Economics and Philosophy, **16** (2000), 267–285 Copyright © Cambridge University Press

THE LOGIC OF BACKWARDS INDUCTIONS

Graham Priest
University of Queensland

1. INTRODUCTION

Backwards induction is an intriguing form of argument. It is used in a number of different contexts. One of these is the surprise exam paradox. Another is game theory. But its use is problematic, at least sometimes. The purpose of this paper is to determine what, exactly, backwards induction is, and hence to evaluate it. Let us start by rehearsing informally some of its problematic applications.

The surprise exam paradox is as follows. A teacher announces to her students that there will be a surprise examination one day next week – a surprise in the sense that on the morning of the exam, the students will not know that the exam will be on that day. The students reason by backwards induction, as follows. The first such exam[1] cannot be on Friday. For if, on Thursday night, the exam has not yet been held, we will know that it must be on Friday; and so it will not be a surprise. But then, it cannot be on Thursday either; for if on Wednesday night the exam has not yet been held, we will know that it must be on Thursday, so it will not be a surprise. But then it cannot be on Wednesday either ... So it must be on Monday; in which case it will not be a surprise either. So there can be no such exam.

Of the many games that could be used to illustrate the situation, let us take the 'centipede game' of Rosenthal (1981). There are two players, you and I. There are 100 one dollar notes on the table (and this is known to us). We take it in turns until either the money runs out or someone

[1] Accounts of the paradox often assume that there will be only one exam. This is unnecessary.

https://doi.org/10.1017/S0266267100000250 Published online by Cambridge University Press

finishes the game. On any turn, one can take either \$1, in which case it is the other's turn; or else take \$2, finishing the game. Each of us wants to come away with as much money as possible. How should we play? To make it easier to visualize the argument, we can depict the game tree as follows:

$$\begin{array}{ccccccccccc} 1_m & \rightarrow & 2_y & \rightarrow & 3_m & \ldots & 98_y & \rightarrow & 99_m & \rightarrow & 100_y \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\ 2,0 & & 1,2 & & 3,1 & & 49,50 & & 51,49 & & 50,50 \end{array}$$

Numbers on the top line are move numbers. A suffix '*m*' represents my move; a suffix '*y*' represents yours. The pairs on the bottom line are the money that each of us obtains should the game end there, mine first, then yours. Note that a player always comes off better if they end the game than if the opponent ends the game at the next move.

The backwards induction is as follows. If the game ever gets to move 99, I will take \$2, ending the game. That way, I am better off. But you know this, so if the game ever gets to move 98, you will end the game. That way, you are better off. But I know this, so if the game reaches move 97 ... Hence on move 1, I will take \$2 and end the game.

Now, both the surprise exam argument and the backwards induction in the centipede game are problematic. In the surprise exam paradox, it seems to be patently clear that there *can* be a surprise exam. And in the centipede game, it is clear that we can *both* be better off if I do not terminate the game at the first move. Despite this, the orthodox reactions to the two arguments are opposite. Philosophers take it that there is something wrong with the surprise exam argument; according to game-theoretic orthodoxy, however, the backwards induction is correct in the centipede game and its ilk, and thus this game should finish at the first move.[2] Some people have noted the similarity between the two, and argued that they should be treated in the same way.[3] What is one to make of this? That is the major question on the agenda here.

In the next section I will specify the abstract structure of a backwards induction, isolating its essential features. In the following two sections, I will formalize both the surprise exam paradox and the centipede game. I do this for two reasons. The first is to show that the arguments do, indeed, instantiate the general schema given. The second reason is that accounts of backwards induction in game theory, in particular, usually give the argument in rather informal terms. Some writers do give more

[2] Though sometimes the endorsement is made with discomfort. See, e.g., Luce and Raiffa (1957, pp. 97–102). And the orthodoxy has been questioned by some in the last 10 years. See the references in Section 1 of Aumann (1995).

[3] E.g., Sorensen (1988, Chs. 7–10), Pettit and Sugden (1989).

https://doi.org/10.1017/S0266267100000250 Published online by Cambridge University Press

formal versions of the argument,[4] but these are a long way from being formalizations in the logician's sense of the term. (Few game-theorists are logicians.) But the purely *logical* moves that the backwards induction argument requires are very sensitive ones. They involve interplays between epistemic operators, deducibility principles, and the conditional, which need very careful handling. The formalization brings these facts – and their consequences – to light.[5]

With this material under our belt, we will be in a position to see what to make of backwards inductions. The next section of the paper analyses the surprise exam paradox. A crucial assumption concerning the persistence of knowledge is isolated and rejected. In the final section of the paper, the lessons learned from this analysis are applied to backwards induction in the centipede argument, and to game theory in general. The net result of the analyses – in a nutshell – is that most arguments that appeal to backwards induction fail. So as not to clutter the text, I will explain mainly in footnotes how all this relates to the literature on the issues.

