Albert, Max; Heiner, Ronald Asher

Working Paper

An Indirect-Evolution Approach to Newcomb's Problem

CSLE Discussion Paper, No. 2001-01

Provided in Cooperation with:
Saarland University, CSLE - Center for the Study of Law and Economics

An Indirect–Evolution Approach to Newcomb's Problem

Max Albert & Ronald A. Heiner*
Center for the Study of Law and Economics
Discussion Paper 2001-01

Players from two populations, predictors and predictees, are randomly matched in a game–theoretic version of Newcomb's Problem. Predictors are able to predict the predictees' choices by observing their type. There are two types of predictees, those who take their predictability into account by using the Backtracking Principle when calculating expected utilities, and those who ignore their predictability by using the Disconnection Principle. Backtrackers are one–boxers, the others are two–boxers. Given predictability, evolution favors the Backtracking Principle. An explicit causal analysis proves that this result does not rest on unusual causal assumptions.

*Albert: University of Koblenz–Landau, Dept of Economics, August–Croissant–Str. 5, D–76829 Landau, Germany, email: albert@uni-landau.de. Heiner: *James Buchanan Center for Political Economy,* MSN 1D3 Carow Hall, George Mason University, Fairfax/VA 22030, U.S.A., email: rheiner@mason2.gmu.edu. For helpful comments and discussions, we are grateful to Alexander Neunzig, Dieter Schmidtchen, and participants, especially Martin Dufwenberg, Werner Güth, Hartmut Kliemt, and Robin Pope, of a seminar of the research group *Making Choices* at the *Zentrum für interdisziplinäre Forschung (ZiF),* Bielefeld, in August 2000.

1 Introduction

1.1 Three Views on Newcomb's Problem

The present paper uses the indirect–evolution approach to shed new light on Newcomb's Problem (NP). In contrast to the standard version of evolutionary game theory, the indirect–evolution approach assumes that players are rational. The evolutionary process affects the determinants of rational choices. The approach has been used to argue that altruistic or moral preferences may develop as a means of achieving cooperation.[1] We connect this line of thought with the literature on foundational problems of rationality, as it developed in response to NP.[2]

NP is usually introduced in the form of a story similar to the following one: Eve has two boxes for Adam to open. The first box is opaque; its contents are not visible. The second box is transparent; it contains one thousand dollars. Eve always gives Adam the opaque box. However, before opening it and taking its content, Adam must decide whether to additionally open the transparent box and take the thousand dollars (be *greedy,* g), or not doing so, foregoing the thousand dollars (*resist* temptation, r). His utility function is strictly increasing in money. When Adam is called upon to decide, Eve has already made her choice. She has put a million dollars in the opaque box (million, m) if and only if she predicted Adam will resist the temptation and reject the transparent box with the one–thousand–dollar note; otherwise, she left the opaque box empty (nothing, n). Now comes the hitch: Eve can perfectly predict Adam's choice (by, for example, making a brain scan of Adam before she decides what to put into the opaque box). Adam knows that his upcoming decision is thereby perfectly predictable to Eve. Should he nevertheless take both boxes?

The term "brain scan" indicates that Eve does not predict Adam by putting herself into his shoes. Rather, she observes an event (the result of the brain scan) that is causally linked to Adam's later decision; which thereby tells her, all by itself, what Adam is going to do. In the metaphorical language of game theory, Eve receives a signal sent by nature—that is, neither voluntarily sent nor observed by Adam—informing her about Adam's later action before it is actually taken. Thus, NP postulates a well–known causal structure sometimes called *Reichenbach's fork.* Adam's and Eve's actions

[1] Cf. Frank (1987) and Güth & Kliemt (1994). For a general characterization of the indirect–evolution approach, see Güth & Kliemt (1998).

[2] NP was invented by a physicist, William Newcomb, and first published by Nozick (1969). Since then, it has been widely discussed, see, e. g., Nozick (1993) and Binmore (1994: section 3.5 and elsewhere).

have a *common cause,* namely, the state of Adam's brain (or mind, the distinction does not matter here) at the time of the brain scan. From Adam's earlier brain state there is a causal chain leading to his later action. Moreover, a second causal chain forks off from Adam's brain state and leading to Eve's action, which occurs earlier than Adam's. The question is: Should the knowledge about this causal structure have an influence on Adam's decision?

Three basic positions are taken in the literature. Two of them propose solutions to NP as it stands, the third rejects the causal structure assumed in NP.

According to the first solution (two–box solution, (n, g)), Adam should take both boxes since this action is *dominant:* At the time Adam decides, the opaque box contains either one million or nothing. Whatever is in the opaque box, taking both boxes means one thousand dollars more.

This consideration focuses on the fact that Adam's action has no causal influence on Eve's action since Eve acts earlier than Adam. According to an important but rarely stated principle of decision theory, any state or event that is not caused by the player's actions should be treated as a "state of nature" by this player.[3] "States of nature" must be assigned a fixed subjective probability by a player, independent of the player's pending action which has not yet been actually chosen. A number of writers have argued in favor of this principle, which we call the *Disconnection Principle* because it requires a player (like Adam in NP) to ignore information about causal links between past events (like the result of Eve's brain scan and her corresponding action) and his choices, as if a player's pending decision were causally disconnected from all past events.

If Adam accepts the Disconnection Principle, the only assignment of probabilities consistent with the assumptions of NP is a probability of 0 for Eve having put money into the opaque box: Whatever fixed subjective probability Adam assigns to the outcome of the brain scan and Eve's corresponding action, he finds that taking both boxes maximizes his expected payoff; hence, he chooses g. Since, however, Adam knows that Eve can perfectly predict him on the basis of the brain scan, he deduces that Eve must have already predicted that he will end up choosing g. So Adam deduces that Eve did not choose m but n. Any other probability assignments violate the central assumption of NP, namely, that Eve in fact has physical means to perfectly predict Adam, *and that Adam knows and hence believes it.*

Thus, the Disconnection Principle does not enforce an assignment of prob-

[3]Of course, the player's actions need not be causally sufficient for bringing about a state or event in order for it not to be counted among the "states of nature". It suffices if their are conditions under which the player's actions make a difference to the realization of the state or event under consideration.

abilities to events contradicting the assumptions of NP. This is because, when Adam computes his expected utilities, he uses the probabilities for Eve's actions conditional on the fact (known to him) that he accepts the Disconnection Principle. Even an observer not accepting the Disconnection Principle agrees that the probability of Eve having put a million into the box conditional on Adam accepting the Disconnection Principle is 0.

The Disconnection Principle is not formally implied by Savage's (1954) approach to decision theory (henceforth, Bayesianism), because this approach does not consider issues of causality.[4] Hence, a coherent Bayesian analysis is possible whether the representation of the decision problem is consistent with the Disconnection Principle or not. Nevertheless, the principle has been defended by some writers as a necessary supplement of Bayesian decision theory, and although it is usually not spelled out, it is never violated in standard game theory.[5]

The second solution to NP (one–box solution, (m, r)) rejects the idea that Adam should use fixed subjective probabilities of Eve's actions when calculating his expected payoffs. Instead, Adam uses the probability of Eve having put a million into the opaque box conditional on himself taking two boxes. This probability can be determined by *backtracking* from the action of taking two boxes along the causal chain to the corresponding state of his brain at the time of the brain scan. Taking two boxes implies that there was a certain outcome of the brain scan in the past. In a second step, Adam can go forward from his earlier brain state and make a postdiction of Eve's action conditional on his taking two boxes. The result is, of course, that the probability of Eve having put a million into the box conditional on Adam taking two boxes is 0. Likewise, the probability of her having put a million into the box conditional on him taking only one box is 1. If Adam uses the probabilities resulting from backtracking, he maximizes his expected payoff by taking only one box.[6]

We refer to the principle of assigning probabilities to events by backtracking as the *Backtracking Principle.* The Backtracking Principle can be

[4]However, the Disconnection Principle is formally implied by Aumann's (1987) game–theoretic definition of Bayesian rationality, cf. fn 24 on p. 23 below.

[5]The combination of Bayesianism with the Disconnection Principle is sometimes called causal decision theory, cf. Gibbard & Harper (1978). Pearl (2000) makes the principle the basis of his causal approach. We avoid the name causal decision theory since we think that the explicit consideration of causal connections does not necessarily imply acceptance of the Disconnection Principle.

[6]There is a variant of decision theory, called evidential decision theory, where one's own action serves as evidence for other events; cf. Gibbard & Harper (1978). We avoid the name evidential decision theory, however, because in contrast to the present approach, evidential decision theory is not explicitly based on a causal analysis.

substituted for the Disconnection Principle in a Bayesian analysis of the problem. It has always been recognized in the literature that the one–box solution can be presented in a Bayesian way: just introduce "Eve predicts Adam correctly" as one "state of nature" and "Eve does not predict Adam correctly" as the alternative state; since the description of NP implies that Adam should assign probability 1 to the first state, taking one box maximizes his expected payoff (cf. Nozick 1969, Bar–Hillel & Margalit 1972). In a sense, this analysis is too simple since it does not make clear how "Eve predicts Adam correctly" can be true independently of Adam's action. However, supplying an explicit causal analysis fills in the gap and provides the basis for a Savage–type approach based on the Backtracking Principle (see Appendix A).

The Backtracking Principle and the Disconnection Principle do not rest on different factual beliefs. Defenders of either principle must agree that the probability of Adam finding a million in the box conditional on Adam using the Disconnection Principle (and taking two boxes) is 0, while the probability of the same event conditional on Adam using the Backtracking Principle (and taking only one box) is 1. This follows deductively from the assumptions of NP. Rather, the debate between the proponents of these principles concerns the question whether it is right to use probabilities for Eve's action conditional on Adam's action (as recommended by the Backtracking Principle), or probabilities for Eve's action not conditional on Adam's action (as recommended by the Disconnection Principle).

The information given in NP unambiguously determines the numerical values of these and other probabilities, and therefore the beliefs about these probabilities held by Adam and Eve. The problem is to select the right kind of probability for computing expected utilities. The selection depends on not the factual but the "normative" beliefs or, better, the decision criteria: How should the decision be made given the information?

A third view rejects the causal structure assumed in NP. It comes with a slightly different version of NP, where Adam does not *know* that he is predictable but has overwhelming evidence of Eve's predictive abilities. For instance, let there be a perfect correlation between Eve's choices and the choices of Adam's predecessors in a long series of NP situations. The third view rests on what we call the Unpredictability Principle: Despite the evidence, Adam should assign fixed subjective probabilities to Eve's actions because it is *impossible,* for some reason or other, for there to exist a causal link between Adam's choices and past events (like the result of the brain scan) that can be used by Eve to predict his choice. Of course, this third view also leads to the two–box solution.

