COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 1128

THE HANGMAN'S PARADOX AND NEWCOMB'S PARADOX AS PSYCHOLOGICAL GAMES

John Geanakoplos

July 1996

The Hangman's Paradox and Newcomb's Paradox as Psychological Games

John Geanakoplos
Yale University

Abstract

We present a (hopefully) fresh interpretation of the Hangman's Paradox and Newcomb's Paradox by casting the puzzles in the language of modern game theory, instead of in the realm of epistemology. Game theory moves the analysis away from the formal logic of the puzzles toward more practical problems, such as: On what day would the executioner hang the prisoner if he wanted to surprise him as much as possible? How should a surprise test be administered? We argue that both the Hangman's Paradox and Newcomb's Paradox are analogous to a well-known phenomenon in game theory, that giving a player an additional attractive (even dominant) strategy may make him worse off.

In the Hangman's Paradox, the executioner is determined to surprise the prisoner as much as possible, yet he cannot surprise him at all because he cannot commit in advance to a random schedule. The possibility of changing his mind (i.e., the presence of alternative strategies) superficially would seem to help the executioner, but because it changes the expectations of the prisoner, in the end it works dramatically to his disadvantage. In Newcomb's Paradox a man given an extra dominant choice is worse off because it changes God's expectations about what he will do.

Our analysis cannot be couched in terms of the standard Nash framework of games, but must instead be put in a recent extension called psychological games, where payoffs may depend on beliefs as well as on actions.

The Hangman's Paradox is an intriguing puzzle, dating back at least to the 1940s, that appears hard to make rigorous in conventional logic, or conventional game theory. (See Gardner [?] for an introduction to the problem, and O'Connor [?], for the first published discussion.) The story goes something like this:

A judge sentences a man to a hanging one morning in the next week; moreover, he assures the prisoner that the hanging will come as a surprise. The very logical prisoner reasons that were he to reach the sixth night alive, he would know that he was to be hanged the next morning, so that it would be no surprise. Hence he deduces that the executioner could not plan to hang him the last morning. But then he reasons that if he reached the fifth night alive, he would deduce from the fact that he could not be hanged on the last morning, that the hanging must be on the sixth morning. But then he would not be surprised. Hence he deduces from the judge's sentence that he could not be hanged on the sixth morning either. Reasoning the same way back to the first morning, the prisoner realizes he cannot be hanged at all. Secure in his calculations, he sleeps peacefully. But then he is completely surprised and hanged on the fourth morning.

Can the judge's sentence be carried out or not? What indeed will become of the prisoner under these circumstances? There is a large literature which analyzes the Hangman's Paradox in terms of the ambiguity of knowledge and surprise, and the self referential aspect of the judge's sentence. A complete bibliography would include over forty articles (see for example [?] for many references). One interesting conclusion reached by several authors (see Gardner [?], following Scriven [?] and O'Beirne [?]) is that the judge's sentence is self-contradictory, hence the prisoner, no matter how logical he might be, cannot rightfully deduce anything about the hanging day from it, and hence will be surprised whatever day the hanging turns out to be, so that in fact the sentence can be carried out, and the sentence is true and self-contradictory! Whatever its other merits or shortcomings, this analysis does not provide any guide to the date of the hanging. We are given no reason, for example, to think the executioner will choose the hanging day with equal probability, or choose the fourth day for sure.

It is typical of many analyses of the Hangman's Paradox that they come to some conclusion about the logical consistency of the judge's sentence, but they do not in the end decide what will happen to the prisoner: their analysis, as it were, leaves the prisoner hanging.

What I wish to do here is recast the puzzle in the familiar game-theoretic language of incentives and decision-making. Implicitly, I argue, the judge has suggested that the executioner will hang the prisoner and try to surprise the prisoner as much as possible. With a well-defined objective thereby specified for the executioner, a recently invented extension of the theory of games, called "psychological games," allows us to calculate explicitly what the executioner will do and what the prisoner will expect. I am thus able to reach the conclusion that the puzzle is indeed paradoxical, but that it can be handled by a traditional Bayesian theory of knowledge.

In my formulation the paradox does not turn on the ambiguity or self-contradictory nature of the judge's sentence. It stems from the remarkable fact that in the properly

formulated psychological game, the executioner's desire to surprise the prisoner as much as possible prevents him from surprising the prisoner at all! The executioner will carry out the hanging on the very first morning, and this will have been perfectly anticipated by both the prisoner and any one in the audience who had heard the judge and knows that the executioner and the prisoner are rational.

To restore the common sense view that the longer the time horizon given by the judge, the more perfectly can the prisoner be surprised, it is necessary and sufficient to imagine that the executioner also gets a tiny pleasure out of not hanging the prisoner whenever he expects to be hanged. Under this hypothesis, the prisoner would indeed be almost completely surprised when the executioner hanged him, if the horizon were long enough.

In Sections 1, 2, and 4, I remind the reader of the essentials of game theory and psychological games, and then turn back in Section 5 to the hangman's paradox. Along the way in Section 3, I also illustrate Newcomb's paradox via psychological games. There too, I argue that a paradox that is usually posed in terms of God's omniscience and individual decision-making has instead a simple (though still puzzling and perhaps instructive) interpretation as a psychological game.

In Sections 6 and 7, I suggest that the Hangman's Paradox provides a practical lesson beyond the challenge some see in it for logical reasoning. Indeed the Hangman's Paradox arose in the context of a very practical problem: how to give a surprise test over a short period of time to students who know the test has to be given. The lesson is that it is best to precommit to choosing the day randomly (as is the case for recent drug testing of international athletes). If the students doubt that the chances are really random, but believe instead the day will be chosen by an instructor bent on surprising them, then they will assume the test will come on the first possible day, and again the first day after each day the exam is not given, making surprise completely impossible. If the examiner cannot convince the class that the day will be chosen randomly, then he must try to convince them instead that he gets a small pleasure out of not examining them when they expect to be examined. If they believe this pleasure is tiny but nonzero, then over a long enough horizon he can indeed give them the exam on a day when they expect its probability to be arbitrarily close to 0.