2. GENERAL STRUCTURE

A standard, forward, induction from some information, β, arises when we have that $\beta \vdash \alpha(1)$ and $\beta, \alpha(x) \vdash \alpha(x+1)$.[6] Induction then gives that $\beta \vdash \forall x \alpha(x)$. (The variables here range over positive integers. We ignore zero.[7]) The simplest understanding of a backwards induction from information, β, is when we have that $\beta \vdash \alpha(n)$ (for some number n), and $\beta, \alpha(x+1) \vdash \alpha(x)$ (for $n > x \geq 1$). It follows that $\beta \vdash \alpha(1)$. This form of backwards induction is unproblematic. Its validity can be established by ordinary induction on n.

The backwards induction that we are concerned with is of a rather different kind, and involves epistemic considerations essentially. Let J be a sentence operator. In presentations of backwards inductions, it is most

[4] E.g., Aumann (1995).

[5] 'It is difficult to evaluate the validity of … [a complex piece of game-theoretic reasoning] using verbal tools only. That is a function of mathematical formalisms. In a formal model, the conclusions are derived from definitions and assumptions. Once one is satisfied that the derivation is mathematically correct, it remains only to examine the appropriateness of the definitions and assumptions. But with informal, verbal reasoning as complex as the above, one never knows for sure whether the argument is sound. One can argue till one is blue in the face, without convincing one another, because there is no criterion for deciding the soundness of an informal argument.' Aumann (1996, p. 142).

[6] For those who are not logicians: the logical symbol $\vdash$ indicates that the premises on its left hand side deductively entail the conclusion on its right hand side.

[7] This is not essential to what follows. It is simply because it is easier to follow the formal arguments in subsequent sections if we start counting at 1.

https://doi.org/10.1017/S0266267100000250 Published online by Cambridge University Press

often read as 'it is known that', and informally I will follow suit here. However, other readings will also be of interest, notably, 'it is believed that' and 'it is rationally believable that'. To formalize the induction we need more than a single operator. It is necessary to be able to represent what is known at stage i of the situation. So if the last relevant situation is the nth, we need a family of modal operators, J_i $(1 \leq i \leq n)$.

I want to assume as little about the behaviour of each J_i as possible, so that we can isolate those features that are essential in the argument. The major such feature is the inference of *Closure*. For each i:

If $\alpha_1, \ldots, \alpha_k \vdash \beta$ then $J_i\alpha_1, \ldots J_i\alpha_k \vdash J_i\beta$

Closure holds for knowledge: knowledge is closed under deducibility.[8] It also holds for justified believability: what is entailed by justified beliefs is itself justified.[9] It is not true of belief, however. Actual beliefs are not closed under deducibility.

We do not need to assume that each J_i is veridical, that is, that $J_i\alpha \vdash \alpha$. It will help to have corresponding operators that are, though. Thus, for $1 \leq i \leq n$, we define $K_i\alpha$ as $\alpha \wedge J_i\alpha$. Each K_i obviously is veridical. Of course, if J is interpreted as knowledge, $J_i\alpha \vdash \alpha$, and so J_i and K_i come to the same thing, but they need not if J is some lesser notion. Note that if a J satisfies Closure, so does the corresponding K. (For then, if $\alpha_1, \ldots, \alpha_k \vdash \beta$, $\alpha_1 \wedge J_i\alpha_1, \ldots, \alpha_n \wedge J_i\alpha_k \vdash \beta \wedge J_i\beta$.)

One further piece of notation will be useful. If $1 \leq i \leq j \leq n$, let K_{ij} be $K_{i+1} \ldots K_j$. Note that K_{ii} is just K_i. Note also that $K_{in}\alpha \vdash K_{i+1n}\alpha \vdash \ldots \vdash K_n\alpha \vdash \alpha$. Given this machinery, we can formulate the abstract structure of a backwards induction as follows.

Backwards Induction Schema

Suppose that for some information, β, we have:

1. $K_n\beta \vdash \alpha(n)$
2. $K_i\beta, K_i\alpha(i+1) \vdash \alpha(i)$ (for $1 \leq i < n$)

Then $K_{1n}\beta \vdash \alpha(1)$.

[8] In principle, anyway. In practice, no one may actually have drawn the appropriate conclusions.

