



Notes on the concept and properties of projected equilibrium

Jean-Pierre Dupuy

1. Rationality and transparency

1.1. Over the recent decades researchers working on the foundations of Rational Choice Theory [RCT] have tended to weaken the demands of rationality with two purposes in mind:

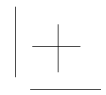
- to bring RCT closer to the psychology of the human mind, which is finite, partially opaque to itself and others, and prone to cognitive errors or illusions;

- to solve a number of paradoxes that seem to undermine the foundations of RCT.

The research that led to the concepts of projected time and projected equilibrium aims in the opposite direction. It is about philosophy, not psychology. It does not seek to describe, explain, or predict behavior. Its ambition is to ground or legitimate normative ethics. Furthermore, it reveals that the paradoxes of RCT do not stem from positing too much rationality, but just the opposite.

1.2. The first demand of rationality is transparency. In economics, game theory, and RCT this demand has usually taken the form of such assumptions as perfect information, perfect forecast, rational expectations, and





the like. The most stringent form corresponds to the following intuition: in a perfectly transparent world, each agent would have the same knowledge of the world as an external omniscient spectator, this fact being common knowledge among the agents. In particular, the agents' foreknowledge does not stem from their having miraculously access to an independent future. If they know the future, it is to the extent that they can compute it, along with the outside modeler, as the fixed point of a reflexive operator: they react to their knowledge of the future, and their knowledge of the others' knowledge of the future, and these reactions jointly bring about the future in question.

The so-called Backward Induction Paradox [BIP] can be interpreted as revealing that orthodox RCT is not, and cannot be made perfectly transparent to itself in the previous sense. There exist situations of interactions - games of the Take-Or-Leave [TOL] form in particular - for which it is logically contradictory to posit that the theory is common knowledge among the players¹. One can characterize the "solutions" usually put forward to this paradox by saying that they commit a metaphysical sleight-of-hand: total transparency is self-contradictory? Never mind since it does not exist in reality. It could very well be the case that although perfect information is unattainable by a finite mind, it is not after all self-contradictory if carefully reformulated, and can serve as a

¹ See Ph. J. Reny, "Common Belief and the Theory of Games with Perfect Information", *Journal of Economic Theory*, 1992; and "Rationality in Extensive-Form Games", *Journal of Economic Perspectives*, vol. 6, no 4, Fall 1992, pp. 103-118.





regulatory ideal, a "horizon" in the Kantian sense. That's the route I have taken.

2. Perfect Foreknowledge and Free Will

2.1. I believe I have shown that the concept of foreknowledge necessary to render total transparency non self-contradictory is the one theologians call *essential* foreknowledge, i.e. foreknowledge in all possible worlds². Essential foreknowledge implies not only the correct prediction of what an agent is going to do, but also the correct prediction of what he would do if he were to act otherwise. I am indebted here to Alvin Plantinga and his solution to Newcomb's paradox. Plantinga treats the Newcomb problem as a challenge to the conventional solutions to the age-old problem of compatibilism, that is, the compatibility between foreknowledge and free will.

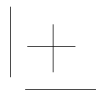
I will briefly review in turn Ockham's solution, Plantinga's demonstration of the inadequacy of the latter due to Newcomb's paradox, and my demonstration of the inadequacy of Plantinga's alternative solution due to the BIP.

2.2. Ockham's Way Out

Let's conventionally call a predictor with essential foreknowledge "God". If God existed at time t_1 and

² Jean-Pierre Dupuy, "Philosophical Foundations of a New Concept of Equilibrium in the Social Sciences: Projected Equilibrium", *Philosophical Studies*, **100**, 2000, p. 323-345. [PFPE henceforth.]





predicted at t_1 that free agent S would do X at t_2 , then the following relation obtains between two events:

(1) "God existed at time t_1 and predicted at t_1 that free agent S would do X at t_2 " **strictly implies** "S does X at t_2 "

where strict implication means material implication in all possible worlds.

On the other hand, with the same two premises:

(2) There is nothing that S can do at t_2 such that, if he were to do it, God would not have predicted at t_1 that he would do X at t_2 .

(2) expresses the **principle of the fixity of the past**: the past is counterfactually independent of present action.

