(22) Both agents should not confess in at least some rounds and both agents should confess in every round.

This contradiction, generated by the conjunction of the conclusions of the two arguments, is considered the core problem in the iterated prisoner's dilemma. Solutions to the iterated prisoner's dilemma must diagnose a flaw in the argument for the above contradiction. While it may be that the flaw lies in the intuition argument, solutions generally seek to identify a flaw in the backward induction argument.

Backward Induction Arguments

Having presented a backward induction argument, I will now detail some general characteristics that are common to all backward induction arguments.

[64] Assume that the agents are playing a three round iterated prisoner's dilemma *n* represents the last round of the game, *n-1* represents the second round of the game, and *n-2* represents the first round of the game.

Because the backward induction argument I have presented is phrased for the iterated prisoner's dilemma, this discussion will be in iterated prisoner's dilemma terms. But the reader should understand that these remarks can be generalized for any backward induction argument.

Backward induction arguments apply the dominance principle to each pair of choices, in a sequence of pairs of choices. Backward induction arguments also require that each agent, at each decision stage of the game being played, have only two choices. If the game structure does not present a sequence of pairs of choices, then the game must be construed in such a way that there is a sequence of choices (and, for each decision point, there must be only two choices available for consideration). The iterated prisoner's dilemma is an example of a game that naturally presents a sequence of decision stages that each have only two possible choices, whereas the traveler's dilemma is a game that must be construed in a specific way for its backward induction argument. In either case the dominance principle eliminates the dominated choice in each pair of choices, starting at one end of the sequence of choices and working towards the other end (in the iterated prisoner's dilemma, starting with the last round of the game).[65]

In the iterated prisoner's dilemma arguing that rational agents should confess in the last round is a critical part of its backward induction argument. This is because, before the first round is played, neither agent has obtained any information to contradict the course of action that the backward induction argument prescribes for

[65] Sorensen, 1999: 279

that round. If one can tenably argue that confessing is not the rational choice in the last round[66], the backward induction argument fails. This can be generalized for all backward induction arguments in the following way. Backward induction arguments all begin by recommending a specific choice (i.e. the dominant choice) for the decision stage at one end of the sequence of choices. If that recommended choice can be denied at that decision stage the backward induction argument cannot commence. Similarly, if that recommended choice can be denied at some other decision stage, then the backward induction argument cannot continue further in the sequence of pairs of choices beyond that particular stage. Also note that backward induction arguments are considered before either agent makes her first choice.

Furthermore, backward induction arguments can be expanded for any number of rounds. Even if the game has a million rounds, the backward induction argument still applies. However the million-round game only seems to reinforce our intuition that not confessing should be, at least some of the time, the rational thing to do.[67]

Having explained the key aspects of the prisoner's dilemma, the iterated prisoner's dilemma, and backward induction arguments, I will now move on to discuss the traveler's dilemma.

[66] I.e. Diagnose a flaw in the basic argument for confessing.

[67] Again, this is because, if the game will last for a million rounds, mutual confession in every round will yield a significantly smaller total payoff for each agent than if they had they not confessed at least some of the time.

4: **Traveler's Dilemma Discussed**

Preamble

My aim in this chapter is two-fold: to present the traveler's dilemma and to evaluate solutions that diagnose a flaw in its backward induction argument. To that end I will raise three solutions, after articulating the key characteristics of the game. The first two solutions were originally designed for similar backward induction argument problems; I will argue that the first one cannot be successfully adapted to the traveler's dilemma while the second can. The third solution, a novel one, focuses on the unique structure of the traveler's dilemma. Because my focus will be on this third solution I will raise several objections to it and offer tenable replies.

Traveler's Dilemma

The Traveler's Dilemma is based on the following case. Two independent travelers (call them *i* and *j*) are returning on the same flight from the same destination. Coincidentally, both these travelers purchased identical antiques while at that destination. However, on their flight both their artifacts were damaged. Because the airline manager is unsure of how much the artifacts cost, he devises a scheme to determine how much compensation to offer the travelers. After the travelers are separated, so that they cannot communication with one another, the manager has them each write down a number between 2 and 100 that reflects the cost of the antiques. But the manager is concerned that the travelers are dishonest and will not write the actual cost of the artifact. So he provides the travelers with the following further information:

If they write the same number, he will conclude that they are both telling the truth and will give each traveler that amount. (So, if the number *i* picks (call it n_i) is identical to the number *j* picks (n_j), both travelers will receive n_j.)

If, however, *i* writes a higher number than *j* (i.e., $n_i > n_j$), the manager will conclude that *j* is telling the truth while *i* is not (or vice versa if *j* picked a higher number than *i*). The manager will consequently take the lower number to be the real value of the artifact.

If that occurs the manager will then distribute the real value to both agents with the following modifications: the traveler determined to be telling the truth (i.e. the traveler who wrote a number lower than the number written by the other traveler) will be reward two additional units of compensation, while the dishonest traveler will be penalized two units (i.e. the dishonest traveler will receive four units less than the total amount the honest traveler received).[68]

So, in this situation, what number should a rational agent choose?

Intuitively a pair of rational agents should each choose some high number (e.g. 99, 98, or 97). However, as I will show, a backward induction argument concludes that both agents should choose 2. Before I explain the arguments that recommend these distinct courses of action I must make some important preliminary remarks about the traveler's dilemma.