Hence, two–boxers come in two variants, and it is not always clear who

is who. There are those who (at least, seem to) accept predictability and argue that, *nevertheless,* Eve's action should be a "state of nature" for Adam. Alternatively, there are those who reject predictability and argue that, *therefore,* Eve's action should be a "state of nature" for Adam. The first variant relies on a *decision criterion* to motivate the two–box solution: According to them, Adam should use the Disconnection Principle to determine the probabilities relevant to his expected–utility calculations, while accepting that he is predictable to Eve. The second variant relies on a different *factual belief* about decisions being inherently unpredictable to motivate the two–box solution. That is, the second variant recommends that Adam should invoke the Unpredictability Principle when forming his probability beliefs. And this principle implies that, no matter what evidence might exist to the contrary, Adam is not predictable by a signal like the one provided by the brain scan.

1.2 The Structure of Our Argument

In our view, the discussion between one–boxers and two–boxers has reached a stalemate. How can the issue be possibly resolved? We contribute two, as we believe, decisive points:

1. The causal structure assumed in NP is implicitly accepted in game theory. Specifically, the causal structure of NP is identical to the causal structure assumed in the indirect–evolution approach to morality. In the indirect–evolution approach, this causal structure is moreover common knowledge among the players. Thus, the Unpredictability Principle is implicitly rejected in game theory: The indirect–evolution approach assumes the existence of causal links between Adam's choices and past events that can be used by Eve to predict his choices.

2. There remains the question whether, under the assumption that the NP causal structure holds, the Disconnection Principle or the Backtracking Principle should be used as a decision criterion (about the probability beliefs relevant to a player's expected–utility comparisons). We argue that this question can be resolved by recourse to an evolutionary competition where some decision makers rely on the Disconnection Principle, accordingly choosing two boxes, while others use the Backtracking Principle, which supports the one–box solution. In such a competition, use of the Disconnection Principle dies out. Therefore, the Disconnection Principle is inferior to the Backtracking Principle as a decision criterion.

In order to make our first point, we introduce a version of NP, called the *moral NP,* combining Selten & Leopold's (1982) game–theoretic version

of NP with Güth & Kliemt's (1994) indirect–evolution approach to moral preferences.

Moral preferences are modeled as additional internal payoffs that are added to (the utility resulting from) material payoffs. In the moral NP, Adam and the likes of him choose one box or two boxes according to their moral preferences. If moral preferences are observable by Eve, that is, she picks up a signal sent by nature informing her about Adam's moral preferences, she can perfectly predict his choice. In an evolutionary setting where moral preferences spread in the population depending on the material success their bearers achieve, moral preferences favoring the two–box solution are at a material disadvantage and die out.

In the *original NP,* Adam and the likes of him choose one box or two boxes depending on their decision criterion (Backtracking or Disconnection Principle), which determines the probabilities they plug into their expected–payoff calculations. If Eve picks up a signal sent by nature informing her about Adam's decision criterion, she can perfectly predict his choice.

The game trees for the moral and the original NP are very different. However, we show that the causal structure of both versions, which can be analyzed with the help of a *causal network,* is identical. Moreover, the causal network turns out to be consistent with the received views on both causality and rationality. This proves our first point.

The second point follows immediately. Whenever Eve can actually predict choices, a decision criterion favoring the two–box solution is at a material disadvantage and dies out. Note that in the original NP, players are motivated by success in material terms alone. Hence, those who use the Backtracking Principle are, in each instance where NP is played, subjectively better off after the game than the adherents of the Disconnection Principle. Accordingly, the Disconnection Principle loses the competition on all accounts: its supporters are subjectively and objectively worse off in each instance.

1.3 Beyond Newcomb's Problem

We introduce causal networks for three reasons. First of all, we need an explicit causal structure for the moral NP in order to show that it corresponds to the causal structure of the original version. This is not obvious from the game trees, which are very different (assuming that a satisfactory game tree for the original NP can be found at all, which is controversial). Secondly, the causal network demonstrates that the assumption of predictability is not in conflict with standard assumptions either about causality or about rationality. Last but certainly not least, we introduce causal networks to draw further lessons concerning predictability of choices and cooperation.

Although it raises interesting philosophical questions, NP in itself is not a very interesting problem for game theorists. Presumably, it attracted attention mainly because it was conjectured that any theory justifying the one–box solution in NP (where Adam resists the temptation to take both boxes) might also justify a cooperative outcome in the much more interesting Prisoners' Dilemma (PD), since cooperation in the PD requires that players resist the temptation to deviate unilaterally from cooperation.

When we come back to this point at the end of the paper, we argue that the causal network of NP, as should by then be quite obvious, can be extended so as to describe a situation of *mutual* predictability. Mutual predictability is a necessary ingredient for cooperation in the PD. However, as discussed below, mutual predictability (in contrast to one–sided predictability as in NP) cannot be modeled using a game tree. Therefore, the causal–network approach in this paper paves the way for an analysis of cooperation in the PD that cannot be based on a game–tree representation.

1.4 Organization of the Paper

Section 2 dicusses the moral NP. Section 3 introduces the causal network for the moral NP. Section 4 interprets the same causal network in terms of the original NP. Section 5 concludes with an outlook on the PD. Appendix A contains a Savage–type analysis for the decision criterion introduced in section 4. Appendix B discusses a counterargument (based on the so–called "medical NP") against the one–box solution mentioned in the conclusions.

2 The Moral Newcomb's Problem

2.1 Preferences, Beliefs, and Equilibrium

We consider an evolutionary setup where players are rational and the evolutionary process affects the determinants of the players' rational choices. We assume the replicator dynamics or any other payoff–monotonic evolutionary dynamics (see, e. g., Weibull 1995).

For our purposes, it is convenient to describe players as being committed to their choices through (a) their preferences, represented by a v. Neumann–Morgenstern (NM) utility function, (b) their beliefs, represented by subjective probabilities, and (c) their decision criterion, which determines how a player combines his or her preferences and beliefs into a subjectively optimal decision; including which probabilities are relevant for calculating expected

utilities. We call this a player's *motivation package.*[7]

The NM utility function defines subjective preferences (payoffs as perceived by the players). Players' subjective preferences may or may not reflect the material payoffs relevant for evolutionary success. According to the indirect–evolution approach, players are rational and, in any individual game, choose the equilibrium strategies on the basis of their subjective preferences. Evolutionary pressure affects the choice of strategy not directly but indirectly, by affecting its determinants, i. e., the motivation package: those motivation packages resulting in higher material payoffs for their bearers spread in the population.

In a Nash equilibrium, all players perfectly foresee the strategy choices of the other players: If a player selects a strategy, the other players assign probability 1 to him playing this strategy. Players' own strategy choices maximize expected payoffs given their expectations. To any player, the choices of other players are states of nature in Savage's sense; however, subjective probability assignments to these states reflect the rules of the game and have an equilibrium property, namely, perfect foresight of strategy choices.

In the case of incomplete information (uncertainty concerning the motivation package), the notion of Bayesian equilibrium applies. We introduce a chance move of nature selecting among different player types for certain player positions. The subjective prior probabilities for encountering a certain type are equal to the probabilities of nature's choices. The game is thus transformed into a game with imperfect information, and the Bayesian equilibrium is the Nash equilibrium of the latter.[8]

2.2 Making a Game of Newcomb's Problem

The insight of Frank (1987) is that a player may profit (in material terms) from being known to have a "conscience", and that this might lead this

[7]Savage (1954) has shown that this package—beliefs, preferences, and decision criterion—is equivalent to a more general kind of preferences. However, it is fundamental to Bayesianism that preferences in the narrower sense and beliefs are separable (Binmore 1993: 207, Aumann 1987: 13 fn 13). This implies that an agent can adopt an NM utility function independently from her beliefs, or beliefs independently of her NM utility function. On the former case, see also Binmore (1993: 207) on Savage and "massaging the priors". The latter case is illustrated by Savage's own use of the sure–thing principle as a device for (implicitly) adjusting his evalutation of NM utilities in the Allais Paradox (where probabilities are given); cf. Pope (1991). Separability of preferences (in the narrower sense) and beliefs implies that there is a third meaningful and separable component of Savage's general preferences, namely, a *decision criterion,* which says how to combine preferences and beliefs in order to find the best decision.

[8]For the equilibrium definitions, cf. Fudenberg & Tirole (1992).

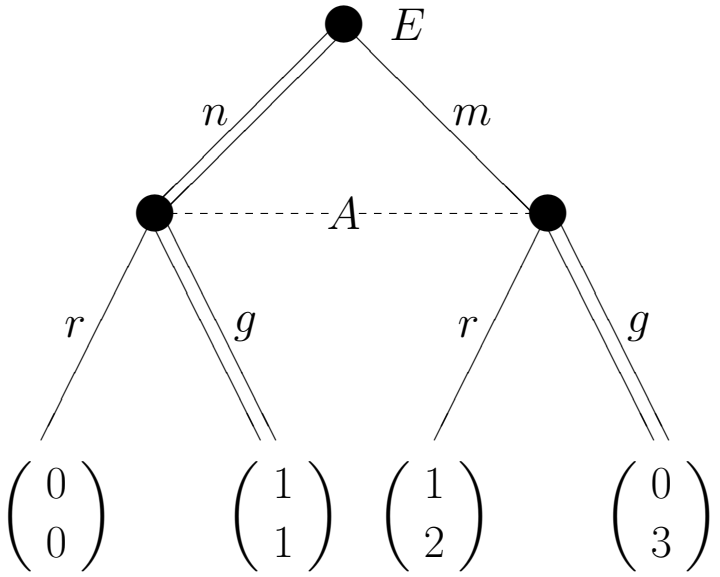

Figure 1: Selten–Leopold Version of Newcomb's Problem

feature to spread in a population. Güth & Kliemt (1994) analyze this idea in a very transparent way for the so–called Game of Trust. We apply their basic argument to Selten & Leopold's (1982) version of NP, which in contrast to the original version, also explicitly considers the decision problem of the predictor (Eve).

In the Selten–Leopoled version of NP, Eve has to decide whether to put a million into the opaque box (m) or not (n); and Adam, not knowing what she did, has to decide whether to resist the temptation to take both boxes (r) or be greedy (g). Let us assume for the moment that material and subjective payoffs coincide. For Eve, the payoffs are $\pi_E(m,r) = \pi_E(n,g) > \pi_E(n,r) = \pi_E(m,g)$, while Adam's payoffs are $\pi_A(m,g) > \pi_A(m,r) > \pi_A(n,g) > \pi_A(n,r)$. For simplicity's sake, we choose numerical values (figure 1):

$$\begin{array}{ll} \pi_E(m,r) = 1 & \pi_A(m,r) = 2 \\ \pi_E(n,r) = 0 & \pi_A(n,r) = 0 \\ \pi_E(n,g) = 1 & \pi_A(n,g) = 1 \\ \pi_E(m,g) = 0 & \pi_A(m,g) = 3 \end{array} \tag{1}$$

Eve is only interested in correctly predicting Adam and choosing the appropriate action. However, no matter what Eve does, being greedy is always better for Adam. Hence, the outcome is the two–box solution (n,g): Adam is greedy, and Eve, anticipating this, does not put money in the opaque box.