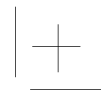
From (1) and (2) one derives:

(3) When an agent acts, there being an essentially omniscient predictor at a given time prior to the time of the action, the agent could not have acted otherwise.

In other words, free will is incompatible with essential foreknowledge.

Ockham's way out of this conclusion consists in posing that the principle of the fixity of the past applies only to events that are truly inscribed in the past in that they constitute *hard facts* about the past. That would not be the case of God's prediction at t_1 if only because that event strictly implies the truth of propositions about future events, such that "S will do X at t_2 " where t_2 is posterior to t_1 .





2.3. Newcomb's Challenge and Plantinga's Way Out

If God is not content with just predicting the future but also changes the world in function of his prediction, like putting or not putting one million dollars in an opaque box, Ockham's way out fails miserably. Contrary to God's prediction, God's action is all but a hard fact about the past.

Plantinga's way out consists in observing that, if God is *essentially* omniscient, then (2) does not obtain. S does X at t₂, all right, and God predicted it at t₁; however:

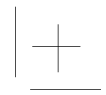
(4) Had S at t₂ taken an action different from X, Y say, God would not have predicted at t₁ that he would do X at t₂, since God would have predicted that he would do Y instead.

In other terms, the principle of the fixity of the past does not apply, not because God's doing is a "soft" fact about the past (his prediction is; the action he takes as a consequence is not), but because free will against the existence of an essentially omniscient predictor entails that the agent is endowed with a **counterfactual power over the past**³.

Newcomb's problem with an essentially omniscient predictor leads to the one-boxer choice. The one-boxer chooses the opaque box and gets 1 million dollars, since the predictor predicted it and put that money inside the box. Had he chosen the two boxes instead, the predictor would have predicted it all the same, and left the opaque

³ Alvin Plantinga, "On Ockham's Way Out", *Faith and Philosophy*, 3, 1986, pp. 235-69.





box empty. The agent's payoff would have been one thousand dollars only⁴.

The dominant-strategy principle (supported by the vast majority of Rational Choice theorists) objects to the one-boxer choice that it rests on an inconceivable *causal* power over the past. Plantinga rejoins that there is no need to posit such a power: a counterfactual power suffices, and it is the logical consequence of compatibilism.

2.4. The Challenge of the Backward Induction Paradox and Projected Time

The BIP proves the inadequacy of Plantinga's way out. There are cases in which the agent's *counterfactual* power over the past *causally* prevents him from taking an action. Such is the essence of the BIP as I believe I have shown⁵.

⁴ See Jean-Pierre Dupuy, "Counterfactual consequences", paper presented at the Workshop on Rationality and Change, Cambridge, UK, 6-8 September 2006.

⁵ **PFPE**.





What is it rational, then, for Peter to play at 1? If he were to play C, given (7) and (5), he would get +1; if he were to play D, he would get 0. Therefore, being rational, he chooses to play C.

Mary's counterfactual power over the past, as illustrated by the disjunction between (5) and (6), seems to have vanished into thin air, along with her free will, since she actually *cannot* choose to play D. What is the nature of this impossibility? Is there a way to save free will against essential foreknowledge?

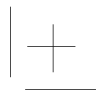
The solution I have proposed is the following. *Before* Mary takes action, she *does* have the choice between C and D. If choosing D is a possibility, it is because as long as Mary has not taken action, her past – here, Peter's choice – is as yet indeterminate (*unbestimmt*). When Mary acts, her choice determines her past. Were she to choose D, she would be prevented from acting. It seems as if she never could choose D, but this impossibility is only retrospective.

What is being jettisoned here is not only the principle of the fixity of the past, but also the **principle of the reality of the past**.

Once Mary takes action, furthermore, it turns out that she could never have acted otherwise – although before taking action, it was true that she could have acted otherwise. The future is necessary but not before it occurs. Once it occurs, the future appears to be fixed, i.e. counterfactually independent of past action.