1 Game Theory and Nash Equilibrium

John Nash's theory of equilibrium in games [**?**] is based on the following question. Suppose there is a group of N agents. Each agent n must choose a strategy s_n, in some finite set S_n. Each agent n has a utility $u_n : S_1 \times \cdots \times S_N \to R$ that is determined by the actions all the agents take (including of course n himself). What strategy should agent n choose in order to make his utility as high as possible, given that what is optimal for him depends on what the other agents are simultaneously deciding to do?

Nash's solution was to find a designated strategy $(r_1, ..., r_N)$ for each agent such that if each agent was informed of everyone else's designated strategy, then he would

want to play his own designated strategy, that is, for all n in N,

$$u_n(r_1, ..., r_n, ..., r_N) \geq u_n(r_1, ..., s_n, ..., r_N) \text{ for all } s_n \in S_n.$$

I shall always restrict attention to games for which there is a unique Nash equilibrium (and later a unique psychological Nash equilibrium). In such circumstances, the force of Nash's theory is that one may predict that out of the huge number of possible combinations of strategies available to the players, they will collectively choose the Nash equilibrium. To the extent that the theory is sound, it explains why in certain circumstances we should expect one kind of outcome over another.

One can give many interpretations or justifications of Nash equilibrium. One is that it is the only agreement or contract that the agents could make (specifying what each will do) which does not give any agent acting alone an incentive to break. Another interpretation is that if the game is repeated over and over, and agents believe their opponents will do what they have done in the past, the only stationary point of this dynamic process is a Nash equilibrium. Lastly we might suppose that agents thinking hard by themselves about what to do and what to believe, and what others will do and believe and so on in a game with a unique Nash equilibrium would come to the conclusion that the only consistent beliefs they could have and attribute to others are the Nash equilibrium beliefs.

The point of Nash's solution was not that such thinking would lead a group of agents to make wise choices. On the contrary, one consequence of his theory is that each agent's pursuit of his own selfish interest might lead the group to make disastrous choices. Consider for instance the following game (which has some points of similarity with the Prisoner's Dilemma):

		Player 2	
		Left	Right
Player 1	Tough	$1, 1$	$5, 0$
	Gentle	$0, 3$	$4, 4$

Each player has two available strategies, called Tough and Gentle for player 1 and Left and Right for player 2. The two numbers in each cell of the matrix give the utility payoffs to agents 1 and 2, respectively, from each possible joint combination of strategies. The best available choice for the group would be for them to choose (Gentle,Right), resulting in a payoff of 4 to each. But player I would have an incentive to defect from his designated strategy. The only combination of strategies from which neither player would want to defect, and hence the unique Nash equilibrium, is (Tough,Left), which provides the miserable outcome (1, 1). The players, trapped by their own self-interest, end up hurting themselves.

Another very important property of Nash equilibrium is that improvements (reductions) in the payoffs from a strategy may actually hurt (help) a player. Suppose for example that player 1 is educated and acquires a strong distaste for Tough behavior. The payoffs might now be:

		Player 2	
		Left	Right
Player 1	Tough	$-4, 1$	$0, 0$
	Gentle	$0, 3$	$4, 4$

Notice that as compared to the previous game, the payoffs to player 1 are either the same or 5 utiles lower at every possible outcome. Yet it is clear that the only Nash equilibrium of the new game is for the players to choose (Gentle,Right), giving each a payoff of 4. Curiously, by hurting some of his own payoffs, player 1 in the end gets a higher payoff than before. The reason of course is that player 2 no longer has to fear that player 1 will play Tough, and so he is content to play Right, which is much more agreeable to player 1.

In the original game, Tough was a dominant strategy for player 1, in that it was better for player 1 no matter what player 2 chose. Any coherent theory based on individual decision-making (including Nash equilibrium) would have player 1 choose Tough. The second game differs from the first only in that the viability of Tough for player 1 has been destroyed; yet this difference makes player 1 better off. This appears paradoxical since in game 1 it is player 1 alone who chooses between Tough and Gentle. But he cannot commit himself to playing Gentle, i.e., he cannot make it known that he will play Gentle, and hence it is truly rational for him to play Tough, and therefore for player 2 to play Left.

This reasoning has been well understood for many centuries. Both William the Conqueror, in his attack on England, and Cortez, in his attack on Mexico, burned their ships so that their men would have no opportunity of retreat. By eliminating their own options beforehand, which in some circumstances might have been optimal, these generals realized they could affect the behavior of their opponents, and thus gain in the end.[1]

Finally, let us turn the story on its head and note that if in the situation of game 2, player 1 finds a way of dramatically increasing his payoff from Tough to reach the situation of game 1, then he will paradoxically find that he gets a lower payoff in the end, moving down from the Nash equilibrium payoff of 4 in game 2 to the Nash equilibrium payoff of 1 in game 1.

[1]It is not easy to pinpoint when this principle that it might be wise to eliminate one's own attractive options in order to affect one's opponents' choices first became fully understood. All exhortations to bravery, such as the fifth century BC Spartan saying "Come back with your shield or on it," might be said to serve this purpose, if perhaps not completely self-consciously. If we go back still further in time, this strategic awareness becomes more ambiguous. A thousand years earlier, Odysseus supposedly tied himself to the mast before his boat got to the island of the Sirens in order to eliminate a later option. But that was for a different reason. He foresaw that his judgment would later be impaired, and was not worried about affecting the behavior of others. And just a decade before that incident, it seems that the Trojans were in fact confused on this very point since they set out to bum the Achaean ships. Where did they suppose the Achaeans were going to go without their ships, except to try harder to conquer their city?