This outcome is inefficient. Adam has a commitment problem: he would like to commit to resisting temptation in order to secure higher payoffs; however, the game contains no commitment possibility.

The Selten–Leopold version of NP deviates from the original version. According to the latter, Eve can predict Adam no matter what he does. In the course of the story, she receives a signal that reliably tells her what Adam is going to do. A better reconstruction of NP as a game has to take this into account. The approach of Güth & Kliemt (1994) provides a simple way of doing so.

2.3 Moral Preferences

We modify the Selten–Leopold version of NP by introducing the possibility that a player has *moral preferences.* Moral preferences are modeled as an additional *internal* payoff that determines the relative attractiveness of Adam's choices for himself but is not directly related to evolutionary success. Material payoffs are still described by (1). Subjective preferences are described by the following revised payoffs:

$$\begin{array}{rlrl} \pi_E(m,r) &= 1 & \pi_A(m,r) &= 2+a \\ \pi_E(n,r) &= 0 & \pi_A(n,r) &= 0+a \\ \pi_E(n,g) &= 1 & \pi_A(n,g) &= 1-a \\ \pi_E(m,g) &= 0 & \pi_A(m,g) &= 3-a \end{array} \tag{2}$$

We assume that there are two populations of potential players: a population of identical individuals from which a player for Eve's role (the predictor) is selected, and a population of players from which a player for Adam's role (the predictee) is selected. In the predictee population, there are two types of players (see figure 2): the amoral type without a conscience ($a = 0$), and the moral type with a conscience ($a = 0.75$). If a predictee is of the moral type, his payoffs imply that he prefers r to g, no matter what the predictor chooses. If Adam is of the amoral type, he prefers g to r, again no matter what the predictor chooses. The fraction of moral players in the predictee population is $p \in (0, 1)$, which is equal to the probability of the predictor encountering such a player.

We assume that the predictor observes a signal sent by nature informing her about the type of the predictee (*type signal*). The type signal is equivalent to a brain scan revealing the moral preferences; it is not observable to the predictee. There is no need to represent the type signal in the game tree explicitly as long as the signal perfectly predicts the predictee's type; we just assume that the predictor observes nature's selection of the predictee's type.

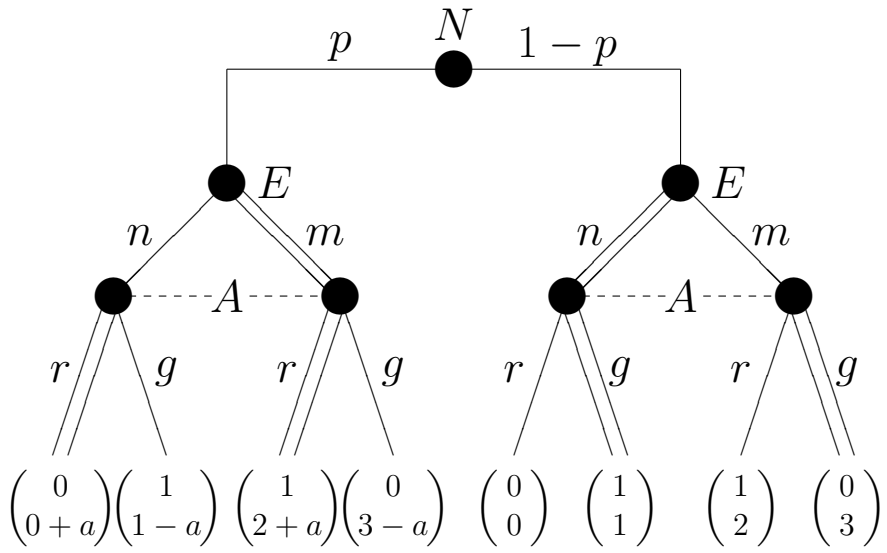

Figure 2: The Moral NP. With probability p, Eve' opponent Adam has a conscience (LHS payoffs: $a = 0.75$); with probability $1 - p$, Adam is only motivated by material payoffs (RHS payoffs = LHS payoffs but $a = 0$).

Given the assumption on the predictee's preferences, the type signal is perfectly correlated with the predictee's choices. The unique Nash equilibrium for play with a moral type is the one–box solution (m, r) (LHS of figure 2), which is efficient. The amoral types are still locked in the old two–box equilibrium (RHS of figure 2).

We call the complete game including nature's initial move the moral NP. In the evolutionary setting, the relative success in terms of material payoffs determines whether the share of moral types p in the predictee population goes up or down. Since moral types always get a higher material payoff, they will spread, and p approachs 1 in the long run. Observability of the conscience is an evolutionary advantage for the moral types; morality solves the commitment problem.[9]

2.4 Subjective Rationality vs Commitment Power

Although the motivation package commits players to their choices, we are not introducing "commitment power". Commitment power means that a player can follow through with a previously chosen strategy even in a situation where

[9]The question of observability is extensively discussed by Frank (1987, 1988).

his motivation package recommends a different strategy as his optimal choice (within a larger set of feasible strategies determined by external constraints). The commitment achieved by commitment power is equivalent to delegation: When a commitment to a strategy is made, choices are delegated to another player (a hypnotized *alter ego*) who is only motivated to play this different strategy.

Commitment power together with the observability of commitment reduces cooperation problems to problems of coordination where efficient equilibria always exist. This can readily be seen from the optimal–delegation versions of NP and PD. If Adam in NP delegates the choice to a player with an unambiguous incentive to take only one box, he earns his million. In a PD, Adam and Eve can each delegate their choices to another player who carries a certifiable order of the respective principal and is known to receive 1 if carrying out this order and 0 otherwise. If Adam and Eve coordinate on the following order, the new players choose cooperation: "Cooperate if and only if your partner carries the same order as yourself." This is an equilibrium.[10]

The indirect–evolution approach leads to similar results as commitment power or delegation (cf. Dufwenberg & Güth 1999). Nevertheless, the causal mechanism is different. Let us call player 1 *causally separated* from player 2 after time t if and only if player 1 receives no information concerning the behavior or motivation package of player 2 after time t.[11] Cooperation by commitment power requires that a players' behavior after causal separation is no longer caused by his or her original motivation package. Thus, in the optimal–delegation versions of NP and PD explained above, the players to whom choice is delegated are motivated by other payoffs than their principals.

In contrast, this paper focuses on explanations of cooperation where the players always follow the recommendations of their motivation packages chosen by nature, even after causal separation. That is, a player always does what is best according to his or her motivation package at the time a decision is actually made. In Güth and Kliemt (1994), players have moral preferences that bind them to certain strategies. In section 4, players use different decision criteria with the same result. In both cases, *players always choose what their motivation package recommends as their best decision at the time of action.* Hence, being committed (to rational behavior) by one's motivation package has nothing to do with (irrational) commitment power; that is, the power to commit to irrational behavior.

[10]See also Rubinstein (1998: ch. 10) on achieving cooperation by delegating choices to Turing machines.

[11]We can identify causal influences with information because perfectly rational players are transparent to themselves; they cannot be causally influenced without being aware of it.

3 A Causal Network

The present section describes the causal structure of the moral NP with the help of a causal network.[12] The moral NP introduces no assumptions violating the strictures of game theory. The game tree of the moral NP (figure 2) summarizes all the information relevant to solving the game. The causal network will contain no surprises; it just illustrates a causal structure taken for granted when drawing the game tree. Why, then, do we introduce the causal network at all?

The answer becomes clear in the next section. The causal network for the moral NP is the same as the causal network for the original NP. Both versions of NP assume exactly the same kind of predictive abilities for Eve. This is the basis for rejecting the idea that the predictability premise of the original NP is somehow illegitimate from a game–theoretic point of view.

The relevant causal connections of the moral NP are illustrated in figure 3. Arrows indicate the direction of a causal influence from one node to another. If all arrows leading to a node are reached (or if the node is an initial node), the node is reached. Since causal influences are always forward in time, the nodes of the network are partially ordered with respect to the time at which they are reached.

The most important elements of figure 3 besides the arrows are square nodes (s–nodes). S–nodes stand for physical or mental states or events with a definite physical or mental description.[13] States or events can be causes of other states or events; nothing else can be a cause. Except for initial or terminal s–nodes, exactly one arrow leads to and exactly one arrow emerges from an s–node. If an s–node is reached and an arrow emerges from it, the arrow is activated, which means that there is a causal influence resulting from this s–node. Two s–nodes connected by an arrow illustrate the simplest case of cause and effect. By building a chain of s–nodes connected by arrows, we can describe a simple causal chain.

Decision nodes (d–nodes), represented by circles in squares, are special instances of s–nodes; they stand for a certain class of mental events, namely, decision making.[14]

[12]Several kinds of such networks can be found in the literature; we have developed our own version adapted to the purposes of decision and game theory. Causal networks (our version) are connected finite oriented 1–graphs without circuits. The latter condition means that they are acyclic in the terminology of Pearl (2000). For a technical reference containing exact definitions and useful theorems and proofs, see Albert & Heiner (2000).

[13]We distinguish between states and events in the ordinary–language meaning of the term and "states of nature" in the sense of Savage (1954).