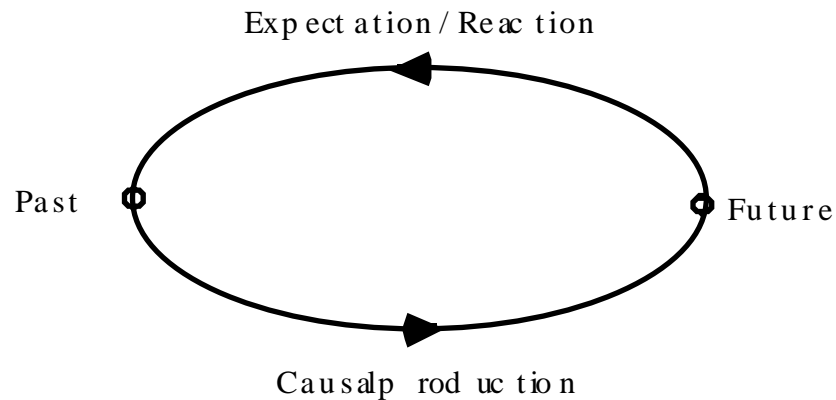
The **indeterminacy of the past** as long as action has not been performed along with the **fixity of the future** once action is taken serve to define a metaphysics of temporality which I have dubbed "**Projected time**."





3. *Projected Time and Two Principles of Choice*

3.1. Since it is the future, rather than the past, which is fixed, the determination of the future is the key problem in projected time. As explained at the outset, this determination is not a prediction in the usual sense; it is rather the computation of a certain fixed point for the following operator that links causally the future to the past, and counterfactually the past to the future:



The theory that leads to the determination of the future as fixed point along with the past that corresponds to the anticipation of it, is perfectly transparent to itself since the theory takes into account that the agents know it and form their predictions and actions according to it. There is no discrepancy whatsoever between the modeler's





computation and the ones performed by the agents themselves.

3.2. Two principles of choice

It may happen that there exist more than one fixed point for the loop linking the future and the past. For instance, in the case of the assurance game, two fixed points exist: [Peter, C; Mary, C], as shown, but we must not forget [Peter, D] which, having no past, cannot be prevented by it. More precisely, there is no way that Peter, having the hand at 1, might be deprived of it as a (counterfactual) consequence of his choice. The future must meet another condition, not reducible to the fixed point, or closure, condition. The assurance game shows what it is: an agent such as Peter at 1, having to choose between two or more actions that are not causally prevented by the past they bring about, chooses the one he prefers.

From now on, we will resort to the terminology of *preemption*. The two principles of choice can thus be summarized as follows:

P1: The actions that are preempted by the past they bring about cannot be chosen.

P2: Between two or several actions that are not preempted by the past they bring about, an agent chooses the one he prefers.

Together, P1 and P2 serve to determine a concept of equilibrium, which I have dubbed **Projected Equilibrium** [PE].





4. *Projected Equilibrium in Extensive-Form Games*

4.1. Although the concepts of projected time and projected equilibrium have relevance in a broad variety of domains⁷, it is mainly in the case of games put in extensive form that we have systematically set out to formalize them and study their properties.

My work published in **PFPE** led me to surmise that the PE always exists, is unique, and is always a Pareto-optimum among the outcomes of the game – making of it, and of the metaphysics of projected time that supports it, the incarnation of a superior form of rationality. This conjecture was later formalized and demonstrated by Ghislain Fourny at Stanford, in the spring of 2004⁸. Fourny's proof was then taken up and made more concise by Stéphane Reiche, equally at Stanford during the spring of 2006⁹.

Fourny's and Reiche's demonstrations are purely algorithmic, which makes them, we hope, acceptable by game theorists or rational choice theorists for whom "metaphysics" may sound like a dirty word. A price had to

⁷ See my *Pour un catastrophisme éclairé*, Paris, Seuil, 2002. See also Jean-Pierre Dupuy, "Two temporalities, two rationalities: a new look at Newcomb's paradox", in P. Bourguin et B. Walliser (eds.), *Economics and Cognitive Science*, Pergamon, 1992, p. 191-220; Jean-Pierre Dupuy, «Common knowledge, common sense», *Theory and Decision*, 27, 1989, p. 37-62. Jean-Pierre Dupuy (ed.), *Self-deception and Paradoxes of Rationality*, C.S.L.I. Publications, Stanford University, 1998.

⁸ Ghislain Fourny, "Equilibrium of Perfect Prediction for Games in Extensive Form, without Indifference", Stanford University, Spring 2004; Publ. Ecole Polytechnique, Paris, April 7, 2005.