[68] Basu, 1994: 391-392

Traveler's Dilemma – Characteristics

First, for simplicity, restrict the choices available to only integers. When the choices are restricted in this way a truncated payoff matrix for the traveler's dilemma is:

						Traveler *i*			
		100	99	98	97	…	4	3	2
	100	*(100, 100)*	*(97, 101)*	*(96, 100)*	*(95, 99)*	…	*(2, 6)*	*(1, 5)*	*(0, 4)*
	99	*(101, 97)*	*(99, 99)*	*(96, 100)*	*(95, 99)*	…	*(2, 6)*	*(1, 5)*	*(0, 4)*
	98	*(100, 96)*	*(100, 96)*	*(98, 98)*	*(95, 99)*	…	*(2, 6)*	*(1, 5)*	*0, 4)*
	97	*(99, 95)*	*(99, 95)*	*(99, 95)*	*(97, 97)*	…	*(2, 6)*	*(1, 5)*	*(0, 4)*
Traveler *j*	…	…	…	…	…	…	…	…	…
	4	*(6, 2)*	*(6, 2)*	*(6, 2)*	*(6, 2)*	…	*(4, 4)*	*1, 5)*	*(0, 4)*
	3	*(5, 1)*	*(5, 1)*	*(5, 1)*	*(5, 1)*	…	*(5, 1)*	*(3, 3)*	*(0, 4)*
	2	*(4, 0)*	*(4, 0)*	*(4, 0)*	*(4, 0)*	…	*(4, 0)*	*(4, 0)*	*(2, 2)*

In addition to that restriction, the travelers' artifacts must be taken to be perfectly similar and to have received identical damage (e.g. both were completely destroyed). And that they are each primarily interested in maximizing their own compensation (i.e. payoff).[69]

Assuming that the agents are each intent on obtaining the greatest individual payoff possible is important because, otherwise, the agents might not "play the game". Instead they might only try to receive compensation equal to what they paid for their artifacts or, alternatively, what they think is fair. But, if they want to receive as much as possible, the actual price they paid for their artifacts is irrelevant. Making the agents primarily interested in maximizing individual payoff also allows us to set

[69] One could say that they are both greedy antique traders who purchased their artifacts with the intention of selling them for as much as possible.

their preferences for outcomes, from most to least preferred, such that they correspond to the outcomes that yield the highest payoff (i.e. a hundred and one) to the lowest payoff (i.e. zero). Having said that, for clarity, understand that the payoff an agent will receive from an outcome denotes how many units of utility she will receive if that outcome is achieved (rather than how many dollars or units of currency she will receive).

Moreover, the traveler's dilemma, like the prisoner's dilemma, must be considered a game which will have no other repercussions on the lives of the agents beyond the payoffs they will each receive from the game. The reader should also note that while the agents will do whatever they believe will yield them respectively the greatest payoff, the agents are not concerned with getting more compensation than the other agent. However there **is** an implicit incentive for each of the agents to try to do better than the other agent. This incentive is that, in some instances, the only way an agent can get the most compensation is at the expense of the other agent. But the traveler's dilemma remains a non-zero-sum game because both agents can gain more, at the same time, if they both make certain specified choices.

Having said that, deconstructed, the traveler's dilemma contains two component games: the compensation game and the punishment/reward game. The compensation game is the primary game, wherein both agents receive a payoff equal to the lowest number selected. The punishment/reward game adds or subtracts a fixed amount (i.e. two) from the payoffs of each agent depending on whether or not the number she chooses is lower than the number picked by the other agent. If an agent's number is lower than that selected by the other agent, she receives a reward of two units of utility while the other agent receives the penalty of two fewer units than the lowest number selected. If the agents choose the same number no reward or punishment is applied to either agent's payoff. So if an agent chooses a number lower than that selected by her opponent, she receives a number of utility equal to the number she chose, plus two. If an agent chooses a number higher than that of her opponent, receives utility equal to **the number selected by the opponent**, minus two. And if both agents select the same number, they will both receive that number of units of utility. For example, if n_i was 66 while n_j was 79, *i* would receive a payoff of sixty-eight and *j* a payoff of sixty-four. In contrast, if both agents picked 66, they would both receive a payoff of sixty-six. Because of the reward/punishment game, both agents must consider what the other agent might do.[70]

Note that even though the pair of choices where both agents select 100 is efficient[71], it is not a Nash equilibrium[72]. This is because, by picking the same number (so long as it is between 3 and 100) one agent **could** have received a greater payoff if she had picked that number minus one (e.g. if one agent selects 100, the other agent receives a greater payoff if she chooses 99 instead of 100). The best response for any agent, if the other agent chooses *k* [73], is to select *k-1* – it is the only choice that will maximize payoff for that agent. Only if *j* selects 2 is it the case that *i* would be best off

[70] Sarangi, 2000: 29-30

[71] By "efficient" I mean the outcome where the agents will maximize joint payoff.

[72] I.e. a set of strategies, one for each agent, such that neither agent could do better by deviating unilaterally.

[73] Where *K* is any specific number between 3 and 100.

picking the same number because doing so is the only way she can get more than zero compensation. However, neither agent knows, before they make their respective choices, what the other agent will choose.

Traveler's Dilemma – Similarities with the Prisoner's Dilemma

As noted at the outset of chapter three, Kaushik Basu identified that the traveler's dilemma is related to the prisoner's dilemma because an alternate version of the traveler's dilemma[74] where the agents' available choices are restricted to only 2 or 3 has the same payoff matrix as that of a prisoner's dilemma game.[75]

But the traveler's dilemma also shares many important characteristics with the iterated prisoner's dilemma. For instance, the backward induction augments in both rely on the application of the dominance principle in a series of successive stages. And, like the iterated prisoner's dilemma, to solve the traveler's dilemma one must provide a justification for why two rational agents would fail to choose 2 despite the seeming persuasiveness of the backward induction argument.

Traveler's Dilemma –Arguments

Before I explain the two arguments that generate a contradiction, I will explicitly note the key assumptions in traveler's dilemma. They are:

[74] Call this version the two-choice traveler's dilemma. Unless specifically noted, I should be understood as referring to the regular version of the game (i.e. the game as it was presented at the outset of this chapter).

[75] Basu, 1994: 392 (footnote 1)

(I) The agents are rational[76] for the duration of the game.
(II) The agents are engaged in a one-round traveler's dilemma game (as described at the outset of this chapter).
(III) Throughout the game the agents have perfect recollection and perfect logical abilities – they do not make certain kinds of mistakes.[77]
(IV) The agents are primarily interested in maximizing their own compensation.
(V) The agents face a decision of uncertainty.[78]
(VI) The game is a contest.[79]
(VII) The game is one of complete imperfect information.[80]
(VIII) Both agents will believe (I-IX) throughout the game.
(IX) At the outset of the game (I-IX) are considered common knowledge.