[14]Decision making of real (boundedly rational) persons should be viewed as a process, which could (but need not) be modeled as a causal chain (because arrows like those

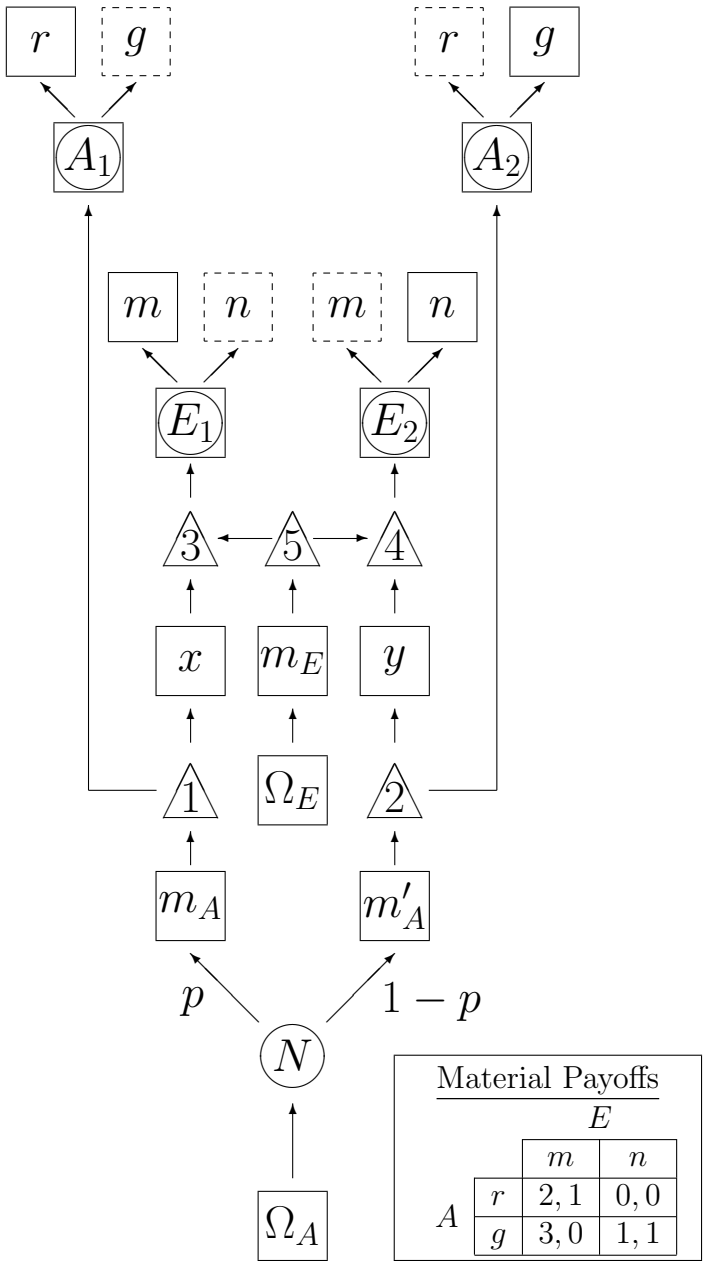

Figure 3: Causal Structure of Newcomb's Problem

Initial s–nodes (or any other initial nodes) are always reached; for all other nodes, it remains to be determined whether or under what conditions they are reached.

Apart from states and events represented by s–nodes, there are other kinds of effects, namely, probability distributions over states/events. For instance, if a player is drawn randomly from a population containing different types of players, this means that certain initial conditions cause a probability distribution over different types. Of course, if some state or event causes a probability distribution, the outcome of this distribution is again a state or event.

Probability distributions are described by circle nodes (c–nodes). Exactly one arrow leads to a circle node but at least two arrows bearing non–zero probabilities emerge from it, each leading to another s–node. The probabilities of the arrows emerging from one c–node add to 1. Exactly one of the arrows succeeding a circle node will be activated; all other causal chains potentially emerging from the c–node are cut off. Circle nodes bear the letter N (for nature), with a subindex for identification if there are several of them.

Triangle nodes (t–nodes) stand for "simultaneous" causation:[15] if a t–node is reached, all causal chains emerging from it are activated. T–nodes are also used to illustrate joint causation: a t–node is only reached if all the arrows leading to it are activated. It is possible for a t–node to have several arrows leading to it and several arrows emerging from it (although such a case does not appear in figure 3). Such a construction can be used to illustrate a situation where several causal influences jointly lead to several simultaneous effects. In order to simplify the discussion of diagrams, t–nodes bear a number by which we may refer to them.[16]

Figure 3 describes the selection of the two players and their choices. It closely corresponds to figure 2. One of the two initial s–nodes is labeled Ω_A; it describes the initial conditions and the drawing of the second player

preceding a d–node can always be viewed as causal chains involving further events like thinking that need not occur explicitly). In the case of perfectly rational players, however, a decision *process* is not necessarily involved: such players may already know (as if they had already calculated) what their optimal choices would be under any potential set of circumstances. In this case, a d–node just stands for the irrevocable selection of an action and could without loss of information (but with a serious loss of readability) be dropped from the diagram.

[15]If several effects are caused *simultaneously,* this means that all must occur and that none of them is causing another. It is not implied that they occur at the same time.

[16]The definition of a causal network automatically rules out certain nonsensical constructions like two arrows emerging from the same c–node leading to the same t–node: This would mean that the t–node is only reached if two mutually exclusive events occur. For the details, see Albert & Heiner (2000).

(Adam) from the population. The arrow emerging from this s–node leads to a circle node because two types of players can be drawn: either a moral type or an amoral type. The probabilities for drawing these types are p for the moral and $1 - p$ for the amoral type. The s–nodes labeled m_A and m'_A stand for the motivation packages of the moral and the amoral type. Thus, with probability p, the moral type (motivation package m_A) is chosen. The only difference between the motivation packages is the players' preferences. As in figure 2, the subjective preferences of the moral type derive from the material payoffs (repeated in the lower right corner of figure 3) by adding or substracting an additional internal payoff $a = 0.75$, while the subjective preferences of the amoral type are described by the material payoffs as they are (that is, $a = 0$).

From each s–node standing for a motivation package, an arrow leads to a t–node. Thus, there is an arrow from the s–node labeled m_A to t–node 1, and an arrow from the s–node labeled m'_A to t–node 2. From each of these t–nodes, two arrows emerge, indicating that there are two simultaneous causal consequences in each case. One is a signal, x in the case of t–node 1 and y in the case of t–node 2. The other causal consequence of a motivation package is Adam's decision. Hence, from each t–node there leads an arrow to a d–node labeled A for Adam.

Each of Adam's d–nodes leads to a single decision, r for d–node A_1 and g for d–node A_2. This illustrates the fact that the motivation package commits a player to his choices. Nevertheless, the other choice is feasible as far as the *external* constraints are concerned; and Adam must hypothetically assume that he actually chooses all the actions from this feasible set in order to determine his optimal decision. Hence, we have used broken–line s–nodes to indicate the other choices, so that all actions allowed by the external constraints appear at each d–node. However, the actions represented by broken–line s–nodes are not actually taken, because they are not optimal according to Adam's motivation package.

This indicates that players are not free to choose just anything; they are only free to choose what they prefer, and what they prefer is determined by their motivation package, i. e., preferences, beliefs, and decision criterion. The circle in the square of a d–node has a similar meaning as the c–node: only one of a set of mutually exclusive and jointly exhaustive alternatives is actually chosen.

There is another initial s–node labeled Ω_E, which stands for the conditions under which the first player (Eve) is drawn from the respective population. While the first player is also randomly drawn from a population, there is no probability distribution because there is only one type of player for Eve's role. Hence, there is just one arrow leading from the s–node labeled Ω_E to

an s–node labeled m_E, which stands for the motivation package of the first player.

We now have to represent the fact that the signals sent by nature are received and acted upon by Eve. From the node labeled m_E an arrow leads to t–node 5, from which two further arrows emerge, each leading to a further t–node (3 and 4). T–node 3 is met also by an arrow emerging from the s–node labeled x, while t–node 4 is met by an arrow emerging from y. Since one of the arrows meeting t–node 3 is always activated (the one coming from the s–node labeled m_E via t–node 5), d–node E_1 (succeeding t–node 3) is reached if and only if the signal is x, in which case Eve chooses m. Likewise, d–node E_2 (succeeding t–node 4) is reached if and only if the signal is y, in which case Eve chooses n. Hence, figure 3 shows that Eve, committed by her motivation package m_E, chooses m if and only if she receives a signal telling her that Adam is of the moral type, whose motivation package m_A will cause him to choose r instead of g.

The possibility of having several initial nodes allows us to visually represent the assumption that players motivation packages are determined *independently* from each other. Moreover, the diagram makes clear that there is no influence from the decision or action of one player to the decision or action of the other player. The network diagram thus demonstrates that no illegitimate causal relations are smuggled into the analysis. In trees with only one initial node (arborescences), it is not possible to represent this kind of independence visually since all events must be positioned in a (partly arbitrary) sequence.[17]

Another difference between the causal network of figure 3 and game trees like those in figure 1 is that figure 3 has no information sets. At each d–node, a player deduces exactly where he or she is.[18] The information set in figure 1 is used to express the fact that Adam cannot react to Eve's choices; it is a consequence of the fact that a game tree cannot be used to depict parallel independent processes. In figure 3, it is clear that Adam cannot react to Eve's

[17] Also, an arborescence representing equilibrium choices becomes trivial because except for nature's choice at the beginning there are no further branchings; cf. e. g., Binmore's (1994: 250, figure 3.10a) representation of NP. While causal networks can always be mapped to arborescences, this results in a loss of (causal) information; see Albert & Heiner (2000).

[18] Of course, there are situations where we would have to draw information sets including several d–nodes. In an imperfect–predictability version of NP, for instance, the signal Eve receives could be an imperfect indicator of Adam's type: there might be a small probability that m_A causes signal y instead of x to be received. Thus, upon receiving signal y, Eve would be uncertain as to the true cause of this signal. Generally, an information set is needed if a relevant part of the history preceding his or her decision is not known to a player.

choice because there is no signal sent by nature indicating Eve's action, on which Adam then could react.

The causal network forces us to be very explicit concerning observation. Observing a state or event means that there is a causal influence from the state or event reaching the player in question. Adam's and Eve's knowledge is described by figure 3. Hence, Adam can conclude what Eve is going to do but he cannot actually observe it.

Moreover, since Adam knows which d–node he is at, he knows his own type (note that there is a causal influence from his motivation package to his d–node). If Adam is the moral type, he knows that he will choose r, and he is satisfied with this knowledge since his motivation package implies that he thinks r is best. Likewise, he knows if he is the amoral type and is equally satisfied with his choice of g. Of course, an amoral Adam would prefer to be able to pose as a moral type and then exploit Eve's trust; however, given that such a deception is impossible, he views g as optimal and does not regret that he is committed to it by his motivation package.[19]

Thus, figure 3 illustrates the indirect–evolution version of NP. Both players are always subjectively rational according to their motivation package. Adam can be either of the moral type or of the amoral type; the difference is the value of the internal payoff a distinguishing the motivation packages m_A and m'_A. Eve can observe Adam's type; this is illustrated by the signal taking on values x (for the moral type with $a = 0.75$) or y (for the amoral type with $a = 0$).

Figure 3 assumes nothing not already assumed in section 2.3. Players are committed to their actions by their motivation package (preferences in the sense of NM utility functions, beliefs, decision criterion); there exists a signal that allows one player to distinguish the type of the other player. The causal information contained in figure 3 can be common knowledge among the players, although not all of the information is relevant to them. The information contained in figure 3 together with the payoff information can be used to construct a game tree, namely, figure 2, representing what players need to know in order to determine their optimal (expected–utility maximizing) choice. There is a loss of causal information when replacing figure 3 by figure 2, but such information is not necessarily required in order to determine players' optimal strategies.