⁹ Stéphane Reiche, "Mathematical Foundations of Projected Equilibrium", Stanford University, Spring 2006; Publ. Ecole Polytechnique, Paris, September 2006.





be paid to achieve that – and this is already visible in the case of the assurance game treated above: what is perfectly rigorous in counterfactual reasoning sounds at times as mere sleight-of-hands in algorithmic language.

In view of future publications, I have tried in what follows **a)** to rephrase Fourny's and Reiche's proofs as simply as possible, using everyday's language, and wielding Ockham's razor in a merciless way; **b)** to relate as systematically as possible the algorithm with its metaphysical underpinnings.

Unfortunately, I have only been able to achieve this task in the two special cases by which Fourny started his search for a general proof: the Take-or-Leave [TOL] games and the simplest case of tree-form game. Although those two cases are of utmost importance in themselves, it remains to be seen whether the same kind of work can be carried out in the general case.

4.2. Fourny's Theory of TOL-Games

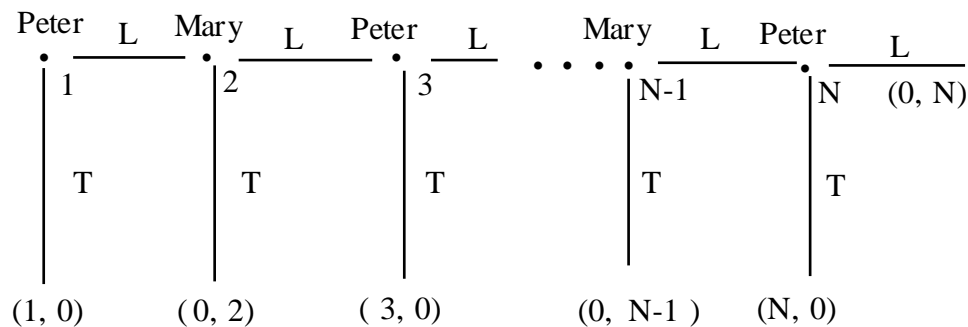
A 2-player TOL-game is such that at every node except the last, the player who has the hand either takes – in which case an outcome is reached that gives each player a payoff – or leaves, i.e. gives the hand to the other player. The player who plays last can only take¹⁰. A 1-to-1 relation exists between the set of nodes and the set of outcomes, as well as between the latter and the set of paths. The chronology of the nodes extends to the outcomes and to the paths.

¹⁰ I am using the phrase "TOL-game" to characterize any game that meets the definition. In the literature, the phrase has been used to designate a subset of such games, defined by a progression of payoffs such as the one shown in the illustration that follows.





For instance:



Extensive-form tree diagram of a TOL-Game, with N odd

4.2.1. Any node either is preempted by the past it brings about, or retains the capacity to preempt.

Proof: <The player who has the hand at 1 takes> cannot be preempted. This is the ground on which the whole proof stands.

The general rule is the following. Let us call J_t the player who has the hand at t . A node/outcome/path t is preempted **iff** a node s previous to it gives J_s a higher payoff than at t without itself being preempted. Since 1 is not preempted, a straightforward forward induction determines step by step which nodes are preempted and which are not.

Along that progression, the players raise their preemptive claim – that is, the payoff under which they





refuse to go by use of preemption – whenever they reach a node at which they play and which grants them a higher payoff (were they to take), unless this node deteriorates the other player's situation to the point that he/she would have preempted it were the player to take.

In the illustration above, one verifies immediately that every [Mary, T] is preempted and that Peter raises his preemptive claim whenever he has the hand again.

By construction, the nodes/outcomes/paths that are not preempted are fixed points of the loop linking the future and the past in projected time.

4.2.2 Any node/outcome/path that is not preempted (i.e. any fixed point) is Pareto-optimal in the set of all the outcomes *previous* to it, itself included.