[9] There is a caveat here, though. I take it that the paradox of the preface shows that there are situations where it may be rational to believe α and $\neg\alpha$, but not to believe $\alpha \wedge \neg\alpha$. (See, e.g., Prior (1971, p. 85).) The failure of Closure in this case is due to the fact that it is different considerations which ground each of the contradictories. In the cases we are concerned with in this paper, there is a unified source of justification. Hence, we may ignore such complications.

https://doi.org/10.1017/S0266267100000250 Published online by Cambridge University Press

Proof

This is proved by the unproblematic kind of backwards induction. Let $\varphi(j)$ be the formula: $K_{jn}\beta \vdash \alpha(j)$. By 1, $\varphi(n)$. Suppose that $\varphi(i+1)$ $(i < n)$, that is, $K_{i+1\,n}\beta \vdash \alpha(i+1)$. Then by Closure of K_i:

$K_{in}\beta \vdash K_i\alpha(i+1)$
But: $K_{i+1\,n}\beta \vdash \beta$
So: $K_{in}\beta \vdash K_i\beta$
Hence: $K_{in}\beta \vdash K_i\beta \wedge K_i\alpha(i+1)$
So by 2: $K_{in}\beta \vdash \alpha(i)$

That is, $\varphi(i)$. Hence, by induction, $\varphi(1)$, i.e., $K_{1n}\beta \vdash \alpha(1)$.

To see what this comes to in terms of J, those with a penchant for combinatorics can establish that $K_{in}\beta$ is equivalent to the conjunction of:

β
$J_j\beta$ for all $i \leq j \leq n$
$J_jJ_k\beta$ for all $i \leq j \leq k \leq n$
$J_jJ_kJ_l\beta$ for all $i \leq j \leq k \leq l \leq n$
$\vdots$
$J_{in}\beta$

Call this conjunction $\hat{J}_i$.[10] Hence, the result of the backwards induction is that $\hat{J}_1\beta \vdash \alpha(1)$.

3. THE SURPRISE EXAM

Let us now see that the surprise exam argument fits this general schema.[11] Suppose that there are n days, 1, ..., n. For $1 \leq i \leq n$ let η_i be the sentence 'The first exam will be held on day i' ('η' for 'exam'). J_i may be read 'It is known on (the morning of) day i that'.

The information known by the students is the conjunction of a number of different components. The first is that there will be a surprise exam, that is, that there will be an exam on one day and it will not be known on that day that it will occur. Call this σ ('σ' for 'surprise'):

$$(\eta_1 \wedge \neg J_1\eta_1) \vee \ldots \vee (\eta_n \wedge \neg J\eta_n) \qquad (\sigma)$$

[10] The proof is by a standard backwards induction on i. If i is n, the result holds by definition. Suppose that the result holds for $i+1$. Then $K_{in}\beta$ is $K_iK_{i+1\,n}\beta = K_{i+1\,n}\beta \wedge J_iK_{i+1\,n}\beta = \hat{J}_{i+1} \wedge J_i\hat{J}_{i+1}$, by induction hypothesis. A little thought suffices to establish that this is $\hat{J}_i$.

[11] The following is not the only possible formalization of the surprise exam paradox. There are different versions, in which the information known by the students is self-referential. ('The students will not be able to infer from *this information* that ...') See, e.g., Kaplan and Montague (1960), Halpern and Moses (1986). There is no self-reference in what follows.

https://doi.org/10.1017/S0266267100000250 Published online by Cambridge University Press

The next piece of information is that the exam in question is the first, that is, if the exam occurs on day i ($1 \leq i \leq n$), it has not occurred before. Call this φ_i ('φ' for 'first'):[12]

$$\eta_i \rightarrow \neg\eta_1 \wedge \ldots \wedge \neg\eta_{i-1} \qquad (\varphi_i)$$

Finally, there is information to the effect that the students do not forget what has happened, that is, if the exam does not happen on day i ($1 \leq i \leq n$), then the students know this on subsequent days. Call this ρ_i ('ρ' for 'remembering'):[13]

$$\neg\eta_i \rightarrow J_{i+1}\neg\eta_i \wedge \ldots \wedge J_n\neg\eta_i \qquad (\rho_i)$$

Now to the α and β of the backwards induction schema. The β in question is the conjunction of all these things:

$$\sigma \wedge \bigwedge_{1\leq i\leq n}\varphi_i \quad \wedge \bigwedge_{1\leq i\leq n}\rho_i \qquad (\beta)$$

For $1 \leq i \leq n$, α_i is:

$$\neg\eta_i \wedge \ldots \wedge \neg\eta_n \qquad (\alpha_i)$$

that is, the statement that the exam does not occur on or after day *i*. The αs are what the students, step by step, deduce.