1.1 Extensive Form Games

Games may involve more than one move by each player. In such "extensive move games," a strategy for a player n is a prescription of what to do in any situation in which n might be called upon to move. As the game unfolds, each player implements more and more of his strategy. For example, a computer program may specify how a player should move no matter what chess position is reached. As the game progresses, the computer appears to "think afresh" with each new position, but in fact it is just implementing the strategy which was a ready programmed into it in the beginning.

Chess is a zero-sum game in the sense that the sum of the payoffs to the two players must always be the same (there can only be one winner). In non-zero sum games such as the Prisoner's Dilemma, the payoffs to the players might sometimes vary together, and at other times be opposed. The Nash equilibria of non-zero sum games are often hard to find. However, for extensive form games in which the players never move simultaneously, a Nash equilibrium can usually be found by a backward induction similar to the kind performed by the prisoner in the Hangman's Paradox. We shall see this shortly.

1.2 Mixed Strategy Nash Equilibrium

Nash (following von Neumann in the case of zero sum games) noted that many games may not have "Nash equilibrium." Consider for example the game of matching pennies:

		Player 2	
		Heads	Tails
Player 1	Heads	$1, -1$	$-1, 1$
	Tails	$-1, 1$	$1, -1$

This is a zero-sum game, since the sum of the payoffs to the two players is always zero, no matter what choices the players make. There is no (pure strategy) Nash equilibrium to this game, for whatever pair of strategies is designated, one of the players will want to defect. By the symmetry of this game, we need only check this for one pair, say (Heads,Tails). At this strategy pair, player 1 gets a bad payoff of -1. But he can defect to Tails and improve his payoff to 1, showing that (Heads,Tails) is not a Nash equilibrium.

Von Neumann pointed out that each player could randomize his choice of strategies. Player 1 would play each strategy i with probability p_i, and player 2 would play each of his strategies j with probability q_j. Given a pair of randomized or "mixed" strategies, one can immediately calculate the probabilistic expected payoff to each of the players. A mixed strategy Nash equilibrium is a designated randomized strategy for each player such that if every player were informed of these designated mixed strategies, then no player could increase his expected payoff by deviating from his designated randomized strategy.

Consider in matching pennies the pair of mixed strategies in which each player n plays Heads with probability 1/2 and Tails with probability 1/2. The expected payoff to each player is then 0. Notice that no player can do better by deviating. If player 1 plays Heads for sure, he still gets an expected payoff of 0, and similarly if he plays Tails for sure. Needless to say, no mixed strategy can do better for player 1 then the best pure strategy. Thus we have a mixed strategy Nash equilibrium.

Nash, extending von Neumann's zero sum theorem to arbitrary games, proved that any game as defined earlier in Section 1 must have a mixed strategy Nash equilibrium. When a game has a unique mixed strategy Nash equilibrium, game theorists often feel comfortable in predicting that the players will behave accordingly, and that the individual incentives thus determine the outcome of their interaction.

1.3 Knowledge and Indifference in Nash Equilibrium

In a mixed strategy Nash equilibrium what can we say player 1 knows about what player 2 will do? Eventually player 2 will choose one strategy. But player 1 cannot know which at the moment he must choose to move. We may suppose that player 1 believes that player 2 will choose his strategies with the probabilities specified by the mixed strategy equilibrium. Thus in the Matching Pennies game, player 1 must assume the probabilities on 2's moves are 1/2, otherwise 1 himself would not be willing to play his designated mixed strategy.

At Nash equilibrium for a game in extensive form, players can use Bayes Law to update their beliefs about which strategies the others are using as the game unfolds. Thus if player 2 is supposed to play each of three strategies with probability 1/3, then after the first observed move of player 2, the other players will be able to revise their beliefs about which strategy player 2 is indeed using. For example, suppose strategy a prescribes Left on every move, while strategy b prescribes Left on the first move and Right on the second move, and strategy c prescribes Right on every move. Then if 2 chooses Left on the first move, the other players can infer that he is using strategies a or b with probability 1/2 each, and hence that the probability is 1/2 that he will continue with Left on his second move. A mixed strategy thus provides a conditional probability of the moves that the agent will make at any point at which he is called upon to move with positive probability.

It is very important to note that at a mixed strategy Nash equilibrium, each player will be indifferent between various alternatives. Thus to player 2 in Matching Pennies, whether he plays Heads or Tails makes no difference to his expected payoff. He might just as well play Heads for sure, or Heads with probability 2/3. Nevertheless, we are bound to assert that player 1 thinks the odds are 1/2 that player 2 will play Heads, for otherwise the Nash equilibrium itself disintegrates.

Indifference can arise even in pure strategy Nash equilibrium (at least for games with infinite strategy spaces). Consider the following bidding game: Each of two players names a real number between 0 and 1. If a player names the strictly highest number, he gets \$1 minus his bid; otherwise he gets \$0. It is easy to see that the unique Nash equilibrium of this bidding game is for both players to bid \$1, so that

each gets nothing. At that equilibrium either one of the players could play anything else without affecting his own payoff. Yet we are right in supposing that the other play knows for sure what will be played.

1.4 Single Decision-Maker Games

It may turn out that the payoffs to all the players in a game do not depend on the strategy choices of players $2, ..., N$. In other words, it may be that player 1's actions alone determine everybody's payoff. In such a one person game, there can be no paradoxical equilibrium. Player 1 will simply choose that strategy which makes his payoff highest. If there is a unique Nash equilibrium, it must be a pure strategy equilibrium.