Proof: let's suppose that \mathbf{t} is a fixed point and that $\mathbf{s} < \mathbf{t}$ is a Pareto-improvement on \mathbf{t} . Since \mathbf{t} is not preempted, and \mathbf{J}_s gets more at \mathbf{s} than at \mathbf{t} , \mathbf{s} must be preempted. Let's call \mathbf{u} the (non-preempted) node that preempts \mathbf{s} , with $\mathbf{u} < \mathbf{s}$. \mathbf{J}_u gets more at \mathbf{u} than at \mathbf{s} (preemption) and more at \mathbf{s} than at \mathbf{t} (Pareto-improvement). Therefore, \mathbf{J}_u getting more at \mathbf{u} than at \mathbf{t} , \mathbf{t} is preempted (by \mathbf{u}), contrary to the assumption.

4.2.3. If there exists a Pareto-improvement on a non-preempted node/outcome/path (i.e. fixed point) among the nodes that come *later*, this fixed point cannot be an equilibrium.





Proof: \mathbf{t} is a fixed point and $\mathbf{v} > \mathbf{t}$ is a Pareto-improvement on \mathbf{t} . Two cases must be considered.

Either \mathbf{v} itself is a fixed point, in which case, according to Principle **P2**, \mathbf{J}_t chooses \mathbf{v} over \mathbf{t} : \mathbf{t} is not an equilibrium.

Or \mathbf{v} is not a fixed point, in which case there is a fixed point \mathbf{w} , with $\mathbf{w} < \mathbf{v}$, which preempts \mathbf{v} . \mathbf{J}_w gets more at \mathbf{w} than at \mathbf{v} (preemption) and more at \mathbf{v} than at \mathbf{t} (Pareto-improvement). Therefore \mathbf{J}_w gets more at \mathbf{w} than at \mathbf{t} . If \mathbf{w} were previous to \mathbf{t} , \mathbf{w} would preempt \mathbf{t} , contrary to the assumption. Therefore, $\mathbf{w} > \mathbf{t}$. Since \mathbf{w} is a fixed point, \mathbf{t} does not preempt \mathbf{w} , and \mathbf{J}_t gets less at fixed point \mathbf{t} than he/she gets at fixed point \mathbf{w} . Principle **P2** demands that \mathbf{J}_t choose \mathbf{w} over \mathbf{t} : \mathbf{t} is not an equilibrium.

4.2.4. A Projected Equilibrium is a Pareto-optimum in the set of outcomes

Proof: A PE is a fixed point, therefore it is not Pareto-dominated by any node prior to it [4.2.2.]; it is an equilibrium, therefore it is not Pareto-dominated by any node posterior to it [4.2.3.].

It is essential to note that the two parts of the demonstration appeal to two different principles, **P1** and **P2**.





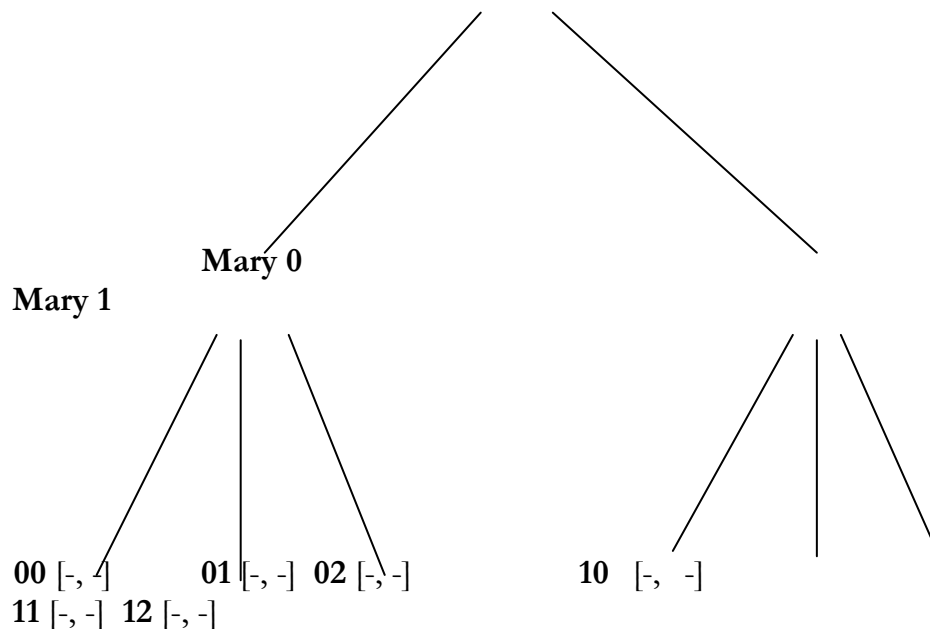
4.2.5. Among the fixed points, the Projected Equilibrium is the latest. It is therefore unique.