With that in mind I can now explain what our intuition tells us a rational agent ought to do in the traveler's dilemma. After presenting that argument I will explain why the backward induction argument prescribes 2 and show how, when the conclusions from these two arguments are put together, they form a contradiction.

Intuitively, each agent would recognize that 100 is the efficient choice. However, as note above, the efficient choice is not a Nash equilibrium. As a result selfishness would lead each agent to try to gain more individually by picking something other than 100.[81] But there is a limit to how far from the efficient choice

[76] Recall that an agent is rational if and only if she makes the choice which she believe will yield the best possible outcome for her having considered: (1) the possible outcomes of all her available choices, (2) her preference ordering among the outcomes of each available choice, (3) the utility functions the other agent attributes to the possible outcomes (over which the first agent has no control), (4) that the other agent is equally rational, and (5) everything that is common knowledge (e.g. rationality among agents, rules of the game, etc.).
[77] For instance, the agents will not fail to realize the consequences of their decision. That is, they will realize (before making their respective decisions) which outcomes a choice could, or could not, lead to.
[78] Each choice, for each agent, leads to a finite number of specific outcomes. The probabilities of those outcomes are unknown (or are not meaningful) to the agents.
[79] The agents are not able to communicate during, before (so that they cannot make agreements about what they will do), or after the game (so that they cannot make side payments or attack the other agent).
[80] Both agent have identical information about the structure of the game.
[81] 100 is ruled out here as an acceptable choice because, at least intuitively, picking 100 is incompatible with what agents interested in maximizing their own payoff would selfishly accept (since, by picking 100, an agent cannot receive a payoff greater to the number she chose – a unique characteristics of 100).

selfishness can cause a rational agent to stray. This is because both agents believe that the other agent will also recognize the above and still pick a high number. They each believe the other agent will still pick a high number because the effect from the reward/punishment game seems too weak to force their respective choices down to an outcome that minimizes their respective payoffs.[82] As such there is some number, which marks a threshold, and the outcome of any number below it would be unacceptable.[83] So the argument from intuition concludes that each agent should choose some high number, but one that is lower than 100.

Note this argument from intuition does not recommend a particular choice or articulate which number, exactly, marks the threshold. It only recommends that an agent choose some high number. Also note that what, exactly, counts as a high number remains undefined. This is crucial to the argument since, if the threshold was specified, the dominance principle[84] could be applied. And nothing in the argument from intuition could stop a backward induction argument from forcing the recommend choice away from high numbers.

To understand why "high number" must be left undefined for the intuition argument, consider the following.[85] If an agent assumes that her opponent is playing a high number, should she choose a high number as well? Clearly yes, since the effect

[82] Basu, 1994: 392

[83] Unacceptable in the sense that, to pick a number below the threshold, would only guarantee that that agent would get more payoff than the other agent (but less payoff than if she had picked any number above the threshold).

[84] I.e. if, for an agent, one choice dominates another choice (either weakly or strongly), then that agent should choose that first choice (rather than any other choice).

[85] Note that Basu categorized the notion of leaving the set of acceptable number ill-defined as a possible solution, rather than part of the argument from intuition. However I have noted it here because I maintain it **is** part of the argument behind our intuition.

of the reward/punishment game gives no force to the dominance principle here. Nothing is given to the dominance principle because, ultimately, it is nonsensical to ask: if one agent is choosing some high number, should the other agent choose a high number minus one?[86]

In contrast the backward induction argument prescribes one particular choice, namely 2. Consider Basu's account of the backward induction argument here:

> At first sight, both players feel pleased that they can get 100 units each. To get this, each player simply has to write 100. But each player soon realizes that if the other player adheres to this plan then he can get 101 units by writing 99. But, of course, both players will do this, which means, that each player will in fact get 99 units. But if both were planning to write 99, then each player will reason that he can do better by writing 98; and so on.[87]

This pattern will continue until the only choice that remains for both agents in 2.[88] So the backward induction argument recommends 2 for both agents, despite the fact that *(2, 2)* will yield less for both agent then almost every other possible outcome.

When the conclusion from the backward induction argument and the conclusion from the intuition argument are combined, a formal contradiction is generated. This is clearest if stated explicitly.

[86] Basu, 1994: 394-395

[87] Basu, 1994: 392

[88] Implicit here is the idea that, for whatever number the opponent picks, the only acceptable choice for the other agent is that number **minus one** (with the exception of when the opponent picks 2, where 2 is the only acceptable choice). So picking the number chosen by the opponent **minus two** would be unacceptable (according to this argument).

Note first the argument from intuition:

(1) 100 is never the optimal choice.

(2) If 100 is never the optimal choice, then neither agent should choose 100.

(3) Therefore, neither agent should choose 100.

(4) The effect from the reward/punishment game is not strong enough to force the choices of both agents to an outcome that minimizes their total payoffs.

(5) If the effect from the reward/punishment game is not strong enough to force the choices of both agents to an outcome that minimizes their total payoffs, then both agents should choose a high number.

(6) Therefore, both agents should choose a high number.

(7) If (3) and (6), then both agents should choose a high number that is not 100.

(8) Therefore, both agents should choose a high number that is not 100.

(9) If both agents should choose a high number that is not 100, then both agents should not choose 2.

(10) Therefore, both agents should not choose 2.

And the backward induction argument:[89]

(11) If *i* should choose *r*, then *j* should choose *r-1*.

(12) If *j* should choose *r*, then *i* should choose *r -1*.

(13) If *i* should choose *r*, then *i* should not choose *r-2*.

(14) *i* should choose *r**. [assume for reductio]

(15) Therefore, *j* should choose *r* -1*.