To verify that the conditions of backwards induction are met, we first establish that $K_n\beta \vdash \alpha_n$. Since φ_n is a conjunct of β:

$$\eta_n, \beta \vdash \neg\eta_1 \wedge \ldots \wedge \neg\eta_{n-1} \qquad \text{(o)}$$

So since σ is a conjunct of β:

$$\eta_n,\beta \vdash \neg J_n\eta_n \qquad \text{(i)}$$

By propositional logic and Closure:

$\neg\eta_1, \ldots, \neg\eta_{n-1}, \eta_1 \vee \ldots \vee \eta_n \vdash \eta_n$

$J_n\neg\eta_1, \ldots, J_n\neg\eta_{n-1}, J_n(\eta_1 \vee \ldots \vee \eta_n) \vdash J_n\eta_n$

But since each ρ_i is a conjunct of β, (o) gives:

$\eta_n, \beta \vdash J_n\neg\eta_1 \wedge \ldots \wedge J_n\neg\eta_{n-1}$

And since $\beta \vdash \eta_1 \vee \ldots \vee \eta_n$

$J_n\beta \vdash J_n(\eta_1 \vee \ldots \vee \eta_n)$

Hence:

$$\eta_n, \beta, J_n\beta \vdash J_n\eta_n \qquad \text{(ii)}$$

[12] If $i = 1$ the consequent is to be understood as any tautology.
[13] Again, if $i = n$ the consequent is to be understood as any tautology.

https://doi.org/10.1017/S0266267100000250 Published online by Cambridge University Press

By (i) and (ii):

$\eta_n, \beta, J_n\beta \vdash \neg J_n\eta_n \wedge J_n\eta_n$

Thus:

$K_n\beta \vdash \neg\eta_n$, i.e., α_n.

Second, we need to show that for $1 \leq i < n$: $K_i\beta, K_i\alpha_{i+1} \vdash \alpha_i$. Since φ_i is a conjunct of β:

$\eta_i, \beta \vdash \neg\eta_1 \wedge \ldots \wedge \neg\eta_{i-1}$ (iii)

$\eta_i, \beta, \alpha_{i+1} \vdash \neg\eta_1 \wedge \ldots \wedge \neg\eta_{i-1} \wedge \neg\eta_{i+1} \wedge \ldots \wedge \neg\eta_n$

So since σ is a conjunct of β:

$\eta_i, \beta, \alpha_{i+1} \vdash \neg J_i\eta_i$ (iv)

And:

$\neg\eta_1, \ldots, \neg\eta_{i-1}, \alpha_{i+1}, \beta \vdash \eta_i$

$J_i\neg\eta_1, \ldots, J_i\neg\eta_{i-1}, J_i\alpha_{i+1}, J_i\beta \vdash J_i\eta_i$

And since each ρ_i is a conjunct of β:

$\beta, \neg\eta_1, \ldots, \neg\eta_{i-1}, J_i\alpha_{i+1}, J_i\beta \vdash J_i\eta_i$

So by (iii):

$\eta_i, \beta, J_i\alpha_{i+1}, J_i\beta \vdash J_i\eta_i$ (v)

Thus by (iv) and (v):

$\eta_i, \beta, \alpha_{i+1}, J_i\alpha_{i+1}, J_i\beta \vdash J_i\eta_i \wedge \neg J_i\eta_i$

Hence:

$K_i\alpha_{i+1}, K_i\beta \vdash \neg\eta_i$

$K_i\alpha_{i+1}, K_i\beta \vdash \neg\eta_i \wedge \alpha_{i+1}$

The consequent is α_i as required.

By the general theorem about backwards inductions, it follows that $K_{1n}\beta \vdash \alpha_1$, that is, there is no exam. But note also that $K_{1n}\beta \vdash \beta \vdash \eta_1 \vee \ldots \vee \eta_n$, i.e., $\neg\alpha_1$. Hence, $K_{1n}\beta$ is inconsistent.

4. THE CENTIPEDE GAME

Let us now turn to the backwards induction in the centipede game. The surprise exam argument assumes no more of the conditional than that it satisfy *modus ponens*: $\alpha, \alpha \rightarrow \beta \vdash \beta$. The centipede game does. I will assume that, as well as *modus ponens*, the conditional satisfies the inference $\alpha \leftrightarrow \beta, \beta \rightarrow \gamma \vdash \alpha \rightarrow \gamma$ (where $\alpha \leftrightarrow \beta$ is $(\alpha \rightarrow \beta) \wedge (\beta \rightarrow \alpha)$).

https://doi.org/10.1017/S0266267100000250 Published online by Cambridge University Press

Call this, *restricted transitivity*. This inference and *modus ponens* hold on virtually all accounts of the conditional: material, intuitionist, relevant, subjunctive (counterfactual).[14] There is much debate in the literature on this paradox as to what sort of conditional is appropriate for the argument.[15] Since the following formalization depends only on the above principles, we may finesse that question here.[16]

For the formalization, let the players be a, who moves first, and b. The other vocabulary we need is as follows:
γ_i – the game reaches the ith move.
$C_i(z, x, y)$ – z has to choose between x and y at move i.
$B(x)$ – it is brought about that x.
$F(z, x)$ – the final result for z is x.
t_i – the action of taking two and terminating the game at move i.
o_i – the action of taking one and handing the move to the next player at move i.
J_i may be read as 'It is known at the ith stage that'. The ith stage should not, here, be interpreted as the ith move of the game. One may never, in fact, get to the ith move. Think of the game as being played over time, each move taking one unit of time. The ith stage is just the ith instant of time. The clock may continue after the actual game ends.