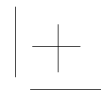
Proof: let's suppose that node \mathbf{t} is an equilibrium and there exists a fixed point $\mathbf{u} > \mathbf{t}$. Since \mathbf{t} does not preempt \mathbf{u} , \mathbf{J}_t gets less at fixed point \mathbf{t} than he/she gets at fixed point \mathbf{u} . Principle **P2** demands that \mathbf{J}_t choose \mathbf{u} over \mathbf{t} : \mathbf{t} is not an equilibrium, contrary to the assumption.

4.3. Towards the General Case

4.3.1. We will consider games of the form:

Peter





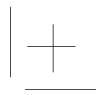
The first player, Peter has the choice between two options, Mary 0 and Mary 1; each one of these, in turn, gets Mary to choose between an indefinite number of options, which all lead to an outcome for the game. Two branches stem from the origin; the depth of the decision-tree is 2. This is enough to usher in a degree of complexity that was not conceivable in the TOL case.

If a 1-to-1 relation still exists between the set of outcomes and the set of paths, neither of them is in a 1-to-1 relation with the set of nodes. What is also lost is a natural chronology on the latter.

In this new framework, the notion of preemption must be complexified. Preemptive actions are no longer actions that lead directly to an outcome. The distinction between preempted outcomes and non-preempted outcomes is no longer an a priori of the search for the PE: preemptions lead to new preemptions that would not have been possible "before", this "before" referring to the sequence of steps that make up the algorithm for the discovery of the PE.

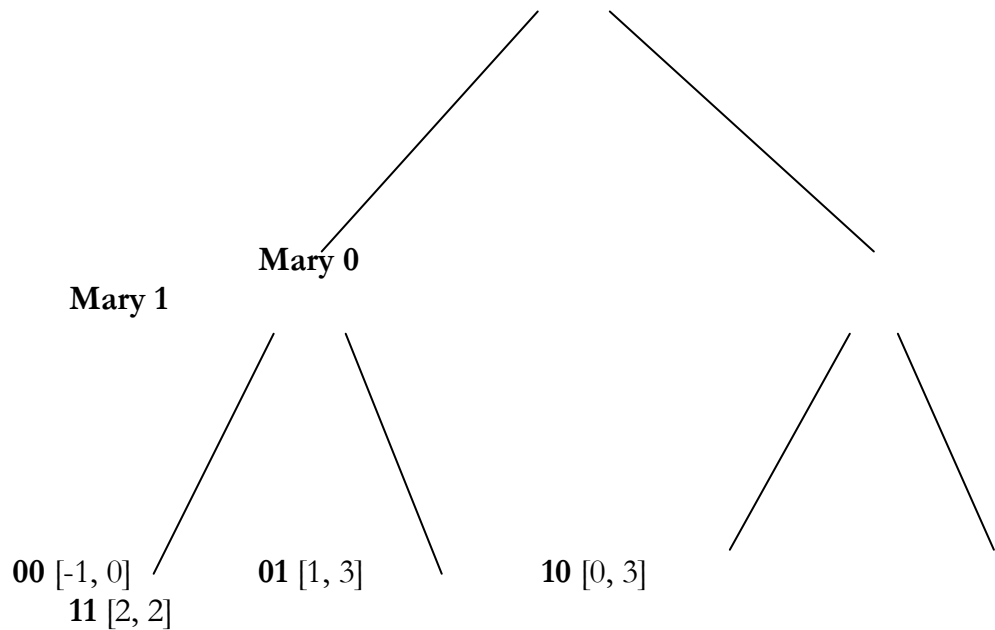
What remains unchanged, though, is the duality of the principles of choice, **P1** and **P2**.





4.3.2 Let us consider the following game:

Peter



There is no hesitation in asserting that [Mary plays **00** at **0**] is preempted by [Peter plays Mary **1**] and that the latter cannot be preempted since it stems from the origin: by playing Mary **1**, Peter secures a minimum payoff of 0 as compared to -1 if Mary plays **00**. We shall say that an outcome is preempted by an action if and only if the agent gets more *whatever* the final outcome of the action than at the outcome in question, provided that the action is not itself preempted by the past it brings about.