2 Psychological Games and Psychological Nash Equilibrium

In the theory of games just described, the payoffs to the agents depend on what the players do. But more generally we may wish to consider games in which the payoffs to the agents also depend on what the players think. Thus in the Hangman's Paradox only the executioner acts, but his payoff might be said to depend on what the prisoner thinks, i.e., on how surprised he is, or perhaps more precisely, on how surprised the executioner thinks the prisoner is.

Psychological games (see [**?**]) were invented to take into account the dependence of utility on beliefs (see also [**?**] for a related but distinct idea called information-dependent games). In general, payoffs might depend on what the players think, or what they think others think, and so on all the way up the whole hierarchy of beliefs. We can simplify these vast possibilities by limiting the thinking to beliefs about what players are going to do. Furthermore, in keeping with the spirit of Nash equilibrium, we suppose that in equilibrium everybody knows what strategies (or mixed strategies) every player is using. Hence we can reduce the entire hierarchy of beliefs about beliefs to the first level, namely the common beliefs about what each player will do.

A psychological game (in this simplified form) is thus given by strategy spaces S_n, for each player n from I to N, and by utility functions

$$u_n : S_1 \times \cdots \times S_N \times S_1 \times \cdots \times S_N \to R.$$

The utility u_n depends on the strategies every agent chooses, which are represented by the first N-fold product of the individual strategy spaces, and by the beliefs commonly held by each player over what the (other) players will do. These beliefs are represented by the second N-fold product of individual strategy spaces. For these n, the element in S_n is the belief the other agents have about what agent n will do (which could equally well be viewed as the belief n has about the beliefs the other players have about what he will do).

Formally, a psychological Nash equilibrium specifies a strategy choice $(r_1, ..., r_N)$ for each agent such that if each agent believed every agent would implement his designated strategy, then he would want to play his own designated strategy, that is for all n in N,

$$u_n(r_1, ..., r_n, ..., r_N,\ r_1, ..., r_n, ..., r_N) \geq u_n(r_1, ..., s_n, ..., r_N,\ r_1, ..., r_n, ..., r_N)$$
$$\text{for all } s_n \in S_n.$$

Thus an agent who considers deviating realizes that the other agents will not expect him to deviate, hence he will hold their strategies and their beliefs (about his actions) fixed when he considers the payoff effect of a deviation.

Since the beliefs are assumed to be correct in equilibrium, it does not appear that psychological Nash equilibrium is any different from Nash equilibrium. We shall see, however, that it is.

3 Newcomb's Paradox

The philosopher Newcomb suggested the following parable of faith, decision-making, and free will. God tells a man that he can have a box, or he can have the box and a diamond. God asks the man to have faith, telling him that if He believes he will choose the box, it will contain four diamonds, but if He believes he will choose the diamond and the box, the box will be empty. God never lies, and will not touch the box after that moment. What should the man do?

The man, being very rational, reasons that whatever is in the box, he would only do better taking the diamond as well. So he takes both. But God, being omniscient, knew the rational man would decide to take both, and so placed nothing in the box. God was true to his word, the man apparently acted rationally, yet he ended up with one diamond instead of four.

Newcomb's Paradox has been ably discussed many times (see [**?**], [**?**], and also the entire book [**?**] devoted to the subject). Nevertheless, it may be of some use to consider the concise presentation of the puzzle that can be obtained via the appropriate psychological game. Accordingly, let us examine a one player game where only the man moves, but his payoffs depend on God's beliefs as described below:

		Beliefs	
		Both	Box
Man	Both	1	5
	Box	0	4

No matter what God's beliefs are, the man would rather choose Both, and so maximizing his payoff, as is required in psychological equilibrium, he will choose Both. But in psychological Nash equilibrium, the beliefs are correct. Hence there is exactly one psychological equilibrium, in which the man ends up with only one diamond and an empty box.

In Newcomb's description, the story requires an omniscient being to guarantee the promised outcome no matter what the man chooses. The wide applicability of psychological games, however, stems from the observation that God can often be replaced by a logical being whose beliefs matter and who, by virtue of logic, not divine power, has the correct beliefs, as we shall see in examples below.

By reinterpreting Newcomb's paradox as a psychological game we shift the emphasis from the *logical* conundrum that choosing a dominant strategy may be bad to the familiar game-theoretic paradox we discussed in Section 1 that the availability of an additional attractive (dominant) choice may make a player worse off. Were the man disgusted by the act of choosing Both, he would have chosen the Box and obtained four diamonds. This paradox is even more starkly apparent in Newcomb's story because there is only one player. In Nash equilibrium, in a one decision-maker game, the paradox could not arise. But in psychological equilibrium the beliefs introduce the same possibilities for paradox as the second player does in a conventional Nash game.

4 The Surprise Game and Mixed Strategy Psychological Games

Consider a man who wants to surprise his very rational girlfriend by bringing her either flowers or chocolate, namely the one she doesn't expect. What should he do? We model the situation as a psychological game, with the payoffs below:

		Beliefs	
		Flowers	Chocolate
Man's Strategy	Flowers	0	1
	Chocolate	1	0

Clearly there can be no pure strategy psychological Nash equilibrium. If the man convinces himself he should bring say flowers, then by the same logic she will expect flowers, and he would do better to bring chocolate instead. The woman forms her beliefs before the man appears at her door, just as in Newcomb's paradox God decided whether to fill the box before the man chose. Yet we are also justified in assuming that she will have the correct beliefs about what he will do, even though she has no divine powers, because she can reason.

The absence of a pure strategy psychological Nash equilibrium leads us as before to consider mixed strategy psychological Nash equilibrium. Consider again a general psychological game. If each player n in N independently chooses his strategy via a probability p_n on S_n, and if each other agent believes agent n will choose according to a probability q_n on S_n, then we can straightforwardly calculate each agent's expected utility by multiplying out all the probabilities. A mixed strategy psychological Nash equilibrium is a vector of such probabilities and beliefs such that no agent can increase his expected utility by unilaterally deviating to another probability, and such that the beliefs q_n are equal to the action probabilities p_n.