The information required for reasoning about the game falls into four groups. Suppose that there are $n + 1$ possible moves in the game: $1, \ldots, n + 1$. (The last one is a forced move, and so no real move at all.) Then, first, there is the information about the structure of the game. Let μ ('μ' for 'moves') be the conjunction of γ_1 and the following:

$$\begin{array}{ll} \gamma_i \leftrightarrow C_i(a, t_i, o_i) & \text{for odd } i \leq n \\ \gamma_i \leftrightarrow C_i(b, t_i, o_i) & \text{for even } i \leq n \\ B(o_i) \rightarrow \gamma_i & \text{for all } 0 < i \leq n \\ B(t_i) \rightarrow \gamma_i & \text{for all } 0 < i \leq n \\ B(o_i) \leftrightarrow \gamma_{i+1} & \text{for all } i \leq n \end{array} \qquad (\mu)$$

The first two say that the game reaches the ith move if and only if the appropriate agent has a choice between taking one and two at the ith move. The second two say that if one or two are taken at the ith move then the game must have reached the ith move. The last one says that the $i + 1$st move is reached if and only if one is taken at the ith move. All of these are obviously true.

[14] Standard transitivity, that is, where the first premise is a simple conditional, fails on all accounts of the subjunctive conditional.

[15] Indeed, some writers, e.g., Binmore (1987), have claimed that the argument is invalid if a counterfactual conditional is employed. This is false, as we shall see.

[16] Though there are good reasons for supposing that, whatever the conditional is, it is not the material conditional of 'classical' logic. See fn. 31.

https://doi.org/10.1017/S0266267100000250 Published online by Cambridge University Press

Next, there is the information about the payoffs. Let π ('π' for 'payoffs') be the conjunction of:

$B(o_n) \rightarrow F(a,(n+1)/2) \wedge F(b,(n+1)/2)$ if n is odd

$B(o_n) \rightarrow F(a,1+n/2) \wedge F(b,n/2)$ if n is even

Together with:

$$\begin{array}{ll} B(t_i) \rightarrow F(a,2+(i-1)/2) \wedge F(b,(i-1)/2) & \text{for odd } i \leq n \\ B(t_i) \rightarrow F(a,i/2) \wedge F(b,1+i/2) & \text{for even } i \leq n \end{array} \qquad (\pi)$$

The first two of these give the payoffs for taking one at the last free move. (If the move is handed over at the nth move, the next, and $n+1$st, move is the last.) The next two state the payoffs for taking two at any move.[17]

We also need the information about the 'rationality' of the actors. Let ρ ('ρ' for 'rationality') be the conjunction for odd i and even j of:[18]

$$\begin{array}{l} (J_i(B(x) \rightarrow F(a,u)) \wedge J_i(B(y) \rightarrow F(a,v)) \wedge u > v) \rightarrow (C_i(a,x,y) \rightarrow B(x)) \\ (J_j(B(x) \rightarrow F(b,u)) \wedge J_j(B(y) \rightarrow F(b,v)) \wedge u > v) \rightarrow (C_j(b,x,y) \rightarrow B(x)) \end{array} \qquad (\rho)$$

In other words, if it is known that, of two actions, one gives a larger payoff than the other, if an agent has to choose between them, they will choose the one with the larger payoff.[19]

Finally we need some facts of elementary arithmetic. Let these be η ('η' for 'erithmetic').

We are now in a position to spell out the α and β of the backwards induction schema. β is $\mu \wedge \pi \wedge \rho \wedge \eta$. This is the information available to the players. α_i is $\gamma_i \rightarrow B(t_i)$, i.e., if the game reaches stage i, it terminates there.

To show that the backwards induction schema is instantiated, we need to show that $K_n\beta \vdash \alpha_n$, and for all $1 \leq i < n$, $K_i\beta, K_i\alpha_{i+1} \vdash \alpha_i$. The argument for the first depends on whether n is odd or even. Here is the case for n odd. The other case is similar. Since π is a conjunct of β we have:

$\beta \vdash B(t_n) \rightarrow F(a,2+(n-1)/2)$

$\beta \vdash B(o_n) \rightarrow F(a,(n+1)/2)$

[17] If the game ends on an odd move, i, an even number of dollars, $i-1$, have been shared out. Then player 1 takes two more. If i is even, player 1 has $i/2$ dollars, player 2 has one less, but then takes two.

[18] Or, strictly, their universal closures. That is, the principles prefixed by a universal quantifier, $\forall z$, for every free variable, z.