(16) Therefore, *i* should choose *r* -2*.

(17) Therefore, it is not the case that *i* should choose *r*-2*.

(18) Therefore, *i* should choose *r*-2*, and it is not the case that *i* should choose *r* -2*.

(19) Therefore, it is not the case that *i* should choose *r**.

(20) If it is not the case that *i* should choose *r**, then *i* should choose 2.

(21) Therefore, *i* should choose 2.

(22) *j* should choose *r**. [assume for reductio]

(23) Therefore, *i* should choose *r* -1*.

(24) Therefore, *j* should choose *r* -2*.

(25) Therefore, it is not the case that *j* should choose *r* -2*.

(26) Therefore, *j* should choose *r* -2* and it is not the case that *j* should choose *r* -2*.

(27) Therefore, it is not the case that *j* should choose *r**.

(28) If it is not the case that *j* should choose *r**, then *j* should choose 2.

(29) Therefore, *j* should choose 2.

(30) If (11) and (19), then both agents should choose 2.

(31) Therefore, both agents should choose 2.

[89] For this argument, let *r* be a variable that ranges over the integers between 3 and 100. Also let *r** be any specific integer greater than 2.

When the conclusions from these two arguments are combined:

(32) Therefore, both agents should not choose 2 and both agents should choose 2.

As stated, we have a contradiction.

Solutions

I will now present three solutions that each attempt to diagnose a flaw in the above backward induction argument; they each argue why rational agents would reject the conclusion reached by the backward induction argument. As noted earlier, two of these solutions were originally created to address backward induction arguments in other problems. I present the first solution to demonstrate why the unique structure of the traveler's dilemma makes its backward induction argument impervious to certain solutions. The second solution will show that there remain certain solutions, originally raised against other backward induction argument problems, which can be adapted to address the backward induction argument in the traveler's dilemma. The third solution is original and so it will be evaluated at length.

Solutions – Adaptive Behavior

Unlike backward induction arguments for other problems, like the iterated prisoner's dilemma, the backward induction argument in the traveler's dilemma occurs entirely in one round of play. Games with multiple rounds enable agents to adapt based on new information. In repeated games the new information can influence the decisions each agent will make in subsequent rounds. This new information is the outcomes of any or all rounds that have been played. For instance, in the second round of an iterated prisoner's dilemma the agents know the outcome of the first round and they can use this information to refine their beliefs before they each decide what to do in the second round.

However this adaptive behavior is not possible in one-round games because the agents must make their decisions with nothing more than the information they have at the outset of the game. This is important because, as a result of having only one round, solutions to backward induction arguments that rely on a game structure having multiple rounds – a common feature of many game theory problems that involve a backward induction argument – are not applicable.

To understand why such solutions fail against the traveler's dilemma, consider the following adaptive behavior solution.

Roy Sorensen argued that the role of reputation building is the key to solving the iterated prisoner's dilemma[90]. He posited that an agent in an iterated prisoner's dilemma can use the tacit communication inherent to the game to influence the beliefs of the other agent. In particular an agent could undermine the confidence the other agent has about whether or not the first agent is rationally self-interested. For instance, a self-interested agent could intentionally damage her reputation as a rationally self-interested agent by making choices that are not rationally self-interested (e.g. choosing to not confess). By doing so in, say, the first round she can increase the other agent's confidence that the first agent will not confess in the third round if and only if she herself does not confess in the second round. While this would be to the detriment of the first agent in the first round, influencing the beliefs of the other agent in this way

[90] Iterated prisoner's dilemmas that are at least three or more rounds long.

could prove beneficial for her in later rounds of the game and yield her a greater total payoff than she might otherwise receive.[91] So for the iterated prisoner's dilemma this solution suggests that the knowledge of the agents can be expanded insofar as they can each learn more about their opponent and use that knowledge as grounds for choosing to not confess at least some of the time.

Observe how this solution relies on multiple rounds, or rather the inherent tacit communication that is available in a repeated game. According to this solution the agents use that tacit communication to convey information between each other; information that they would otherwise lack (e.g. willingness to engage in a mutually advantageous strategy). Because this tacit communication is only present in games with multiple rounds, this solution, and others like it, cannot be successfully adapted to the traveler's dilemma. The traveler's dilemma, as a one-round game, lacks the tacit communication that exists in repeated games. Without that mechanism neither agent can influence the way in which they are perceived by the other agent. And so, in virtue of the unique structure of the traveler's dilemma, such solutions to backward induction arguments cannot be successfully adapted to the traveler's dilemma. Similarly, any other backward induction argument solution that makes use of tacit communication, or any other feature that comes with repeated games, would be inapplicable to the traveler's dilemma.

[91] Sorensen, 1985: 162-166

While the traveler’s dilemma is not open to such solutions, it remains vulnerable to other kinds of solutions. The next solution is one that can be successfully adapted to the traveler’s dilemma despite its unique structure.

Solutions – Iterations of Knowledge

Timothy Williamson argued that each additional stage in a backward induction argument invokes an additional iteration of the knowledge held by the agents. As Williamson noted the backward induction argument for the iterated prisoner’s dilemma concludes that a rational agent would never not confess since, not confessing in round *z* might occur only if it is not common knowledge that there is no round after *z* at which it is rational to not confess.[92] The flaw in the backward induction argument, he argues, is that with every additional stage in the argument the agents must add another “knows that” to that which they know. For example, when deciding to confess in round *z*, an agent knows that her opponent should confess in *z*. And then, when she decides to confess in *z-1*, she knows that her opponent should confess in *z-1* and also knows they should both confess in *z*. When that agent then decides to confess in *z-2* she knows that her opponent should confess in *z-2* and that her opponent also knows that they should both confess in *z-1*, and knows that they should both confess in z. And so on until she decides to confess in the first round of the game because her opponent will confess in that round (and all subsequent rounds).

[92] Let *z* represent the last round of the game.