At a mixed strategy psychological Nash equilibrium, just as at a mixed strategy Nash equilibrium, we can say precisely how "surprised" an agent m is to find that another player n has played strategy i: it is $1 - p_{ni}$. His expected surprise from strategy i is $p_{ni}(i - p_{ni})$.

In the surprise game there is a unique psychological Nash equilibrium; the man randomizes and chooses flowers and chocolate each with probability 1/2. The surprise, and hence his expected payoff, from Flowers and Chocolate when she believes he is so randomizing is the same (namely 1/2), so he is indeed doing as well as he can by randomizing, and her beliefs are indeed correct. The notion of psychological Nash equilibrium allows us to analyze a commonplace but puzzling situation and prescribe the sensible outcome: the man simply cannot completely surprise his girlfriend, but he can partly surprise her, and he chooses the most surprising course of action available.

Note once again that given her beliefs, he might just as well bring flowers for sure, or chocolate for sure, rather than randomizing. But we have seen this over and over again in Nash equilibrium; logical consistency breaks down unless she assumes that he will bring each with probability 1/2.

Psychological games always have (possibly mixed strategy) psychological equilibria (see [**?**]). This is because the beliefs which affect payoffs are taken to be probabilistic, i.e., on a par with actions. One might instead have taken the view that the payoffs should depend only on what agents know or do not know, not on their degree of belief. In that case one would often confront games that did not have equilibria. (This is the situation in [**?**], for instance.) Psychological games thus not only provide a framework for describing certain kinds of situations, they also prescribe, via psychological equilibrium, what will happen.

Even one-player psychological games can have curious equilibria, as the surprise game makes clear, where the only equilibrium is a mixed strategy equilibrium despite the fact there is only one decisionmaker.

5 The Hangman's Paradox

We are now ready to formalize the hangman's paradox as a psychological game in extensive form.

At each of a finite number of time periods $t = 1, ..., T$ the executioner must choose a move, either to Hang the prisoner, or Not to hang him. If at any time $t < T$ the prisoner is hanged, the game ends. Otherwise it proceeds until time T, when it ends whether or not the prisoner has been hanged. At each date t, if the prisoner has lived past $t - 1$, he will have expectations about what is going to happen to him on date t. The payoff at date t to the executioner (who is the only agent to act) will depend on his actions and the beliefs of the prisoner, given in the matrix below. The total payoff to the psychological game is the sum of the payoffs from all the periods 1 through T (or until the point the game ends).

		Beliefs	
		Hang	Not
Executioner	Hang	B	A
	Not	0	0

The judge has ordered the executioner to Hang the prisoner, and so he gets a payoff $B \geq 0$ for carrying out the sentence. He gets a greater payoff $A > B$ if he also surprises the prisoner. (If $B = 0$, then the interpretation is that the executioner gets no payoff unless he hangs the prisoner and surprises him.) In a psychological Nash equilibrium the executioner will by definition strive to make his payoff as high as possible.

5.1 Maximizing Surprise

It would seem a simple matter to surprise the prisoner. For example, suppose the executioner programs a computer or Automaton to flip a fair T-sided coin and hang the prisoner on the day that the coin designates. If the prisoner believed this strategy, the night before the $(t+1)$st day he would assign probability $1/(T-t)$ that he would be hanged later that morning. The expected payoff to the executioner can easily be calculated to be the sum of $[B + (A - B)(1 - 1/t)]$ from $t = 1$ to $t = T$, divided by T, which is just slightly more than $A - (A - B)(\log(T + 1)/T$ For large T this converges to A, so in fact the prisoner is almost completely surprised if the horizon is long enough. For $T = 7$, the payoff is approximately $[B + .63(A - B)]$.

In fact the executioner could design a slightly better Automaton if he could rely on the prisoner believing he would stick to his announced strategy. Let S_{-T} be the maximal payoff the executioner can achieve given that his actions are correctly anticipated by the prisoner, when the horizon is of length T Let P_{-T} be the probability of hanging the prisoner with T periods left to go, conditional on his still being alive. The maximal surprise, for each horizon T, must satisfy the recursive relation:

$$S_{-T} = p_{-T}[p_{-T}B + (I - p_T)A] + (1 - p_{-T})S_{-(T-1)}$$

This can be maximized recursively by setting $p_{-1} = 1$, $S_{-1} = B$, and in general, $p_{-T} = (A - S_{-(T-1)})/2(A - B)$, and $S_{-T} = S_{-(T-1)} + (A - S_{-(T-1)})^2/4(A - B)$. It is easy to see that S_{-T} converges monotonically up to A, and slightly faster than with the random coin flip. For $T = 7$, the conditional probabilities for hanging the prisoner on each day are respectively (.20, .225, .26, .305, .375, .50, 1.0), and the expected payoff is $[B + .64(A - B)]$.

Yet we shall show that in fact the executioner cannot surprise the prisoner at all, i.e., he cannot get a payoff above B. The reason is that the executioner is trapped by his own zealous effort to surprise the prisoner. Suppose for example that the executioner thought that the prisoner believed that he was going to follow the optimal scheme defined above. From the above calculation, we know that the farther from the end, i.e., the higher is T, the lower will be p_{-T}. Hence the executioner could strictly increase his payoff by hanging the prisoner on the very first day. With the

$T = 7$ period horizon, he could countermand his orders to the Automaton, hang the prisoner on the first day, and gain a payoff of $[B+.8(A-B)]$ instead of $[B+.64(A-B)]$. Realizing this, the prisoner would not believe the executioner's announced plan. We now show precisely what the prisoner would believe, and what the hangman would do.