[19] In fact, at odd moves it is only a's knowledge that is relevant, and the opposite at even moves. Since the knowledge of the players is the same, this makes no difference. Alternatively, we could have two families of operators, J_i^a, J_i^b, and just *define* J_i to be J_i^a if i is odd, and J_i^b if i is even. Note that the principles constituting ρ seem just as valid if J is interpreted as belief or justified believability.

https://doi.org/10.1017/S0266267100000250 Published online by Cambridge University Press

Hence:

$J_n\beta \vdash J_n(B(t_n) \rightarrow F(a, 2 + (n-1)/2))$ (i)

$J_n\beta \vdash J_n(B(o_n) \rightarrow F(a, (n+1)/2))$ (ii)

And since η is a conjunct of β:

$\beta \vdash 2 + (n-1)/2) > (n+1)/2$ (iii)

But, by (i)–(iii), since ρ is a conjunct of β:

$\beta, J_n\beta \vdash C_n(a, t_n, o_n) \rightarrow B(t_n)$

But since μ is a conjunct of β:

$\beta \vdash \gamma_n \leftrightarrow C_n(a, t_n, o_n)$

Hence, by restricted transitivity:

$K_n\beta \vdash \gamma_n \rightarrow B(t_n)$

as required.

Next, suppose that $1 \leq i < n$. There are two cases, depending on whether i is odd or even. Again, I will do the case for i odd. The other is similar. As before, we have:

$\beta \vdash B(t_i) \rightarrow F(a, 2 + (i-1)/2)$

$J_i\beta \vdash J_i(B(t_i) \rightarrow F(a, 2 + (i-1)/2))$ (i)

Now, by definition of α_{i+1}:

$\alpha_{i+1} \vdash \gamma_{i+1} \rightarrow B(t_{i+1})$

and since μ is a conjunct of β:

$\beta, \alpha_{i+1} \vdash \gamma_{i+1} \leftrightarrow B(t_{i+1})$

Again, since μ is a conjunct of β:

$\beta \vdash B(o_i) \leftrightarrow \gamma_{i+1}$

Hence, by restricted transitivity:

$\beta, \alpha_{i+1} \vdash B(o_i) \leftrightarrow B(t_{i+1})$

Since π is a conjunct of β (and $i+1$ is even):

$\beta \vdash B(t_{i+1}) \rightarrow F(a, (i+1)/2)$

So by restricted transitivity again:

$\beta, \alpha_{i+1} \vdash B(o_i) \rightarrow F(a, (i+1)/2)$

$J_i\beta, J_i\alpha_{i+1} \vdash J_i(B(o_i) \rightarrow F(a, (i+1)/2))$ (ii)

But since η is a conjunct of β:

https://doi.org/10.1017/S0266267100000250 Published online by Cambridge University Press

$\beta \vdash 2 + (i-1)/2) > (i+1)/2$ (iii)

Hence, by (i)–(iii), since ρ is a conjunct of β:

$\beta, J_i\beta, J_i\alpha_{i+1} \vdash C_i(a, t_i, o_i) \rightarrow B(t_i)$

But, since μ is a conjunct of β:

$\beta \vdash \gamma_i \leftrightarrow C_i(a, t_i, o_i)$

So by restricted transitivity:

$\beta, J_i\beta, J_i\alpha_{i+1} \vdash \gamma_i \rightarrow B(t_i)$

Hence:

$K_i\beta, J_i\alpha_{i+1} \vdash \gamma_i \rightarrow B(t_i)$

So, $K_i\beta, K_i\alpha_{i+1} \vdash \alpha_i$, as required.

By the backwards induction schema, it follows that $K_{1n}\beta \vdash \alpha_1$. And since $K_{1n}\beta \vdash \beta \vdash \gamma_1$ (γ_1 is a conjunct of μ), $K_{1n}\beta \vdash B(t_1)$, that is, the game stops at the first move.

5. THE PERSISTENCE OF KNOWLEDGE

We are now in a position to analyse the backwards induction in the surprise exam paradox. In this, as we have seen, $K_{1n}\beta$ is inconsistent, since it entails both $\neg\alpha_1$ and α_1: there will and there will not be an exam. $K_{1n}\beta$ is therefore false.[20] It seems undeniable, however, that $K_1\beta$ can be true, however J is interpreted. (Provided that $n > 1$: if $n = 1$, $K_1\beta$ is a simple contradiction.) Certainly, β can be true; it can be believed; and the evidence for it can be as strong as you like. (The teacher is of impeccable character, never lies, etc. She could even be God.)

So $K_1\beta$ may be true, but $K_{1n}\beta$ is not. We can extract further information from this. For there are certain principles which, together with the former, entail the latter. For a start, consider the principle:

K0 $K_i\delta \vdash K_iK_{i+1}\delta$

(for all δ, and $1 \leq i < n$). This entails that for all $1 \leq i \leq n$, $K_i\delta \vdash K_{in}\delta$. The case for $i = n$ is true by definition. Suppose that $K_i\delta \vdash K_{in}\delta$. Then by Closure, $K_{i-1}K_i\delta \vdash K_{i-1\,n}\delta$. Since $K_{i-1}\delta \vdash K_{i-1}K_i\delta$, the result holds for $i - 1$. Hence, we have the result.