[19]One might consider the question whether an amoral Adam would wish to be moral. Analyzing this problem presupposes that Adam is able to make interpersonal comparisons of utilities. Nowhere in this paper such comparisons will be made.

4 The Original Newcomb's Problem

4.1 Setting Up a Game Tree

There is still a key difference between the moral and the original NP. In the original version, Adam has no additional internal payoffs; he is only considering what he should do in the light of the material payoffs and his knowledge that Eve can predict his choices.

Let us assume, then, that there are two types of predictees who are both motivated by material payoffs alone but differ in some other respect. Both types are predictable, both believe that they are, and both share all other factual beliefs. However, their decision criteria differ, leading them to use the same probability beliefs differently, as explained in the introduction. One type relies on the Backtracking Principle and, therefore, uses probabilities for Eve's action that are conditional on his own action; these probabilities are determined by backtracking considerations. The other type takes Eve's action as a fixed past event which is thereby disconnected from his own action, as required by the Disconnection Principle. Adam's other type therefore uses a fixed probability for Eve's action that is independent of his action.

We first try to represent these assumptions with the help of a game tree (figure 4). If Adam uses action probabilities for Eve that are conditional on his own action, this means that he chooses as if he were the first mover (LHS of figure 4). If, on the other hand, he views Eve's action as causally disconnected from his own, we are just back to the Selten–Leopold version of NP (RHS of figure 4).

However, this game tree is problematic. First of all, remember that two–boxers argue that Adam, if he is rational, must consider Eve's action as fixed because it happened in the past. Hence, assuming on the LHS that Eve is second mover and Adam is rational amounts to rejecting their position without argument.

Secondly, the LHS looks as if Eve could somehow observe Adam's choice after the fact even though she actually chooses before Adam chooses. In contrast, the RHS shows a simultaneous game between Adam and Eve, as if Eve could use only payoff information to predict Adam's future choice. Yet, Eve uses *the same means* to predict Adam on both sides; namely, receiving a signal revealing Adam's type of decision criteria. Why, then, does Adam's way of thinking about the problem determine the representation of Eve's information in the game tree?[20]

[20]The symmetries of NP speak against representations where more than one action is available at one node. Since Eve is always in the same position relative to Adam, she should be second mover on both sides or on neither side. If she is second mover on

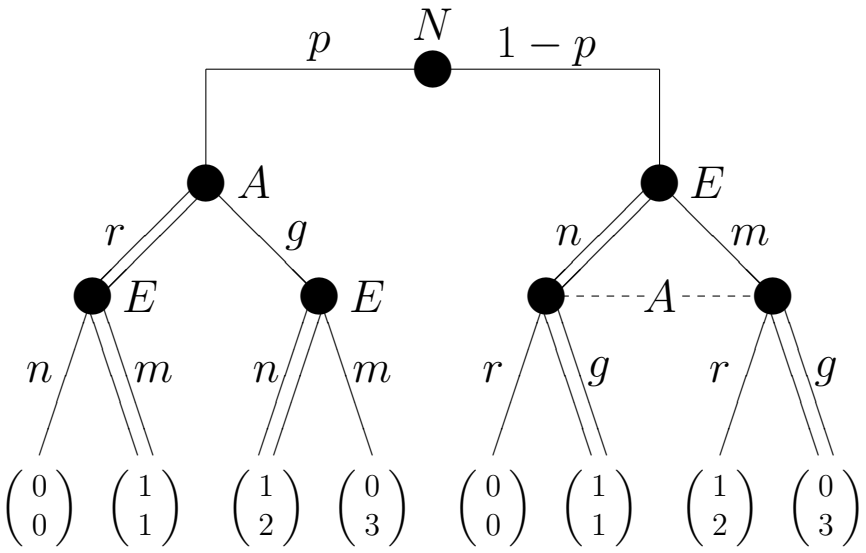

Figure 4: The Original NP. With probability p, Eve's opponent Adam relies on the Backtracking Principle (LHS); with probability $1-p$, he relies on the Disconnection Principle (RHS). Eve's payoffs are given first on both sides.

This question cannot be answered by analyzing the game tree. We have set up this tree such that, without differences in payoffs between the RHS and the LHS, the same equilibrium results as in the moral NP (figure 2). This, however, is not an analysis. We have started from some result and have constructed a quite arbitrary–looking game tree that delivers the result. The tree itself cannot be used to decide whether it represents a legitimate causal structure or not because it has been designed to deliver the result without reference to causal structures.

Of course, after analyzing the causal structure of the original NP, we may come to the conclusion that, after all, figure 4 represents the original NP in a satisfactory way. But this conclusion must obviously be based on an analysis that goes beyond game trees.[21]

both sides (neither side), Adam's action r (g) must be removed on the RHS (LHS) since otherwise he would not choose g (r). Adam is equally free to choose on both sides of the tree; hence, symmetry requires that we remove Adam's non–chosen actions everywhere. However, Adam and Eve are equally free to choose; hence, symmetry requires that we also remove Eve's non–chosen actions. This leaves us with a trivial and useless tree just displaying the equilibrium choices. Cf. also Binmore (1994: section 3.5).

[21] We will not discuss the question whether such a conclusion is indeed warranted. In any case, a game tree for NP involves some potentially controversial innovation. Traditionally,

If, on the other hand, the reader already feels that figure 4 is *not* a good representation of the original NP, this is grist to our mill. Game-tree representations tend to proliferate when discussing NP since all trees proposed have some problematic features. While causal networks for NP are unambiguous, the attempt to encode the same information within the strictures of game–tree modeling does not, in our experience, lead to agreement on representations. Therefore, we use causal networks to make progress beyond figure 4.

4.2 Reinterpreting the Causal Network

In section 3, we have shown that the causal structure behind the moral NP is perfectly reasonable. Subsequently, we show that the same structure can be used to describe the original NP. The reason why, nevertheless, the game trees of the moral NP (figure 2) and the original NP (figure 4) look so different lies in the restrictions of the game–tree representation.

We focus on a version of the original NP where players rightly assume (or, in other words, know) that the predictee is in fact predictable but where there are still two types of predictees coming to different conclusions about their optimal choice.[22]

Let us summarize the main points of the analysis so far. Adam as the predictee does whatever he is committed to by his type, that is, his motivation package. A type who is observably committed by his motivation package to resisting the temptation and rejecting the transparent box does consistently better than a type who is not so committed. Nothing in this basic argument depends on *what part* of Adam's motivation package, preferences, beliefs, or decision criterion, does the job of committing him to taking only one box. The only important point is observability: Eve receives a signal revealing Adam's type, as illustrated by the causal network 3.

Let us assume, then, that there are two types of predictees who are both motivated by material payoffs alone but differ with respect to the decision criterion they use. Both types are rational in the sense of maximizing expected payoffs, although the expected–payoff formulas differ (otherwise both

a game tree is a description of the rules of the game that is independent of the way players make decisions and solve the game. Adjusting the representation of subgames to the decision criterion used by a player in the subgame transcends a game tree's traditional meaning and purpose.

[22]If there were a third type who, motivated for instance by the Unpredictability Principle, did not believe in his own predictability, this type would still end up with the same action as the type who, despite believing in his own predictability, chooses two boxes. Hence, ruling out the disbelievers does not change the results of the analysis.

would choose the same action). Both types accept figure 3 as a correct description of their situation: There is a signal that tells Eve which decision criterion they use.

Since both types are perfectly rational, they are transparent to themselves, that is, they know their own motivation package and already know beforehand which action they prefer in any kind of situation they might themselves find in. Therefore, they do not need to calculate expected payoffs in order to *find out* what they should do. The action they prefer is always the one recommended by the relevant version of expected–payoff maximization.

In the next two subsections, we show that there are two decision criteria involving expected–payoff maximization that can serve as the distinguishing elements of motivation packages m_A and m'_A in figure 3. These decision criteria are based on the Backtracking Principle (in the case of m_A) and the Disconnection Principle (in the case of m'_A). When we have shown that these decision criteria lead to the respective choices of one box or two boxes, our reinterpretation of figure 3 in terms of the original NP is complete. We have then *proved* that figure 3 illustrates the causal structure of the original NP.

4.3 The Disconnection Principle

According to the Disconnection Principle, Eve's actions must be assigned fixed probabilities by Adam, despite the fact that Adam's motivation package is a common cause of his impending action and Eve's earlier action. Formally, this is achieved by holding constant Adam's motivation package, which fixes Eve's action probabilities, while still hypothetically changing Adam's actions.

Let us pretend that we as game theorists do not know which motivation package is consistent with the Disconnection Principle. In order to answer this question, we need to calculate the expected payoffs for Adam's actions according to the Disconnection Principle. That is, we must use probabilities for Eve's action that are conditional on a given motivation package, m_A or m'_A, while varying Adam's chosen action from g to r.[23] Thus, we (not Adam) have to compute four expected payoffs. A motivation package is consistent with the Disconnection Principle if and only if the action caused by the package is also the action recommended by the expected–payoff calculations conditional on the same package.

So first consider motivation package m_A in figure 3. Probabilities for Eve's actions conditional on m_A can be found as follows. If the s–node labeled

[23]In the present case, Adam's motivation package can be used in as a stand–in for Adam's information. In general, conditioning on the d–node (or the information set) would be appropriate. See also fn 24 below, where the Disconnection Principle is derived from Aumann's (1987) definition of Bayesian rationality.

m_A has been reached, the arrow emerging from this s–node is activated; t–node 1 and the s–node labeled x have also been reached. Eve's inital node Ω_E, s–node m_E and t–node 5 have been reached independently of Adam's motivation package. Since both t–node 5 and s–node x have been reached, Eve made her decision at the d–node E_1. Hence, she chose m and not n. Thus, the relevant probabilities are $P(n|m_A) = 0$ and $P(m|m_A) = 1$.

If we plug these probabilities into expected–payoff calculations, we get the following results:

$$\begin{array}{rcl} \mathrm{EU}(r|m_A) & = & P(n|m_A)\pi(n,r) + P(m|m_A)\pi(m,r) \; = \; \pi(m,r) \\ \mathrm{EU}(g|m_A) & = & P(n|m_A)\pi(n,g) + P(m|m_A)\pi(m,g) \; = \; \pi(m,g) \end{array} \tag{3}$$

Notice that the two expected–payoff formulas in (3) change Adam's action from r to g with no change in Eve's probabilities of choosing n or m, because her action probabilities are conditioned on holding Adam's motivation package constant at m_A regardless of his chosen action. Thus, from our assumptions on payoffs, it follows that the expected payoff of action g is higher than for action r. Hence, these calculations recommend g. Since, however, motivation package m_A causes Adam to actually choose action r, we have proved that motivation package m_A is inconsistent with the Disconnection Principle.