Because any agent we can conceive of has imperfectly accurate epistemic capacities, an agent's knowledge of her opponent's knowledge is less exact than her knowledge of her own knowledge. Williamson concludes that an agent's knowledge about what her opponent knows becomes eroded after a lengthy number of iterations (e.g. in an iterated prisoner's dilemma that will last one hundred rounds). When such erosion occurs her knowledge becomes unreliable, at which point the backward induction argument cannot continue. Simply put, no sufficiently lengthy iteration of knowledge can be considered reliable. So, in an iterated prisoner's dilemma of sufficient length, the backward induction argument will be unable to persuade the agents (at the outset of the game) to confess in every round.[93]

Note that while Williamson's original argument was designed as a solution to the surprise exam problem, he argued that his solution is applicable to the iterated prisoner's dilemma (as well as other backward induction argument problems). While using the iterated prisoner's dilemma for context in this presentation of his objection, the way in which it can be applied to the traveler's dilemma should be sufficiently clear.

Because the backward induction argument in the traveler's dilemma is of sufficient length to be considered a game that involves a lengthy iteration of knowledge, Williamson's solution can be applied. Here the effect would be that the agents will not reach the conclusion of the backward induction argument – they will not conclude that choosing 2 is the rationally prescribed course of action. Observe

[93] Williamson, 1992: 230-232

how this solution does not require that the structure of a game has multiple rounds; it only requires that backward induction arguments repeatedly successively add the same pattern in each stage of its argument – a quality that the backward induction argument of the traveler's dilemma shares with other backward induction arguments.

Note, though, that there are concerns with this solution that are worthy of attention. For instance it is unclear which specific premise, in backward induction arguments, this solution rejects. Instead of attacking a specific premise this solution rejects the cumulative effect of backward induction arguments – the successively growing iterations of knowledge. And, the merits of Williamson's solution notwithstanding, a solution that specifies which premise must be denied is preferable. Thus, I will set this solution aside.

These two solutions were raised in order to demonstrate that some solutions, previously argued against other backward induction argument problems, can be adapted for the traveler's dilemma while others cannot.

Solutions – Suboptimal Payoffs

Before I can present the third solution, I must go on a brief, but crucial, tangent. I must articulate the critical differences between the traveler's dilemma and the prisoner's dilemma.

Solutions – Suboptimal Payoffs - Key Differences

Unlike the prisoner's dilemma, the traveler's dilemma does not have a single dominant choice.[94] Only when restricting consideration to two numbers does one choice dominate another. Specifically, every number that the agents could choose is dominated by the highest number that is less than it – 100 is dominated by 99, while 99 is dominated by 98, and so on.[95]

If an agent considers three or more choices, for example 99, 98, and 97, there is no dominant choice. To understand why, observe that if *i* were to pick one of those three numbers, and *j* was certain that *i* was going to pick one of them (but was unsure of which one in particular), *j* is not always best off picking the lowest number in that set or even the number just outside the bottom of the set (i.e. for this example, 96). This is because, if *i* had picked 99 and *j* pick either 97 or 96, *j* would receive one or two fewer units of utility then if she had pick 98 or even 99.

In the traveler's dilemma, the backward induction argument artificially narrows the range of choices that an agent considers to only two neighboring numbers (e.g. 99 and 98, or 66 and 65). Because the backward induction argument depends on an application of the dominance principle, the range of choices for each agent in the traveler's dilemma needs to be restricted in this fashion. As shown above, the backward induction argument supposes that one agent will make a specific choice and forces the other agent to determine which would be best: that same specific choice or a specific alternative (i.e. the next lowest number). This is in contrast to the

[94] Sarangi, 2000: 28

[95] With the exception of 2 since there is no number lower than it that either agent could choose.

iterated prisoner’s dilemma where no such artificial restriction is required. In that game, every round offers only two choices for each agent.

With that in mind I can now present the third solution and my evaluation of it.

Solutions – Suboptimal Payoffs – Argument

Given the way the traveler's dilemma is construed[96], if an agent chooses 2, or any integer between it and 96, she has made a mistake. Her mistake will occur in one of two ways, she will either have: (a) failed to realize that making such a choice will yield her a suboptimal payoff, or (b) assumed that the other agent will commit (a)[97]. Recall that our agents are taken to be ones that never make mistakes. So if (as I will argue) the backward induction argument requires that the agents make a mistake, it misconstrues the agents as other than what they actually are assumed to be.

What do I mean when I speak of "suboptimal payoffs"? Simply:

> Suboptimal payoffs: *x* is a suboptimal payoff for *i* if and only if *x* is the payoff for *i* from an outcome at which *i* either **won or tied** in the reward/punishment game, but that payoff is less than a payoff *i* **could** have received from some other outcome whereat *i* would have **lost** the reward/punishment game.[98]

With that in mind, I will introduce the optimality principle:

> Optimality principle: A rational agent not will make a choice that will guarantee her a suboptimal payoff.

The optimality principle is simply an extension of the underlying assumptions that the agents are both rational and primarily interested in getting the most compensation possible.

So, why, specifically, would choosing a suboptimal payoff yielding number be a mistake? Simply because if a rational agent picks 2, or any other suboptimal

[96] Specifically that the choices available to the agents are restricted to integers between 2 and 100 inclusively.

[97] I.e. Mistakenly misconstrue the other agent as someone who will make a mistake.

[98] Note that this usage of suboptimal differs from its typical usage in game theory.

yielding choice, she must have failed to realize that such a choice will guarantee her a suboptimal payoff.[99] The nature of the compensation game (i.e. that the lowest number picked between the two agents becomes a benchmark from which the payoffs are determined) makes it the case that, by picking 2 (for instance), there is no way in which her payoff could be higher than four. A failure to recognize the maximum payoff that a choice can yield cannot be considered anything other than a mistake since each agent knows that both agents are rational, interested in maximizing their own compensation, and are cognizant of the game structure.