Hence, K0 must fail. This is, perhaps, not terribly informative. The content of K0 is opaque enough for us not to have strong intuitions about

[20] There is always, of course, the dialetheic option: $K_{1n}\beta$ is true, and so is the contradiction. But that has no appeal in this case. The contradiction involved is not some recondite and unobservable state of affairs like the Liar sentence being true and false. Either there will be an exam or there will not; and that is an exclusive disjunction.

https://doi.org/10.1017/S0266267100000250 Published online by Cambridge University Press

it.[21] More significant, is the fact that K0 follows from the two simpler principles:

K1 $K_i\delta \vdash K_iK_i\delta$
K2 $K_i\delta \vdash K_{i+1}\delta$

From K2 and Closure, it follows that $K_iK_i\delta \vdash K_iK_{i+1}\delta$. Whence K1 entails the result. Hence, one of K1 and K2 must fail. Note also that K1 and K2 are entailed by the corresponding principles for J:

J1 $J_i\delta \vdash J_iJ_i\delta$
J2 $J_i\delta \vdash J_{i+1}\delta$

That K2 follows from J2 is obvious. For K1:

$\delta, J_i\delta \vdash \delta \wedge J_i\delta$
$J_i\delta, J_iJ_i\delta \vdash J_i(\delta \wedge J_i\delta)$
$J_i\delta \vdash J_i(\delta \wedge J_i\delta)$ (by J1)
$\delta \wedge J_i\delta \vdash \delta \wedge J_i\delta \wedge J_i(\delta \wedge J_i\delta)$
$K_i\delta \vdash K_iK_i\delta$

Hence, either J1 or J2 must fail.

What to make of this? For a start, J1 and J2 both fail for belief. A person may believe something without ever having reflected on their beliefs, and so believing that they believe them (J1); and the fact that a person believes something at a certain time does not entail that they believe it later: they may change their mind for all sorts of reasons (J2).

When J is interpreted as a stronger notion the situation is more interesting. Take J1 first. If J is interpreted as justified belief or as knowledge – in which case, it is often called the *KK* thesis – this looks a plausible principle. Whatever it is that grounds our justified belief (knowledge) of δ grounds the claim that we are justified in believing (know) that δ. The principle has been defended along these lines by various philosophers.[22] But it has also been criticized on various grounds. Now, it may well be that the principle (for both justified belief and knowledge) fails in general. However, the most persuasive examples of its failure turn on various cognitive limitations of the agent, such as the failure of self-awareness. In the context of the surprise exam paradox, we are at liberty to suppose that there are no such limitations: the agents are fully aware of all relevant factors, have full information, are self-aware, etc. All the relevant evidence is on the table, as it were, for

[21] Though a version of it is defended in the context of the surprise exam paradox in Binkley (1968, p. 133f).

[22] For example, for justified belief – and in the context of the surprise exam paradox – it is defended by Wright and Sudbury (1977, p. 53). The *KK* thesis is defended in Hintikka (1962, Ch. 5).

https://doi.org/10.1017/S0266267100000250 Published online by Cambridge University Press

all to see; and there is no bar to employing it. In such a case, it is difficult to produce a cogent counter-example to J1.[23]

J2, the persistence of knowledge (or justified belief) through time, is much more problematic. Consider, first, the case where J is interpreted as justified belief. Evidence that is sufficient to give rational ground for belief may become insufficient in the light of subsequent evidence. It is rational for me to believe that life in Sydney is, at the time that I write, normal. But if I see a newspaper headline saying that there was a melt-down at Lucas Heights last night, it will no longer be rational to believe this. β illustrates this failure exactly. On day $i+1$ we have more information than we had on day i. And this may well give us rational ground for changing our beliefs. Even if it is rational to believe that there will be a surprise exam on (the morning of) day 1, the fact that the exam has failed to materialize by day $n-1$ gives reason to believe that the teacher has forgotten, or changed her mind (especially if n is very large). And hence, it is no longer rational to believe β.[24] Thus, the paradox is solved by the failure of J2.[25] Note that, in the present case, the increase in information that vitiates J2 is occasioned by the passage of time; but this is not essential. The crucial point is simply that what is justified with respect to certain information may not be so with respect to augmented information.[26]

The principle J2 looks plausible if J is interpreted as knowledge, and one endorses the condition that knowledge requires conclusive evidence. For having *conclusive* evidence for δ ensures that nothing that happens subsequently will undercut the belief in δ. Hence, knowledge about something entails later knowledge. But the condition is too strong. Since conclusive evidence is virtually never obtainable, the condition entails that one never knows anything. That is, it results in scepticism. Conclusive evidence is not required for knowledge: the evidence is merely required to be sufficient, as fallibilists about knowledge have emphasized.[27] But in that case, the situation is much the same as that for

[23] The point (with references to the literature) is argued in detail by Sorensen (1988, pp. 313–7).