Analogous reasoning for motivation package m'_A leads to the following expected payoffs:

$$\begin{array}{rcl} \mathrm{EU}(r|m'_A) & = & P(n|m'_A)\pi(n,r) + P(m|m'_A)\pi(m,r) \; = \; \pi(n,r) \\ \mathrm{EU}(g|m'_A) & = & P(n|m'_A)\pi(n,g) + P(m|m'_A)\pi(m,g) \; = \; \pi(n,g) \end{array} \tag{4}$$

Similar to (3), Eve's probabilities of choosing n or m are conditional on holding Adam's motivation package constant at m'_A regardless of his chosen action. And so the two expected–payoff formulas in (4) again change Adam's action from r to g with no change in Eve's probabilities of choosing n or m. We thus again have Adam's expected payoff higher for action g than for action r. In this case, motivation package m'_A causes Adam to actually choose action g. Hence, the assumption that motivation package m'_A is based on acceptance of the Disconnection Principle is consistent with figure 3.

The rationale for conditioning on one's own motivation package is that conditioning on given information is the standard procedure in decision theory.[24] Still, readers may wonder whether, not mathematically but on some metalevel, it is consistent to both accept figure 3 and condition on one's own

[24]Thus, the Disconnection Principle follows from Aumann's (1987: 7) definition of Bayesian rationality, which can be applied to figure 3. Possible worlds are maximal sets of nodes that can be reached jointly. Only s–nodes (states and events) matter.

motivation package (as if one's motivation package is constant while also assuming one chooses different actions). In particular, when Adam computes his expected payoff for taking only one box, he still uses probabilities conditional on his having a given motivation package m'_A *even though the causal network implies that having this motivation package makes it impossible for Adam to actually take only one box.* That is, the causal network tells Adam that only a type with a *different* motivation package (m_A instead of m'_A) actually resists the temptation, and for this type the probability of Eve putting a million into the opaque box is necessarily also *different:* it is not 0 but 1.

Although we believe that conditioning on one's own motivation package is the best way of rationalizing the two–box choice in the case of figure 3, it does not really matter for our overall argument how, exactly, one can derive the two–box solution.

4.4 The Backtracking Principle

In contrast to the Disconnection Principle, the Backtracking Principle requires that the causal history of Adam's actions must be taken into consideration when computing probabilities for other events like Eve's actions.

Let us pretend that we as game theorists do not know which motivation package corresponds to the Backtracking Principle. In order to answer this question, we need to calculate the expected payoffs for Adam's actions according to the Backtracking Principle. That is, we must use probabilities for Eve's action that are conditional on Adam's hypothetically chosen action (derived by backtracking from this action).[25] A motivation package is consis-

The inclusion of broken–line s–nodes leads to actually impossible worlds. The set of possible (or rather: conceivable) worlds consists of two actually possible worlds $\omega_1 = \{\Omega_A, m_A, x, A_1, r, \Omega_E, m_E, E_1, m\}$ and $\omega_2 = \{\Omega_A, m'_A, y, A_2, g, \Omega_E, m_E, E_2, n\}$, and six worlds ω_i, $i = 3, \ldots, 6$ that are actually impossible because they derive from $\omega_{1,2}$ by changing only players' actions, but not their motivation packages. For instance, $\omega_3 = \{\Omega_A, m'_A, y, A_2, r, \Omega_E, m_E, E_2, n\}$ (where Adam chooses r while Eve chooses n because she receives signal y and so, wrongly, predicts g) is impossible since Adam cannot choose r when his motivation package is m'_A (as indicated by a broken–line s–node for r at A_2). By assumption, the impossibility of $\omega_{i>2}$ is known to the players. So in Aumann's analysis, the common prior P on Ω must assign probabilities $P(\omega_{i>2}) = 0$ and, of course, $P(\omega_1) = 1 - P(\omega_2) = p \in (0, 1)$. Players recognize the possible world they are in. According to Aumann's definition, Adam is Bayes rational at ω_2 (but not at ω_1): Adam's action at ω_2, g, is better than action r, because he receives higher payoff from ω_2's action profile (g, n) than he receives from action profile (r, n). However, (r, n) occurs only in impossible worlds like ω_3. Cf. also (4); conditioning on m_A (m'_A) or ω_1 (ω_2) is equivalent.

[25]Generally, the Backtracking Principle requires conditioning on the action and current information as far as it is consistent with the action's causal history as found by backtracking. Specifically, information about the player's decision principle must be ignored if

tent with the Backtracking Principle if and only if the action caused by the package is also the action recommended by the expected–payoff calculations.

We begin by selecting one action for consideration, say, g. We then "backtrack" from the assumed occurrence of g, collecting all of g's predecessors (immediate or not) that must occur in order to cause g to be chosen in the first place. This collection includes the nodes A_2, t–node 2, m'_A, N, and Ω_A, because these nodes have to have been reached before g is reached. Starting from this collection, one can go forward again to determine all the nodes that have to be reached if g, A_2, t–node 2, m'_A, N, and Ω_A are reached. Of course, Ω_E, m_E, and t–node 5 are always reached. Since t–nodes 4 and 5 are reached, nodes E_2 and n have also to be included. All other nodes are definitely not reached if g is reached.[26]

Hence, figure 3 reveals that n (but not m) is reached iff g is reached. Analogously, m (but not n) is reached iff r is reached. This means that figure 3 implies the following conditional probabilities: $P(n|g) = 1$, $P(m|g) = 0$, $P(m|r) = 1$, $P(n|r) = 0$. So in contrast to expressions (3) and (4) above, changing Adam's action from r to g necessarily implies that Eve's probabilities of choosing n or m must also change from 1 to 0.

Plugging these probabilities into expected–payoff calculations, we get the following results:

$$\begin{array}{rcl} \mathrm{EU}(r) & = & P(n|r)\pi(n,r) + P(m|r)\pi(m,r) \; = \; \pi(m,r) \\ \mathrm{EU}(g) & = & P(n|g)\pi(n,g) + P(m|g)\pi(m,g) \; = \; \pi(n,g) \end{array} \tag{5}$$

From our assumptions on payoffs, it follows that the expected payoff of action r is higher than for action g. Hence, these calculations recommend r. Since only motivation package m_A leads Adam to actually choose action r, we have proved that only motivation package m_A is consistent with the Backtracking Principle.

Again, one may wonder whether this procedure is consistent on some metalevel. If Adam knows that he has motivation package m_A, how can it be relevant to consider a case where he has a different motivation package m'_A? However, making a decision necessarily means comparing different actions, as if each different action was actually chosen from those possible

this information identifies a motivation package that is inconsistent with the action. Other information about the motivation package must be used (cf. appendix B).

[26]For a general philosophical defense of backtracking inferences see Carroll (1994). In general, backtracking cannot reasonably be avoided in making predictions. For instance, we always backtrack from what we perceive to be the causes of our perceptions and then use the result to predict what might happen next or what has to have been going on without us perceiving it. The only question is whether players should use backtracking with respect to their own pending decisions.

according to external constraints. And the causal network necessarily implies that different actual actions have different causal histories, including different motivation packages. Ignoring this fact and assuming, counterfactually, that different actions can actually result from the same motivation package is the problematic aspect of the Disconnection Principle. The Backtracking Principle is the only alternative.

4.5 The Two–Boxers' Dilemma

As in the case of the moral NP, figure 3 can be common knowledge. Everybody knows (knows that everybody knows, and so on) what everybody else prefers, believes, and computes as the best action, which is then actually taken, to nobody's surprise. The only fact not known to Eve beforehand is the type she is playing against; she cannot recognize it independently from the type signal. The type signal reveals Adam's type, where the only difference between types is the way they use the information of figure 3. Both types agree with respect to the various probabilities that can be computed from the causal network but opt for using different probabilities in their expected–payoff calculations; depending on the type of decision criteria contained in their respective motivation packages (either the Disconnection Principle or the Backtracking Principle).

In an evolutionary competition between the Disconnection Principle and the Backtracking Principle, the latter wins, because the backtracking types always do better than the non–backtracking types. The Backtracking Principle wins for the same reason that morality spreads in the moral NP: backtracking types and moral types are observably committed to the one–box solution, the former by their decision criterion, the latter by their moral preferences. The assumptions concerning observability are the same in both cases because the causal structure is identical.

Our analysis implies that two–boxers face a dilemma. They can reject the causal assumption of the original NP, namely, that Adam is predictable on the basis of a type signal. Or they can accept predictability. If they reject predictability, they also have to reject the indirect–evolution approach (the moral NP) because the causal structure is the same as in the original NP. If they embrace predictability, they must state a decision criterion that delivers the two–box solution under predictability conditions. We think that we have already spelled out the best version of such a criterion. But never mind the exact form of the criterion. However the two–boxers' rationality analysis might be reformulated, if it does its job and rationalizes the two–box solution, it is doomed in an evolutionary competition against the Backtracking Principle.

5 Conclusion

The present paper has demonstrated that the causal structure of NP is identical to the causal structure implicitly assumed in the indirect–evolution approach. While previous indirect–evolution models assumed that player types differ with respect to their preferences, we have reconstructed NP by assuming that player types differ with respect to decision criteria: One type uses the traditional Disconnection Principle leading to the two–box solution; the other type employs the Backtracking Principle and ends up with the one–box solution. In an indirect–evolution model where the predictor receives a signal indicating the predictee's type, the Disconnection Principle dies out in the long run and the one–box solution prevails. In our view, this suggests that the Backtracking Principle embodies rationality, not the Disconnection Principle: A decision principle that is consistently less successful than a rival principle should not be viewed as a principle of rational decision making.

Our analysis is open to two criticisms. First of all, one could claim that, in the real world, there are no signals telling one person about the type of another person. However, this criticism would apply to the indirect–evolution approach in general.

Secondly, one could argue that the problem posed by NP must be solved by an appeal to first principles; it cannot be excluded from the outset that evolution may favor irrational behavior.

While we agree with the latter part of this argument, we do not quite see how it could apply in the present case. Adam, the predictee in NP, has well–defined preferences over outcomes: he prefers more money to less. Our analysis shows that, in terms of these preferences, the Backtracking Principle is better than the Disconnection Principle in every single instance of NP, no matter whether the evolutionary equilibrium is reached or not. Moreover, this is known to Adam. Therefore, normative questions of first principles should not arise. Adam prefers more money to less, and it is a fact that he makes more money by adopting the Backtracking Principle.

There is a third criticism which is sometimes discussed under the heading of the "medical NP". Ronald A. Fisher explained data showing a correlation between smoking and lung cancer by the hypothesis that there is a genetic common cause for both, the preference for smoking and the tendency to get lung cancer.[27] If Fisher's explanation for the correlation were correct, it would be irrational to stop smoking because one feared to get lung cancer. While Fisher's hypothesis is probably not true, it seems that his argument is sound. This example is sometimes used to argue that the one–box solution

[27] See Levi (1985) for Fisher's argument and the connection to NP.

to NP is unsound: It is claimed that (i) the causal structure of Fisher's smoking–and–cancer problem and of NP are the same, (ii) the decision to stop smoking is analogous to the one–box solution, and, therefore, (iii) the one–box solution is as irrational as the decision to stop smoking.