For clarity, consider the payoff *i* would receive from the pair of choices *(2, 2)* (where the first number represents the choice of *i* and the second that of *j*). Here her payoff is suboptimal in virtue of her potential payoff from the pair of choices *(98, 97)*. The first set of choices is suboptimal because, from it, *i* would only receive two units of utility whereas the second set of choices would yield her ninety-five. Similarly 96 is a suboptimal choice for *i* because, if she tied in the reward/punishment game with this choice she would only receive a payoff of ninety-six. Whereas there are better possible outcomes for her that could have obtained, had she chosen 100 and lost the reward/punishment game.[100]

Having said that, I will now explain why it would be a mistake to assume that the other agent will make a choice that will necessarily yield her a suboptimal payoff.

[99] Unless she realized that her choice was sub-optimal but picked it because she believed her opponent was also picking a sub-optimal yielding choice (i.e. she was forced into picking a suboptimal yielding choice). However in this situation the agent has misconceptualized her opponent as someone who might make a mistake.

[100] Whereat she would have received a payoff of ninety-seven.

Simply, if *i* believes *j* will pick 2, or any other specific suboptimal yielding choice, then *i* must also believe that *j* will not act in accordance with the optimality principle. That is, *i* would need to believe that *j* will knowingly make a choice that will guarantee *j* suboptimal payoff – a belief that is incompatible with the assumption that both agents are interested in maximizing their own respective payoffs. Because an agent who chooses a suboptimal yielding choice failed to realize that that choice was a suboptimal yielding one, an agent who assumes her opponent will make a suboptimal yielding choice must also assume that that opponent will fail to realize that that suboptimal yielding choice is actually a suboptimal yielding one.

As noted earlier, neither agent is primarily concerned with getting more payoff than the other agent – they are each only interested in maximizing their own compensation generally. Only in virtue of the reward/punishment game will the agents attempt to gain more individually at the expense of the other agent. But even when considering the effect of the reward/punishment game, it would be a mistake for an agent to consider any choice that would guarantee her a suboptimal payoff. This is because saying that the agents are predisposed to avoid suboptimal payoffs is just a rephrasing of the primary aim of the agents – to receive more, rather than less, compensation. The effect of supposing that *j* will pick a suboptimal yielding choice is inconsistent with the assumption that the agents are primarily seeking more compensation generally, rather than just more than their opponent.

Since the backward induction argument assumes that the other agent has decided to make a suboptimal choice (i.e. she will choose 2) when it prescribes 2 as the rational choice for the agent considering the argument, an agent who adheres to the conclusion of the backward induction argument makes the mistake of assuming that the other agent will make a mistake. In short, the backward induction argument requires that the agents assume that their respective opponents will each fail to realize that choosing 2 will yield them each a suboptimal payoff.[101]

But, the reader should note that **if** *i* had good reason to believe that *j* was going to pick 2 (or even 3) it would be irrational for *i* to pick a number other than 2. Even though picking 2 in that situation would be rational, it would still be suboptimal. And in virtue of the game structure (i.e. no communication, simultaneous play) *i* **never could** have sufficiently good reason for believing that *j* will make a (suboptimal) choice, to which 2 could be considered the rational choice for *i*. More broadly, neither agent could ever have sufficiently good reason to believe that picking a suboptimal yielding choice is in her best interest.

[101] Note that this solution does **not** say that the assumption in the backward induction argument about what the other agent will do is a mistake. Such a move would be nonsensical since the backward induction argument contains *reductio ad absurdum* arguments and assuming the negation of what one seeks to establish is part of any *reductio*. Instead this solution argues that the **conclusion** reached by the backward induction argument is inconsistent with the assumption that the agents will realize that picking a suboptimal yielding number would be a mistake. Regardless of *reductio ad absurdum* arguments, to succeed the backward induction argument cannot succeed by dictating a course of action that a rational agent (who never makes mistakes) would deem to be a mistake in virtue of the maximum payoff for all the possible outcomes of that choice could yield (compared to other potential payoffs from other choices).

This solution can be explicitly stated in the following way:

(i) The backward induction argument concludes that an agent will pick 2.
(ii) If the backward induction argument concludes that an agent will pick 2, then an agent will pick 2.
(iii) If an agent picks 2, then an agent did not realize that 2 will guarantee her a suboptimal payoff.
(iv) If an agent did not realize that 2 will guarantee her a suboptimal payoff, then an agent will make a mistake.
(v) Therefore, if the backward induction argument concludes that an agent will pick 2, then an agent will make a mistake.
(vi) If (v), then the backward induction argument misconstrues the agents as ones who make mistakes.
(vii) Therefore, the backward induction argument misconstrues the agents as ones who make mistakes.
(viii) If the backward induction argument misconstrues the agents as ones who make mistakes, then it is not the case that both agents should choose 2.
(ix) Therefore, it is not the case that both agents should choose 2.

Note that the above solution dictates a specific set of choices that the agents can consider optimal: 100, 99, 98, and 97.[102] These choices are optimal because, if an agent chooses any one of them and either ties or wins the reward/punishment game, she has not guaranteed herself a suboptimal payoff. That is, with any one of these choice it could not have been the case that she could have received a greater payoff had she chosen some other number and lost the reward/punishment game. 96 falls just outside the set of optimal choices because, as stated above, if an agent chooses 96 and ties in the reward/punishment game her payoff would be ninety-six. She could

[102] Note also that while the conclusion of this argument is the negation of (31) in the explicit presentation of the contradiction as presented above (i.e. the contradiction generated by the conjunction of the conclusion of the intuition argument and the conclusion of the backward induction argument), premise (13) (from the backward induction argument of that same presentation) is also implicitly rejected by this solution. This premise is implicitly rejected since, according to the solution, an agent who chooses the number selected by her opponent minus two has not necessarily made an unacceptable choice (i.e. so long as her choice is an optimal one).

have received a better payoff had she chosen 100 and lost the reward/punishment game (i.e. the pair of choices *(100, 99)* would yield her a payoff of ninety-seven).