5.2 The Psychological Nash Equilibrium for the Hangman's Game

The charm of the Hangman's paradox is that the executioner does so badly. Were he content with a little surprise with nonzero probability, he could easily arrange it. In fact, as we saw in the last section, with enough periods to work with, he could almost totally surprise the prisoner. Yet his efforts to get still more give rise to the paradox.

Proposition 1 *The hangman's game has a unique psychological Nash equilibrium; in that equilibrium the prisoner is hanged for sure on the first day, the hanging is completely anticipated, and the payoff to the executioner is B. There is no surprise at all.*

Proof of Proposition 1 One proof would be to argue by backward induction exactly as in the statement of the hangman's paradox. We present a shorter proof that does not require backward induction.

Let the hangman's mixed strategy prescribe the hanging on the tth day with conditional probability p_t that is, conditional on the prisoner still being alive on the night of the $(t-1)$st day. Let the corresponding conditional beliefs of all the players be q_t. The payoff for a hanging on date t is $B + (A - B)(1 - q_t)$, which is strictly increasing in $(1 - q_t)$. Let T' be the set of t for which $(1 - q_t)$ is a maximum, i.e., for which q_t is a minimum. We shall prove that $q_t = 1$ for all t in T, for which it suffices to show that $q_t = 1$ for at least one t in T'. Suppose otherwise. Since the prisoner can only be hanged once, and there is no payoff for not hanging him, and since $A > B$, optimality requires the executioner (given the q) to choose $p_t = 1$ for some t in T'. But in psychological equilibrium $p_t = q$, for all t in T. □

Several comments are in order. In the unique psychological Nash equilibrium it is true that the executioner could deviate and hang the prisoner on the third or fourth day without lowering his payoff. But since we proved that $q_t = 1$ for all t (not just $t = 1$), that would come as no surprise to the prisoner. His beliefs are that he will be hanged on the first available day whenever he is alive; if by some miracle he is not hanged on the first day, the prisoner would expect to be hanged on the second day, and so on.[2] Thus postponing the hanging beyond the first day would not help the executioner's payoff.

[2] Many commentators seem to have overlooked that it is perfectly logical for the prisoner to believe he will be hanged the first day and then again the second day, since the relevant opinion about the second day is the one he forms *after* seeing that he was not hanged the first day.

If the executioner is indifferent to which day he does the hanging, then why should the prisoner expect the hanging on the first day, and why should the executioner hang him on the first day? The answer is, because that is the unique psychological Nash equilibrium. Exactly as we were justified in figuring that in matching pennies player 2 was equally likely to play left or right, even though many other mixed strategies would give him the same payoff, or that in the bidding game it could be expected that player 2 would bid \$1 even though any other bid would have given him the same payoff, so we are justified here in presuming that each player knows the other's strategy.

Finally, let us make clear that the paradox can be seen as another instance of the principle of disadvantageous addition of attractive strategies that we described at the end of the introductory section on Nash games. We can do this by comparing the Automaton executioner game to the hangman's game. An Automaton has no choices, and hence effectively one strategy. An executioner who (is motivated) to get pleasure out of surprising the prisoner and is free to change his mind at any moment has many available and attractive strategies. One might naturally suspect that he should be able to increase the surprise beyond what the automaton could achieve in our discussion from Section 5.1. Indeed, if the prisoner continued to believe that the executioner was the same Automaton, then the motivated executioner could increase his payoff, as we saw when $T = 7$, from $B + .64(A - B)$ to $B + .8(A - B)$, by hanging the prisoner on the first day. But the judge not only motivates the executioner to surprise the prisoner, he informs the prisoner that the executioner is trying to surprise him (and has the freedom to choose the hanging whenever he wishes); this changes what the prisoner is prepared to believe. It makes it common knowledge that the game being played is the motivated executioner game, not the automaton executioner game. The executioner knows that the prisoner knows that the executioner is trying to surprise him, and this makes it impossible to surprise the prisoner at all.

5.3 Restoring the Surprise

Common sense suggests that if the horizon is long enough, then it will be possible to hang the prisoner and (almost) completely surprise him, even if the prisoner knows the executioner is trying to surprise him. We now show that this common sense view can be justified if we suppose that the executioner gets some tiny pleasure from surprising the prisoner by not hanging him when he believes he will be hanged. Consider the following payoff matrix in each period t:

		Beliefs	
		Hang	Not
Executioner	Hang	B	A
	Not	ε	0

Suppose now that $0 < \varepsilon < \min[A - B, B)$.

Proposition 2 *The modified Hangman's psychological game defined above has a unique psychological Nash equilibrium for any horizon T For any $\varepsilon > 0$, as T gets larger, the payoff $S(T, \varepsilon)$ to the executioner converges to A. Furthermore, as ε gets smaller, and T gets larger, the payoff not only converges to A, but it comes almost exclusively from surprising the prisoner by hanging him, and in the limit, not at all from surprising him by not hanging him.*

The modified hangman's game thus shows the circumstances under which a willful executioner can adopt a credible strategy that hangs the prisoner for sure, and does so on a day when he rationally expects the probability of hanging to be nearly 0.

Proof of Proposition 2 Suppose that the horizon T is fixed, and consider a psychological Nash equilibrium, i.e., a probability p_t that the executioner will hang the prisoner on date t, assuming that the prisoner is still alive, for each date t. As before, it is convenient to define $p_{-t} = P_{T+1-t}$ by the probability of hanging in the tth period before the end. As before, we also define S_{-t} as the expected payoff to the executioner during the last t periods of the psychological game.

In the last period, $t = T$, the hangman will hang the prisoner for sure, since this is a dominant strategy, and of course the prisoner will expect it. We write that $p_{-1} = 1$, and $S_{-1} = B$.