However, according to Fisher's assumptions, the decision to stop smoking has no effect on the chances of survival. The causal structure is not the one of NP: If you condition on your action, material payoffs (with the exception of the gratification derived from smoking) are the same, whether you stop smoking or not. Thus, premise (i) is false and, therefore, conclusion (iii) does not follow. If Fisher's assumptions are correctly converted into a causal network, the result is that one should not stop smoking. This result is demonstrated in Appendix B; by presenting a causal network for the smoking–and–cancer example that is consistent with the constraint that it makes no difference to survival whether one stops smoking or not.

As mentioned in the introduction, the next question is whether the present argument could make a difference for the PD: Will players motivated by backtracking considerations cooperate in a one–shot PD with observability of types? We believe that the answer to this question is "yes". While we must leave the complete argument to another paper, we can at least indicate the direction our argument takes.[28]

One important point in our analysis of NP is that Eve, although she is able to predict Adam, is not a "superior being".[29] Therefore, we can go over to a symmetrical situation where the players' types are mutually observable. Note that this cannot be illustrated with the help of a game tree where predictability of one player by another requires that the predictor is a second mover. In the case of mutual predictability, both players would have to be second movers, which is impossible.

However, the causal network of figure 3 can easily be extended to additionally include a type signal correlated with Eve's motivation package and received by Adam. Both players then have four strategies since they can condition their action on the signal they receive. It can then be shown that a motivation package leading to the strategy of only cooperating with those players that have the same motivation package as oneself is consistent with the Backtracking Principle. This solution is similar to the strategic–

[28]The earliest paper discussing the formal connections between NP and PD seems to be Brams (1975). See Pettit (1988) for a sceptical position concerning this connection. However, we have found no papers that completely deny the relevance of NP for the PD. A formal model of mutual predictability in the PD is analyzed for the first time in Heiner & Schmidtchen (1995). A complete analysis can be found in Heiner, Albert & Schmidtchen (2000).

[29]See Binmore's (1994: esp. 245-6) discussion of NP.

delegation solution to the PD briefly discussed in section 2.4. However, like the one–box solution to the NP developed in this paper, it does not rely on any (irrational) commitment power that somehow overcomes, at the time a decision is actually made, a player's motivation to maximize his or her expected payoff.

References

Albert, Max (1998), *The Logic of Risk and Uncertainty,* Konstanz: unpublished manuscript.

Albert, Max/ **Heiner,** Ronald A. (2000), "Causal networks for decision and game theory: a technical reference", Landau and Fairfax/VA: unpublished manuscript.

Aumann, Robert J. (1987), "Correlated equilibrium as an expression of Bayesian rationality", *Econometrica* 55, 1-18.

Bar–Hillel, Maya/ **Margalit,** Avishai (1972), "Newcomb's paradox revisited," *British Journal for the Philosophy of Science* 23, 295-304.

Binmore, Ken (1993), "De–Bayesing game theory", in: Ken **Binmore**/ Alan **Kirman**/ Piero **Tani** (eds), *Frontiers of Game Theory,* Cambridge/ MA: MIT Press 1993, 321-339.

Binmore, Ken (1994), *Playing Fair. Game Theory and the Social Contract* vol. 1, Cambridge/MA: MIT Press.

Brams, Steven J. (1975), "Newcomb's problem and prisoners' dilemma", *Journal of Conflict Resolution* 19, 596-612.

Campbell, Richmond/ **Sowden,** Lanning (1985), *Paradoxes of Rationality and Cooperation,* Vancouver: University of British Columbia Press.

Carroll, John W. (1994), *Laws of nature,* Cambridge: Cambridge University Press.

Dufwenberg, Martin/ **Güth,** Werner (1999), "Indirect evolution vs. strategic delegation: a comparison of two approaches to explaining economic institutions," *European Journal of Political Economy* 15, 281-295.

Frank, Robert H. (1987), "If *homo oeconomicus* could choose his own utility function, would he want one with a conscience?" *American Economic Review* 77, 593-604.

Frank, Robert H. (1988), *Passions Within Reason,* New York and London: Norton.

Fudenberg, Drew/ **Tirole,** Jean (1991), *Game Theory,* Cambridge/MA: MIT Press.

Gibbard, Alan/ **Harper,** William L. (1978), "Counterfactuals and two kinds of expected utility," in: C.A. **Hooker**/ J.J. **Leach**/ E.F. **Mc-**

Clennen (eds), *Foundations and Applications of Decision Theory* vol. 1, Dordrecht: Reidel 1978, 125-162.

Güth, Werner/ **Kliemt,** Hartmut (1994), "Competition or co-operation: on the evolutionary economics of trust, exploitation and moral attitudes," *Metroeconomica* 45, 155-187.

Güth, Werner/ **Kliemt,** Hartmut (1998), "The Indirect Evolutionary Approach," *Rationality and Society* 10, 377-399.

Heiner, Ron A./ **Schmidtchen,** Dieter (1995), "Rational cooperation in one–shot simultaneous PD–situations", Fairfax/VA and Saarbrücken: unpublished manuscript.

Heiner, Ronald A./ **Albert,** Max/ **Schmidtchen,** Dieter (2000), "Rational contingent cooperation in the one–shot Prisoner's Dilemma," Fairfax/VA, Landau, and Saarbrücken: unpublished manuscript.

Jackson, Frank/ **Pargetter,** Robert (1985), "Where the tickle defense goes wrong", in: **Campbell** & **Sowden** (1985), 213-219.

Levi, Isaac (1985), "Smoking, causes, and lung cancer", in: **Campbell** & **Sowden** (1985), 234-247.

Nozick, Robert (1969), "Newcomb's Problem and two principles of choice", in: Nicholas **Rescher** et al., *Essays in Honor of C. G. Hempel,* Dordrecht: Reidel 1969, 114-146.

Nozick, Robert (1993), *The Nature of Rationality,* Princeton: Princeton University Press.

Pearl, Judea (2000), *Causality. Models, Reasoning, and Inference,* Cambridge: Cambridge University Press.

Pettit, Philip (1988), "The prisoner's dilemma is an unexploitable Newcomb problem," *Synthese* 76, 123-134.

Pope, Robin E. (1991), "The delusion of certainty in Savage's sure–thing principle", *Journal of Economic Psychology* 12, 209-241.

Rubinstein, Ariel (1998), *Modeling Bounded Rationality,* Cambridge/Ma: MIT Press.

Savage, Leonard J. (1954), *The Foundations of Statistics,* New York: Wiley.

Selten, Reinhard/ **Leopold,** Ulrike (1982), "Subjunctive conditionals in decision and game theory", in: Wolfgang **Stegmüller**/ Wolfgang **Balzer**/ Wolfgang **Spohn** (eds), *Philosophy of Economics,* vol. 2, Berlin: Springer 1982, 191-200.

Weibull, Jörgen W. (1995), *Evolutionary Game Theory,* Cambridge/MA: MIT Press.

A A State–Space Analysis

In game theory, strategy choices of players are "states of nature" for other players. Beliefs about these strategy choices are, however, nevertheless endogenous, although not from the perspective of the players: probability assignments to strategy choices are determined by equilibrium conditions postulating rational expectations (even for parts of the game tree that are never reached in equilibrium). How players learn to have rational expectations is not considered in the traditional approach. The same goes for the usual assumption that the rules of the game and the rationality of the players are common knowledge, an assumption that in general is neither necessary nor sufficient for equilibrium play. Therefore, we assume rational expectations and common knowledge without any reference to learning processes. Extensions of the present approach covering learning are of course possible.

First of all, consider the Selten–Leopold version of NP from Adam's perspective (figure 5). The relevant states are Eve's decisions (m or n). We do not bother with the derivation of the NM utility function because only the coherence of beliefs is contentious; thus, we assume that the consequences are already described in terms of utilities. Moreover, we consider only the two actions (acts, in Savage's terminology) actually available to Adam (r or g).

A	m	n
r	2	0
g	3	0

Figure 5: State Space, Act(ion)s and Consequences for the Selten–Leopold version of NP

The consideration of more acts is only necessary if one wants to derive unique beliefs (a unique subjective probability distribution) from the preferences over acts. Given the restriction to two acts, it is only possible to check whether certain beliefs are consistent with the preference ordering on the set of acts. Game theory imposes equilibrium beliefs, namely, a subjective probability of 1 for Eve choosing n. These beliefs are consistent with Adam preferring g over r. In this case, this is trivially so since g is better than r no matter what the beliefs might be.

We now show that the same kind of analysis is feasible for a backtracker on the basis of figure 3, where we assume that only material payoffs are relevant.

We need some logical notation to introduce states of nature as they are described by figure 3. Let s be any statement of fact. Let $\Diamond s$ denote the

statement that the fact described by s is causally possible. Let $\Box s$ denote the statement that the fact described by s is causally (or, as an extreme case, logically) necessary. Let $s \rightarrow t$ denote the material implication "if s, then t", a statement that is false iff s is true and t is false. Moreover, let $\wedge$ denote "and". We use these symbols to define the operator $\Box\!\!\rightarrow$: $s \Box\!\!\rightarrow t$ iff $\Diamond s \wedge \Box(s \rightarrow t)$. This means: Given the causal structure of the situation in question, s is possible and, necessarily, if s is realized, t must also hold.[30]

Let r, g, m and n stand for statements describing the corresponding actions by Adam or Eve, respectively. Figure 3 implies $(r \Box\!\!\rightarrow m) \wedge (g \Box\!\!\rightarrow n)$, meaning that, given the causal structure displayed, both r and g are possible, and, necessarily, if r (g) is realized, m (n) must hold. This statement is a state of nature in the Bayesian analysis of NP. Notice that states of nature are causal connections between different events, rather than simply the events themselves. This also means that such a (causal–connection) type of state $(r \Box\!\!\rightarrow m) \wedge (g \Box\!\!\rightarrow n)$ holds independently of whatever Adam might do. However, Adam's action interacts with certain causal connections, thereby producing certain consequences. This, however, is not different if we formalize a situation where Adam presses a button: depending on whether he presses the button or not, different things may happen. The difference to a button–pressing example is that in NP the consequences m or n are not causal consequences of Adam's actions. Hence, the analysis violates the Disconnection Principle; the causal connections are derived by backtracking. Nevertheless, a Bayesian state–space analysis is possible, and from a logical point of view, the states are not more problematic than in a button–pressing example.