Solutions – Suboptimal Payoffs – Objections and Replies

One might object that, because the solution implies a specific set of choices, the backward induction argument still applies. That is, the backward induction argument would end on the lowest number in the set of optimal yielding choices, or even continue and conclude that the agent should choose the next number (i.e. 96, since 96 weakly dominates 97). But, then, the backward induction argument appears capable for continuing down to 2. In short, since the solution fails to prescribe a specific choice that would always be rational, the backward induction argument can still be applied - either to just the numbers in the set or all the way down to 2.

This objection has two parts: First, that a backward induction argument can be applied to the numbers that make up the optimal set. And, second, if the first part is true, the solution fails to stop the backward induction argument at all (insofar as 2 would remain the prescribed rational choice).

A proponent of this solution must respond to this objection by first addressing the second part of the objection (then turn to address the first part).

Regarding the second part of the objection, if a backward induction argument prescribes 97 as the rational choice, it is not the case that another stage in the argument can be applied (or at least not the kind of stage the objection suggests). To understand why, consider the following. If a backward induction argument concludes that 97 is the rational choice, it cannot then continue and prescribed 96 as the rational choice (or continue further and prescribe any other lower number) because both agents know, through their respective introspective backward induction arguments, that **if** they were to continue the backward induction argument and, for instance, both agents would end up with an outcome neither would be satisfied with. For example if they ended with the pair of choices *(92, 93)*, *i* would get ninety-four while *j* would get ninety. Neither agent would be content with this outcome.

j would be dissatisfied with the outcome (and thereby reject, from consideration, her choice that resulted in it) because she lost the reward/punishment game. *i* would also reject her choice that resulted in this outcome since her payoff from it would be suboptimal. This outcome yields a suboptimal payoff for *i* since, even though she won the reward/punishment game, she remains worse off than had she ceased her backward induction argument at the *(100, 99)* stage (even though she would have lost the reward/punishment game). Had she pursued this other outcome she would have received a payoff of ninety-seven. And since an agent prefers a payoff of ninety-seven over ninety-four, she should have chosen the ninety-seven yielding choice instead of the ninety-four yielding choice (or any other option that yields a worse outcome).

Both agents would realize that continuing the backward induction argument beyond the set of optimal choices is against their own respective best interest. They would also both realize that it would be against the best interest of the other agent for her to continue **her** own backward induction argument because continuing it beyond the set of optimal yielding choices would be against **her** own best interest. As a result, both agents would abandon the continuance of the backward induction argument beyond the set of optimal choices.

Note that the above response to the second part of the objection is tenable because, in the traveler's dilemma, it can be in an agent's best interest to reconsider choices that were previously dismissed. What makes the choices worth reconsidering is the information each agent discovers at later stages of the backward induction argument. The agents can discover what would occur if she chose a suboptimal payoff yielding choice, and, as a result of that discovery, reconsider the optimal choices.

Having established why an agent would not continue the backward induction argument beyond the set of optimal choices, I will now explain why the first part of the objection can also be denied. As just noted, it can be in an agent's best interest to reconsider previously dismissed choices. So if an agent is considering a backward induction argument, after realizing that it should not be continued beyond the set of optimal yielding choices, the argument would seem to start over again in the set of optimal yielding choices (i.e. the agents would reconsider the choice pair where one chooses 100 and the other 99). But then the one that would have chosen 100 would determine that she should consider 99 instead of 100. Then she would realize that she should consider 98, but then so should the other agent. And so on.

However when the backward induction argument reaches the bottom of the set of optimal choices, it would loop back to the higher numbers in the set of optimal choices (just as it did initially). In short, to try to apply a backward induction argument on top of the optimality principle is fruitless as the argument continues in a continuous loop. The backward induction argument would never reach an end point or stop at a single choice to prescribe as the best course of action. So while one can try to apply a backward induction argument to the set of optimal choices, because it would fail to prescribe a course of action, doing so would be futile.

Moving on to a second objection, one might argue that this solution prescribes 100 as an acceptable choice. And that this generates a contradiction with intuition since the argument from intuition concludes that 100 is not an acceptable choice.

This objection simply misunderstands the solution. The solution does not actively suggest which choice a rational agent should make. Instead it argues why a rational agent would **not** make certain choices (specifically, ones that would yield a suboptimal payoff). The solution eliminates some choices from consideration but leaves the specific decision of which choice to make unanswered. This solution only argues which choices would be irrational.[103]

Does this entail that the solution is incomplete? No, because the aim of the solution was to provide an escape from the contradiction generated by the conjunction of the intuition argument conclusion and the backward induction argument conclusion. This solution does so by offering an account of why a rational agent would not adhere to the conclusion reached by the backward induction argument.

[103] One still requires some other argument to dictate a specific course of action for a rational agent.

Solutions – Suboptimal Payoffs – Final Remarks

Despite my discussion and defense of this novel solution it seems evident that there may be other promising objections to it. However I have considered the important objections that I believe are the strongest possible. Regardless, the prospect of other possible objections should not be considered a roadblock for what I intended to accomplish in this thesis. Even if this third solution is problematic beyond salvation my aim here was to demonstrate the fact that the traveler's dilemma, in virtue of its unique structure, closes the door on some backward induction solutions[104], while also opening other doors that house new and unique solutions.

[104] E.g. ones reliant on a game structure having multiple rounds.

5: **Summary**

In this chapter I will briefly summarize this thesis and make some last peripheral remarks.

I began by explaining the key terms and how they are used in the analysis of the traveler's dilemma, the prisoner's dilemma, and the iterated prisoner's dilemma. I then prefaced my discussion of the traveler's dilemma by presenting the prisoner's dilemma and the iterated prisoner's dilemma. I did so to facilitate my discussion of the traveler's dilemma and its the backward induction argument. This was a worthwhile approach because the traveler's dilemma is related to the prisoner's dilemma. Additionally, it was worthwhile because my presentation of the backward induction argument for the iterated prisoner's dilemma allowed me to note some important general characteristics of all backward induction arguments before discussing the traveler's dilemma and its backward induction argument.