Let us then prune the tree from the branch **00** without qualms. We repeat the operation and observe that **10** is





now preempted by [Peter plays Mary **0**], an action that cannot be preempted since it too stems from the origin. Weeding that branch out, we are left with two paths stemming from the origin, **01** and **11**, neither of them being susceptible of being preempted. Principle **P2** demands that Peter choose **11**.

In this particular case, we observe that Mary's payoffs play no role. That is not the case with the subgame perfect equilibrium obtained by backward induction: [Mary **0**, **01**; Mary **1**, **10**; Peter, Mary **0**], which leads to the outcome **01**.

It must be observed that there was no way to conclude that **10** would be preempted (by [Peter plays Mary **0**]) *before* we removed **00**. Preemptions are performed on top of each other. Are we entitled to reasoning that way? The only way to check is by returning to the language of counterfactuals.

Translating back the algorithm into the latter we get:

1. If Mary at 0 were to play 00, Peter at the origin would have played Mary 1, and Mary wouldn't have the hand at 0.
2. Therefore, if Mary were to play at 0, she would play 01, and that wouldn't be preempted by Peter playing Mary 1 at the origin.
3. Therefore, if Peter at the origin were to play Mary 0, the outcome would be 01 and Peter would get **1**.
4. If Mary at 1 were to play 10, Peter at the origin would have played Mary 0, since, because of 3, he would get **1** against **0** at 10. Mary would not then have the hand at 1.





5. Therefore, if Mary were to play at 1, she would play 11, and that wouldn't be preempted by Peter playing Mary 0 at the origin.
6. Therefore, if Peter at the origin were to play Mary 1, the outcome would be 11 and Peter would get 2.
7. Comparing 3 and 6 and applying P2, we come to the conclusion that Peter chooses 11.

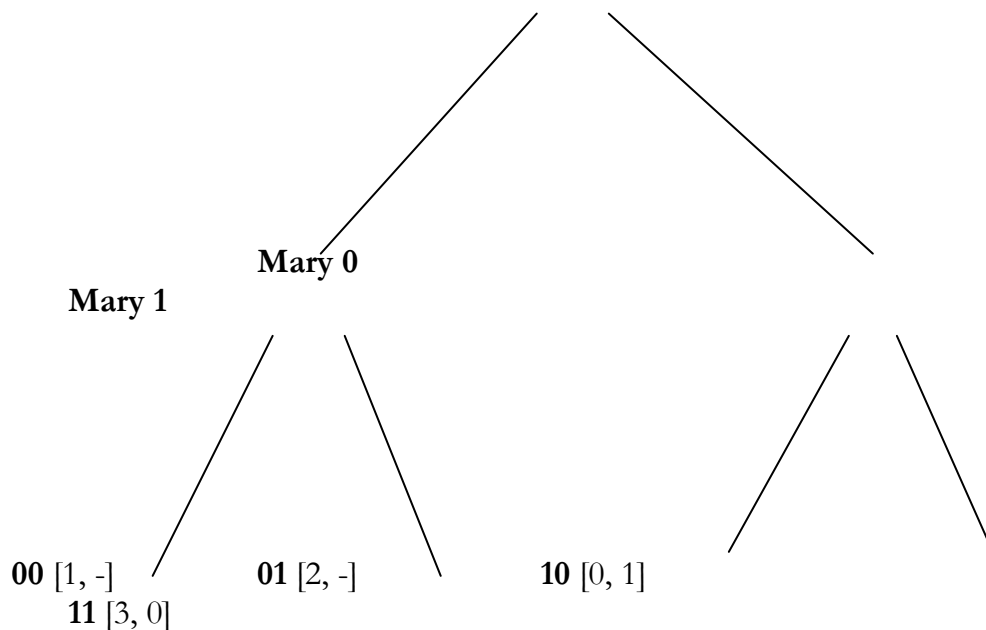
One could be tempted in this case to try to subsume the two principles of choice under a single one and, using the language of preemption, assert instead of 7 that [Mary at 0 plays 01] is "preempted" in turn by [Peter at the origin plays Mary 1, and Mary at 1 plays 11.] However, because of 3, [Mary at 0 plays 01] is equivalent to [Peter at the origin plays Mary 0], and the latter cannot be preempted.