For $t < T$, in stark contrast to the standard version of the Hangman's Paradox, it can never be optimal to hang the prisoner for sure. If it were, and if this were expected, then the maximal total payoff from that point on would have to be equal to the payoff from hanging the prisoner immediately, namely B, whereas by not hanging him, the executioner could get an immediate $\varepsilon > 0$, and then collect B afterwards by hanging the prisoner in the very next period.

Similarly, for $t < T$ we cannot have that the executioner chooses not to hang the prisoner for sure. If there were such a t, consider the last time it happens $\bar{t}$. From period $\bar{t}$ onward (which, from the last paragraph, has positive probability of being reached), the payoff to the executioner of following his plan is 0 (for period $\bar{t}$) plus whatever he gets afterward. But by our choice of $\bar{t}$, in every period after that the executioner must be planning to hang the prisoner with some positive probability, which, being expected, would give a payoff strictly less than A. By hanging the prisoner instead in period $-t$, when it is completely unexpected, the executioner can get a payoff of A.

Thus we can conclude that for all $t < T$, the executioner must be randomizing, and that therefore

$$S_{-t} = p_{-t}B + (1 - p_{-t})A = p_{-t}\varepsilon + S_{-(t-1)}.$$

From this we deduce that

$$S_{-t} = A - (A - B)p_{-t} \text{ and } p_{-t} = (A - S_{-(t-1)})/(A - B + \varepsilon)$$

and that

$$p_{-t} = (A - B)p_{-(t-1)}/(A - B + \varepsilon) = (A - B)^{t-1}/(A - B + \varepsilon)^{t-1}.$$

Fix an amended hangman's game with very large T. It is evident that at the beginning stages, when $-t$ is large, the probability of hanging the prisoner will be very small, since $\varepsilon > 0$. Hence the expected payoff to the executioner is nearly A. As time progresses it becomes exponentially more likely that the hangman will act, until the last period when the hanging takes place with probability 1.

The executioner can get his payoff of nearly A in two ways: either by chance he hangs the prisoner early when he is expecting it with almost 0 probability, or the hanging comes much later, when the payoff is closer to B, but along the way the executioner picks up nearly $(A - B)$ for all the times he did not hang the prisoner when he attached a small probability to being hanged. We shall now show that for small enough ε, all the payoff actually comes from the first way: the hanging, though still destined to happen, is nearly sure to occur in one of those early periods when it is little expected.

Let the hanging payoff be given by s_{-t}, which is equal to S_{-t} minus the amended payoff from not hanging the prisoner. Let $k = (A - B)/(A - B + \varepsilon)$. Then it is easy to see that s_{-t} must satisfy the recursion

$$\begin{aligned} s_{-t} &= p_{-t}[p_{-t}B + (1 - p_{-t})A] + (1 - p_{-t})s_{-(t-1)} \\ &= k^{t-1}B + (1 - k^{t-1})[k^{t-1}(A - B) + s_{-(t-1)}]. \end{aligned}$$

This is uniquely solved by

$$s_{-t} = B + k(1 - k^{t-1})(A - B).$$

For any fixed $0 < k < 1$ and large t, the payoff from hanging is approximately $B + k(A - B)$. By choosing a very small $\varepsilon > 0$, and hence k nearly 1, and then choosing large t, we see that the payoff is nearly A. $\square$

When ε is chosen large, the executioner gets his payoff of nearly A by setting the probability of hanging so low at the beginning that the actual hanging is very likely to take place toward the end when it is not so unexpected. (This is fine for the executioner, because along the way he is picking up a little bit of payoff for not hanging the prisoner.)

In summary, we have shown that when there is a tiny payoff $\varepsilon > 0$ from not hanging the prisoner when he fully expects to be hanged, the executioner will be led to follow a strategy that hangs the prisoner for sure, and when the number of periods T is very large, nearly completely surprises him. To get any desired expected level of surprise just below complete surprise, however, the horizon T must be much longer that would be required by the Automaton executioner under the scheme defined in Section 5.1.

6 The Surprise Test

The Hangman's Paradox was discovered in the 1940s when the Swedish Broadcasting Company wanted to monitor the preparation of its civil defense units.[3] The SBC recognized that if the agents knew in advance when the test would take place, they could "cram" for it, thereby disguising their true preparedness. Testing is time consuming and costly, and the SBC therefore did not wish to test them often. On the other hand, cramming also takes effort and the agents would do it only in proportion to the probability they expected a test. The obvious solution was to randomly schedule an exam. Of course, given any such posted schedule of probabilities, the agents would adjust their behavior, cramming most when the probability of the exam was highest. The SBC therefore hoped that by putting an examiner in charge who was motivated to catch the agents when they were least prepared, it could improve the reliability of the test beyond that achievable by a fixed (but random) schedule. They saw that the effort seemed to lead to logical difficulties, and the Paradox was born. Our analysis indeed shows that the strategy of putting a human agent motivated to surprise the agents, but bound to give a test sometime in a finite horizon, is bound to fail badly.

The problem of when to test is familiar to any teacher, and also to organizers of athletic competitions who do tests for performance enhancing drugs. In the latter case an athlete already using illegal drugs can mask them by taking auxiliary masking drugs, or by reducing the dosage of the illegal drug he takes, in the period just before the test will take place. These masking efforts are costly to the athlete, however. For example, reducing the dosage of the illegal drug might force him to curtail the most rigorous parts of his training (which it was the purpose of the illegal drug to maintain). This reluctance on the part of the subject (student or athlete) to constantly take efforts to hide his true condition is the reason why it would appear that a surprise test would be successful.

To see why, on the contrary, it is inadvisable to put the testing in the hands of a willful agent intent on surprising the subjects, we give a highly stylized description of the situation that will allow us to carry over our previous analysis without change. We suppose the athlete is taking the illegal drug, and that the examiner is trying to find the test that will reveal the most.