Viewed from Adam's perspective, figure 3 illustrates a single state of nature in this sense, namely, $(r \Box\!\!\rightarrow m) \wedge (g \Box\!\!\rightarrow n)$, the state of nature corresponding to the equilibrium. We can rewrite this statement, which is a reduced form, by mentioning the type signal explicitly: $(r \Box\!\!\rightarrow x) \wedge (g \Box\!\!\rightarrow y) \wedge (x \Box\!\!\rightarrow m) \wedge (y \Box\!\!\rightarrow n)$. We now explicitly see that this state assumes equilibrium reactions of Eve. Let us introduce another state of nature (inconsistent with the diagram) with non–equilibrium reactions of Eve: $(r \Box\!\!\rightarrow x) \wedge (g \Box\!\!\rightarrow y) \wedge (x \Box\!\!\rightarrow n) \wedge (y \Box\!\!\rightarrow m)$. Both states of nature can again be displayed in a state–space diagram (figure 6).

A preference of Adam for r over g is consistent with him assigning probability 1 to the first state of nature. If we wanted to derive unique beliefs from preferences over acts, we would have to introduce further irrelevant states.

[30]The system of modal logic assumed here is called S5 (definition of $\Box\!\!\rightarrow$ added). In our view, the subjunctive and counterfactual conditionals in decision and game theory are statements like $s \Box\!\!\rightarrow t$; see Albert (1998), who also extends S5 to include probabilistic statements.

		$(r \,\square\!\!\rightarrow x) \wedge (g \,\square\!\!\rightarrow y) \wedge (x \,\square\!\!\rightarrow m) \wedge (y \,\square\!\!\rightarrow n)$	$(r \,\square\!\!\rightarrow x) \wedge (g \,\square\!\!\rightarrow y) \wedge (x \,\square\!\!\rightarrow n) \wedge (y \,\square\!\!\rightarrow m)$
A	r	2	0
	g	1	3

Figure 6: State Space, Act(ion)s and Consequences for a Backtracker's Evaluation of Figure 3

Figure 3 already tells us which acts are possible, and it is these acts that appear in the states of nature. Hence, one might feel that it is not possible in this analysis to introduce further acts that are in fact not available to Adam. There are ways around this difficulty: the decision problem can always be blown up by introducing irrelevant acts and states of nature until all requirements for Savage's (1954) analysis are fulfilled.[31] Yet even if enlarging the set of acts were impossible, it would not be important: Acts beyond those actually available in a certain situation are only needed to ensure that the NM utility function needed for the expected–utility representation is unique. The uniqueness results of Savage (1954) have nothing to do with rationality.

B Smoking and Cancer

Fisher's assumption was that there is a genetic common cause for both, the preference for smoking and the tendency to get lung cancer. It is claimed that the causal structure of Fisher's smoking–and–cancer problem and of NP are the same and that therefore the conclusions should be the same.

There is no agreement on this argument. In our view, the first step to resolve the issue is to find a causal network that actually represents Fisher's explanation of the data. Let us consider an extreme version. Assume that there are two genetic types. The genes of the first type (g) have two effects: they cause the person to develop cancer (c) with certainty, and they cause the person to enjoy smoking. Thus, the cancer gene also causes a person to have a certain preferences (in the sense of an NM utility function) that would make smoking (s) the dominant choice, no matter whether the person

[31] Figure 3 can be made consistent with the assumption that infinitely many acts are possible. For instance, Adam may wave his hand, scratch his head, and so on. All these actions can be introduced as modifiers: scratching one's head and taking both boxes (g_1) is a different act from waving one's hand and taking both boxes (g_2). Of course, in a suitably modified version of figure 3, all these variants of g (g_1, g_2, and so on) would have the same consequences. Suitable motivations can also be introduced. These additional acts might then have different consequences under further (actually irrelevant) states of nature.

develops cancer or not. For the other genetic type (g'), the genes cause the person to stay healthy and not to enjoy smoking. Thus, for the second genetic type, it is the dominant choice not to smoke (n), whether the person develops cancer or not.

So far we have described only the preferences (NM utility functions) of the two types. The causal setup implies that all smokers get cancer. Some persons think that the correlation provides a good reason not to smoke. Of course, for those who do not enjoy smoking anyway the correlation is irrelevant. However, some of those who do enjoy smoking might decide to abstain from smoking because of this correlation between smoking and cancer. Hence, there are four kinds of motivation packages. Some people enjoy smoking and do not think that the correlation provides a reason to abstain (package m). Another group enjoys smoking but abstains in view of the correlation (package m'). A third group does not enjoy smoking anyway but concurs with the first group that the correlation provides no reason to abstain (package $\bar{m}$). A fourth group also does not enjoy smoking but thinks that the correlation provides a further reason not to smoke (package $\bar{m}'$).

The criticism of the one–box solution runs as follows. In the smoking–and–cancer example, it is not rational to view the correlation as a reason not to smoke. It is claimed that the same considerations leading to the one–box solution in NP would lead those who enjoy smoking to stop smoking in the smoking–and–cancer example. Thus, it is claimed that the package m' describes decision makers who employ backtracking in order to justify their decisions. It would follow, then, that backtracking, which is behind the one–box solution in NP, is irrational because it leads to an irrational decision to stop smoking in the case of motivation package m'.

We consider a causal network (cf. figure 7) in order to demonstrate that this criticism is unfounded.

The network describes the extreme case described above, with the four different motivation packages. However, the network implies that backtracking is not consistent with the package m', as shown next.

Assume that Adam knows that he is at A_2, that is, he is equipped with motivation package m' and is going to actually choose n. If he hypothetically assumes that he chose s instead, backtracking leads him to conclude that the nodes A_1, m, N_2, t–node 1, g, N_1 and Ω have already been reached; therefore, c is going to be reached. If, on the other hand, he backtracks from n, *on the LHS, where he actually is,* the only difference is that backtracking finds A_2 instead of A_1 and m' instead of m. Therefore, c is predicted for both actions. Since c is therefore given, expected–payoff maximization leads to s according to the payoffs for the LHS given in the lower left corner of figure 7. Hence, it is impossible that motivation package m' involves backtracking

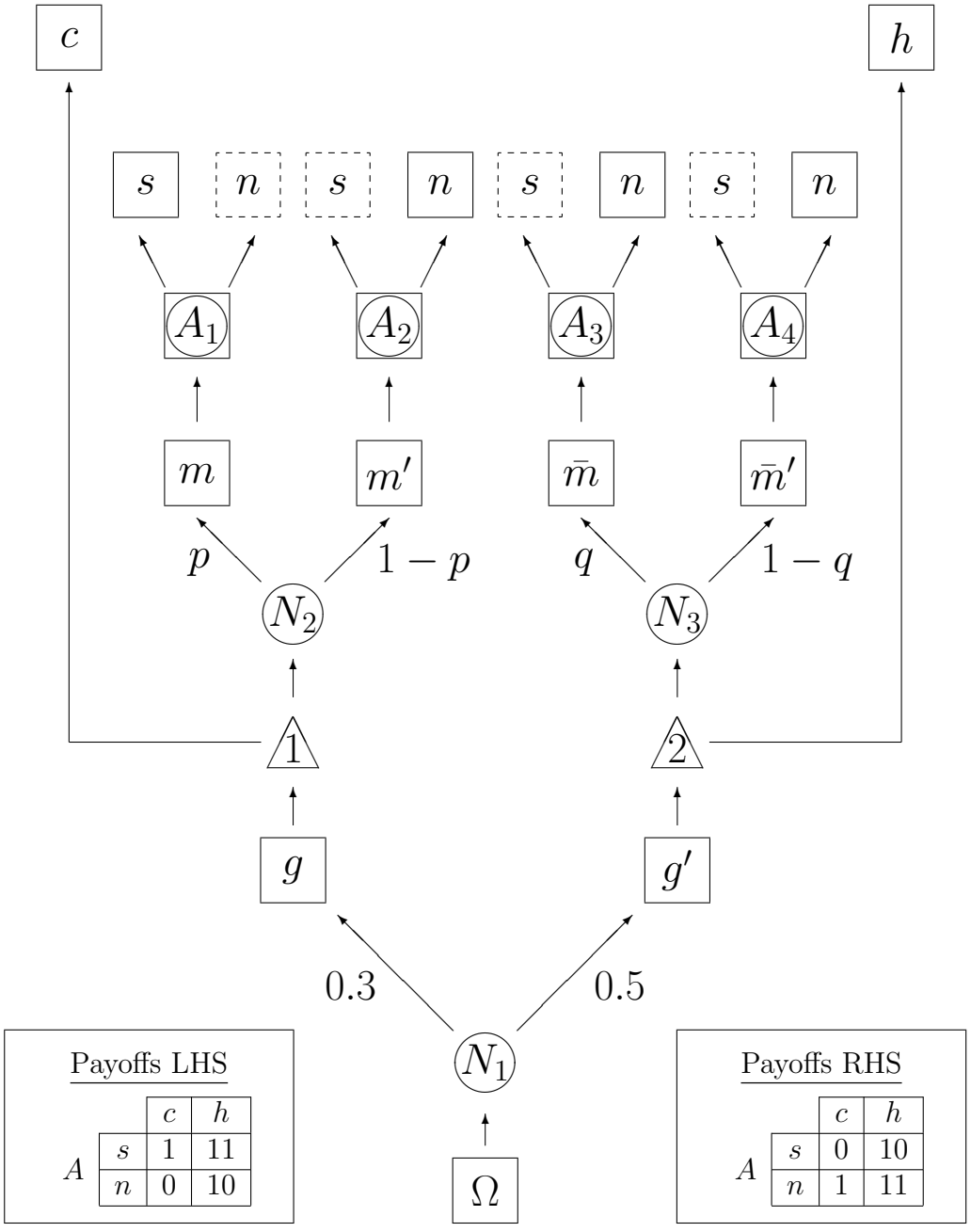

Figure 7: The Smoking–and–Cancer Problem

because backtracking recommends s and not n.[32]

Only motivation package m is consistent with backtracking, where it is moreover required that Adam compares s with n on the LHS. Such a comparison seems reasonable since Adam knows that he enjoys smoking. However, whether any of the other motivation packages is consistent with backtracking is not the issue here. The point is that the decision to stop smoking even though one enjoys it is *not* analogous to the one–box solution of NP.

Thus, the causal network not only makes sense as a simplified extreme version of the smoking–and–cancer example; but in addition, it also implies that the problematic motivation package m' is necessarily *inconsistent* with maximizing expected payoffs on the basis of backtracking.

[32]This is our version of the so–called "tickle defense" of the one–box solution. Cf. Jackson & Pargetter (1985) for references on this line of argument and a critical discussion. Note that we have applied the Backtracking Principle along the more general lines indicated in fn 25 on p. 24 above.