While discussing the traveler's dilemma, I clarified its essential qualities and explained how the course of action prescribed by intuition and the course of action prescribed by the backward induction argument are incompatible. In my discussion of the traveler's dilemma I have critically examined three solutions that sought to diagnose a flaw in its backward induction argument. These solutions, and why they were raised, are summarized below.

While Basu created the traveler's dilemma to emphasize an important characteristic of backward induction arguments - that they do not require that the games have multiple rounds to operate – The first solution I raised demonstrated this, insofar as it relied on backward induction arguments having multiple rounds so that the agents could engage in adaptive behavior. The presentation of this solution elucidated the importance of the traveler's dilemma. The second solution I raised argued that no sufficiently lengthy iteration of knowledge can be true. While originally argued for other backward induction arguments, this solution can be successively adapted to the traveler's dilemma. note that I raised this solution to show that, despite the unique structure of the traveler's dilemma, it remains vulnerable to some solutions raised against other backward induction argument problems. I then offered a novel solution to the traveler's dilemma, one that relies on the unique structure of the game. I explicated this solution in full and offered a few possible objections to it as well as replies to those objections. As explained in the fourth chapter, this solution was raised to demonstrate that the unique structure of the traveler's dilemma creates unique ways of addressing its backward induction argument.

In closing, the reader should understand that discussions about problems generated by backward induction arguments are important because they highlight how those arguments can conflict with intuition. The traveler's dilemma is also an example of a game where non-Nash play is intuitively the rational course of action. Ultimately, because I gave an account for why agents should engage in non-Nash play in the traveler's dilemma, one could say that I am agreeing with Basu's conclusion that the traveler's dilemma gives reason to revise typical game theory assumptions.[105]

Having said that, I believe it would be worthwhile to examine whether or not the optimality principle can be successfully adapted to other game-theoretic backward induction argument problems (e.g. the iterated prisoner's dilemma), or even more broadly to other game theory problems that do not contain backward induction arguments (e.g. the prisoner's dilemma). At least initially, the answer seems that it cannot be. This is because the optimality principle is grounded in the unique structure of the traveler's dilemma. The structure of the iterated prisoner's dilemma, for instance, differs in such a way from the structure of the traveler's dilemma that it would never be the case that an agent ought to reconsider earlier stages of the backward induction argument in light of purely introspective information she would discover at later stages of the backward induction argument.

Moreover, it is unclear how one ought to apply the optimality principle in games with multiple rounds. Should it be applied at every round of the game? Or just with regard to the total possible payoff? Either way its recommendation seems to depend on the specific payoffs structure of the particular game being played (i.e. the specific utilities attributed to the different potential outcomes by the agents involved). This requirement seems contrary to the essence of games like the iterated prisoner's

[105] Basu, 2007; 95

dilemma since, at least traditionally, the preference rankings (rather than the specific utility functions) are considered to be of primary importance.

But these are only preliminary remarks; further study would be required to definitively determine whether or not the optimality principle can be successfully generalized and applied to other games.

References

Basu, K. (1990) “On The Nonexistence of a Rationality Definition for Extensive Games”, *International Journal of Game Theory*, Vol. 19, pp. 33-44.

Basu, K. (1994) “The Traveler’s Dilemma: Paradoxes of Rationality in Game Theory”, *American Economic Review*. Vol. 84, No. 2, pp. 391-395.

Basu, K. (2007) “The Traveler’s Dilemma”, *Scientific American*, Vol. 296, No. 6, June, pp 90-95.

Binmore, K. (1990a) “Aims and Scope of Game Theory”, in *Essays on the Foundations of Game Theory*, Cambridge, USA: Basil Blackwell, pp. 1-42.

Binmore, K. (1990b) “Nash Equilibrium”, in *Essays on the Foundations of Game Theory*, Cambridge, USA: Basil Blackwell, pp. 43-77.

Binmore, K. and Brandenburger, A. (1990c) “Common Knowledge and Game Theory”, in Binmore, K. (ed.) *Essays on the Foundations of Game Theory*, Cambridge, USA: Basil Blackwell, pp. 105-150.

Campbell, R. (1985) “Background for the Uninitiated”, in Campbell, R. and Sowden, L. (ed.) *Paradoxes of Rationality and Cooperation*, Vancouver: UBC Press, pp. 3-44.

Chapman, J.M. and Butler, R.J. (1965) “On Quine’s ‘So-Called Paradox’”, *Mind*, Vol. 74, No. 295, pp. 424-425.

Davis, M. (1983) *Game Theory: A Nontechnical Introduction*, New York: Basic Books.

Goeree, J. and Holt, C. (2005) “Ten Little Treasures of Game Theory and Ten Intuitive Contradictions”, *The American Economic Review*, Vol. 91, December, pp. 1402-1422.

Luce, D. and Raiffa, H. (1989) *Games and Decisions,* New York: Dover Publications.

Olin, D. (2003) *Paradox*, Montreal: McGill-Queen's University Press.

Pettit, P. and Sugden, R. (1989) “The Backward Induction Paradox”, *The Journal of Philosophy*, Vol. 86, No. 4, pp. 169-182.

Rapoport, A. (1960) *Fights, Games, and Debates,* Ann Harbor: University of Michigan Press.

Sarangi, S. (2000) *Exploring Payoffs and Beliefs in Game Theory*, Virginia Polytechnic Institute and Virginia State University. Blacksburg, Virginia: PhD (Economics) dissertation.

Sorensen, R. (1985) "The Iterated Versions of Newcomb's Problem and the Prisoner's Dilemma", *Synthese,* Vol. 63, pp. 162-166.

Sorensen, R. (1999) "Infinite "Backward" Induction Arguments", *Pacific Philosophical Quarterly,* Vol. 80, pp. 278-283.

Williamson, T. (1992) "Inexact Knowledge", *Mind*, Vol. 101, No. 402, April, pp. 217-242.