Suppose that an examiner has a horizon of T days during which he can make one test (say for drug use). It is assumed too expensive to conduct more than one test during the horizon.[4] Each night the performer can take some effort e to mask the drug. The more effort taken to mask the drug, the harder it is to detect. We suppose that the payoff to the examiner on a day t on which he gives the exam is $A-e$, where e is the effort taken by the performer to mask the drug on the night before. Note that we assume efforts from several nights previously have no benefit in masking the

[3] Here we rely on the account presented by Lennart Ekbom in a letter he wrote to Martin Gardner, described in [?].

[4] If there were a fixed upper bound greater than one on the number of tests, our analysis would not change.

drug. The payoff to the examiner on day t is zero if he does not give the test that day. The total payoff to the examiner is the sum of the daily payoffs from days $t = 1$ to $t = T$.

The benefit to the athlete from making the masking effort e on the $t-1$st night is e itself if a test is given on the tth day (since it raises his score by e), and nothing if no test is given on the tth day. The cost to the athlete from making the effort e on the $t-1$st night is $e^2/2(A-B)$, which he bears whether or not the test is given the following morning, where $A > B \geq 0$. The total payoff to the athlete is the sum of all his expected benefits minus the sum of all his costs, over all the days and previous nights from $t = 1$ to $t = T$.[5]

Having described the action spaces for each agent, and their payoffs as a function of their joint actions, we have a well-defined game (albeit with an infinite number of strategies for the athlete). We can therefore analyze the game for its (conventional) Nash equilibria. These are described below.

Proposition 3 *The drug testing game has a unique Nash equilibrium in which the examiner gives the test the first day, and if he forgets, then the first day afterwards and so on, and the subject takes effort $e = A - B$ on the night before the first day, and the same effort every successive night until the test is actually given.*

Proof of Proposition 3 Observe that if the subject believes on the night before day t that he will be tested on day t with probability q_t, then he will act to maximize his expected payoff $q_t e_t - e_t^2/2(A-B)$, which he does by choosing $e_t = q_t(A-B)$. The examiner will therefore play the same strategy in Nash equilibrium of the drug testing game as he would in psychological Nash equilibrium of the Hangman's game. □

7 Conclusion

The moral of the hangman's paradox is thus seen to be quite practical and down to earth. If an examiner wants to test his students for sure in a fixed period of time, but surprise them as much as possible, he should announce a random examination schedule defined by the algorithm in Section 5.1. Each morning during the allotted horizon he should come in and make a grand show of drawing a ball from an urn, where the number of balls is adjusted each day so that the probability of getting an "exam" ball is equal to that specified in Section 5.1. As the days go on, the relative number of "nonexam" balls will steadily decrease until the last day, when there will only be exam balls in the urn. As an approximation to the optimal schedule, the examiner could begin with 1 black "exam" ball and $T-1$ white "no exam" balls in

[5] Another example of our game might fit the stylized facts more exactly. Consider an examiner who wants to ascertain the weight of a subject, who desires to appear as light as possible. The night before a weighing, the subject could go to a sauna and sweat out e pounds, but at the cost described above. Within a day or so, the weight will be back on him, and if he were not tested when he expected, his efforts would have gone for naught.

the urn. Each day he would publicly pick a ball out of the urn. If it is black, he gives the exam. If white, he discards the ball and the next morning chooses from an urn with one less white ball. Eventually the exam must be given, since if it has not been given before, by the Tth morning the urn will only contain one black ball.

If the examiner has no way to commit himself credibly to this schedule then he will have problems. For example, if the class thinks he will cheat and change the number of balls in the night, then he will not be able to surprise them at all.

In that case, the examiner must convince the class that he gets a little pleasure out of not testing them when they think he will test them. If that is common knowledge, then once again he will be able to test them, and surprise them almost completely, although in order to do so he will need a much longer time horizon in which to give the exam.

References

[1] Ells, Ellery (1982). *Rational Decision and Causality.* Cambridge: Cambridge University Press.

[2] Gardner, Martin (1986). *The Unexpected Hanging and Other Mathematical Diversions*, 2nd ed. New York: Simon & Schuster, Inc. (1st ed., 1969).

[3] Geanakoplos, J., D. Pearce, and E. Stacchetti (1989). "Psychological Games and Sequential Rationality," *Games and Economic Behavior*, 1: 60–79.

[4] Gilboa, Itzhak and David Schmeidler (1988). "Information Dependent Games: Can Common Sense be Common Knowledge?" *Economics Letters*, 27: 215–221.

[5] Halpern, Joseph Y. and Yoram Moses (1986). "Taken by Surprise: The Paradox of the Surprise Test Revisited," *Journal of Philosophical Logic*, 15: 281–304.

[6] Margalit, A. and M. Bar–Hillel (1984). "Expecting the Unexpected," *Philosophia*, 13: 263–288.

[7] Nash (1950). "Equilibrium Points in N-Person Games," *Proceedings of the National Academy of Sciences*, 36, 48–49.

[8] Nozick, R. (1969). "Newcomb's Problem and Two Principles of Choice." In N. Rescher et al. (ed.), *Essays in Honor of Carl G. Hempel*. Dordrecht: D. D. Reidel Publishing Company, pp. 114–146.

[9] O'Connor, D. J. (1948). "Pragmatic Paradoxes," *Mind*, 57: 358–359.

[10] O'Beirne, T. H. (1961). "Can the Unexpected Never Happen?" *The New Scientist*, 15: 464–465; letters and replies, June 8, 1961, pp. 597–598.

[11] Quine, W. V. (1953). "On a So-called Paradox," *Mind*, 62: 65–67.

[12] Scriven, Michael (1951). "Paradoxical Announcements," *Mind*, 60: 403–407.

[13] Skyrms, B. (1980). *Causal Necessity: A Pragmatic Investigation of the Necessity of Laws*. New Haven: Yale University Press.