4.3.3. A second case study will wind up convincing us that the algorithmic language, although at times counterintuitive or even outrageous, is perfectly in tune with counterfactual reasoning:

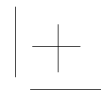
Peter



The algorithm proceeds as follows, using shorthand notations:

10 is preempted by **Mary 0** and removed from the tree.
 Both branches stemming from **Mary 0**, in short **Mary 0**, are preempted by **Mary, 1**.
 Therefore, Peter plays **Mary 1**, and Mary at **1** plays **11**.





[Backward induction leads to Peter playing Mary **0** in order to avoid **10** which Mary would play were she to get the hand at **1**.]

This seems an outrageous sleight-of-hand! [Peter plays Mary 0] looks like a stooge that one gets rid of once it has carried out its dirty work.

Let us check that the algorithm, "scandalous" though it may seem, is supported by a rigorous counterfactual reasoning:

1. If Mary at 1 were to play 10, Peter at the origin would have played Mary 0, and Mary wouldn't have the hand at 1.
2. Therefore, if Mary were to play at 1, she would play 11, and that wouldn't be preempted by Peter playing Mary 0 at the origin.
3. Therefore, if Peter at the origin were to play Mary 1, the outcome would be 11 and Peter would get **3**.
4. If Mary at 0 were to play 00, Peter at the origin would have played Mary 1, since, because of 3, he would get **3** against **1** at 00. Mary would not then have the hand at 0.
5. Therefore, if Mary were to play at 0, she would play 01.
6. If Mary were to play 01 at 0, Peter at the origin would have played Mary 1, since, because of 3, he would get **3** against **2** at 01. Mary would not then get the hand at 0.
7. If Mary were to play at 0, whatever she did, she wouldn't have the hand at 0. Therefore, Mary does not have the hand at 0 at the equilibrium.





8. At the equilibrium, Peter plays Mary 1, and Mary plays 11 at 1.

In this case, there is no need whatsoever to appeal to principle **P2**. The logic of preemption suffices to determine the PE.

4.3.4. Determination of the PE

It is straightforward to generalize the previous algorithm to the whole class of games that we are considering.

At each step of the procedure, having already removed from the tree a number of preempted branches, we determine which remaining branch gives Peter the smallest payoff. This branch is preempted by the node [Mary n , $n = 0$ or 1] from which it does not stem. This operation is iterated until 3 branches are left, 1 on one side, 2 on the other. Two cases are possible:

- a) the node from which the lonely branch stems does not preempt any of the outcomes pertaining to the other side (or, equivalently, the opposite node preempts the lonely branch). The PE is the branch pertaining to the latter that maximizes Mary's payoff [Principle **P2**];
- b) that is not the case, and the node from which the lonely branch stems, which cannot be preempted, preempts at least one of the outcomes pertaining to the other side. The PE is the branch that maximizes Peter's payoff [Principle **P1**].





4.3.5. The PE is a Pareto-optimum

Let us call K the PE which we get at through the algorithm just described, and let's suppose that another outcome, K^* , Pareto-dominates K . Two cases must be considered:

- a) Either K and K^* stem from the same node, say Mary \mathbf{n} . K^* is not preempted by the other node $\underline{\mathbf{n}}$: if it were, K would be also since Peter gets more at K^* than at K (Pareto-improvement). We are in case a) above, and K maximizes Mary's payoff were she to play at \mathbf{n} . However, this contradicts the fact that Mary gets more at K^* than at K (Pareto-improvement).

- b) Or K and K^* stem from two different nodes, respectively \mathbf{n} and $\underline{\mathbf{n}}$.

Either K is a singleton for \mathbf{n} . K being an equilibrium is not preempted by $\underline{\mathbf{n}}$, therefore we are in case b) above. K maximizes Peter's payoff. This contradicts the fact that Peter gets more at K^* than at K (Pareto-improvement).

Or K^* is a singleton for $\underline{\mathbf{n}}$. \mathbf{n} does not preempt K^* since Peter gets less at K than at K^* . We still are in case b) above, and the conclusion is the same.