[24] There are shades of a sorites paradox here, however. Given the sequence $J_1\beta, J_2\beta, \ldots, J_n\beta$, it may be very difficult to say which is the first false member of the sequence. The point is developed further in Sorensen (1988).

[25] This is the solution of Wright and Sudbury (1977). A similar solution is espoused in Jackson (1987).

[26] The point is nicely illustrated by Sorensen (1988, pp. 317–20), who formulates a synchronic version of the paradox where the increase in information is occasioned by different spatial locations.

[27] According to Jackson (1987, Ch. 7), the *hard* surprise exam paradox is when the information of the students at every stage 'is certain', that is, conclusively demonstrated. But the solution then is simply that $K_1\beta$ is false. This is not a paradox. *Conclusive* demonstration of anything about the future *is* impossible.

https://doi.org/10.1017/S0266267100000250 Published online by Cambridge University Press

rational belief: knowledge may disappear in the light of further information; and the case of the surprise exam illustrates exactly this with β. At any rate, given that $K_{1n}\beta$ fails, and that this is entailed by $K_1\beta$, together with J1 and J2; and given that the first two of these are correct, J2 must fail. We therefore have an argument for fallibilism about knowledge. Such fallibilism is, I think, a major lesson to be learned from the surprise examination paradox.

A final interpretation of the operator J is illuminating and worth considering.[28] According to this, $J\alpha$ is interpreted as 'α is undefeatable by true information', where this means that there is no true information, the receipt of which would cause one to revise one's belief that α. This looks like a plausible candidate for the satisfaction of *J*2, at least if one starts off with beliefs that are true (as we may suppose that they are in the case at hand). For how could the truth cause us to ditch true beliefs? A little thought, however, shows that this could indeed happen. For example, suppose that there is some species – let us call the creatures swans – such that all known swans are white. You believe (rationally), therefore, that there are no black swans. You then start to explore hitherto untraversed regions, travelling further towards the equator. As you go, the swans start to change colour, through light grey, to dark grey, and approaching black. There are still many miles to go to the equator, but by now any rational person would surely ditch their belief that there are no black swans. It is at least a good bet that there are black swans closer to the equator. For all that, it may yet be true that there are no black swans:– very dark grey is as dark as they get.[29]

Whether *J*, interpreted in this way, satisfies J1 is also dubious. If α is undefeatable by true information, does it follow that 'α is undefeatable by true information' is itself undefeatable by true information? It would seem not. Let us suppose that you believe α, say that all swans are white. Let us suppose that α is true, and that there is, in fact, no true evidence that would cause you to revise the belief in α: it is undefeatable by true information. Does it follow that that claim is itself undefeatable by true information? No. Suppose that the reason that you believe α is, in part, that all the swans you have seen are white, but it is also because you have been told that α by an authoritative source – maybe even God. On the basis of this, you believe that α is undefeatable by true information. You are then given (true) evidence that your authoritative source *might* be a fake. Though you do not revise your belief that α, you should revise your belief that α is undefeatable by true information. That α is

[28] This interpretation was suggested to me, in correspondence, by Wlodek Rabinowicz.

[29] Of course, if it were certain, for some reason, that there were no black swans, then the evidence would not lead one to change one's mind. But if we require certitude, the problems lie elsewhere, as we have seen.

https://doi.org/10.1017/S0266267100000250 Published online by Cambridge University Press

undefeatable by true information is itself, therefore, defeatable by true information.

In case one wants to contest the foregoing conclusions, let me reiterate one final point. Let 'know' be any interpretation of J. Then, dialetheism aside, there can be no sense – the last considered, or any other – in which you can know the conditions of the surprise exam paradox, β, and in which knowledge satisfies J1 and J2. For these, as we have seen, give contradiction.

6. GAME THEORY

Let us now move on and apply what we have learned to the centipede game, and, more generally, to game theory.

In the centipede game, the backwards induction gives us that $K_{1n}\beta \vdash \alpha_1$, with the appropriate β and α. In particular, $K_{1n}\beta \vdash B(t_1)$. Now, this is highly counter-intuitive, as I have already observed; but this time, $K_{1n}\beta$ is consistent.[30] Perhaps this is why many game theorists have bitten the bullet and accepted the result.[31]

But the application of the backwards induction is still not out of the woods. Why should one suppose that $K_{1n}\beta$ is true? It is not at all obvious. $K_1\beta$ is, we may grant, true. (β can certainly be true; it can be believed; and the evidence for it can be as strong as one wishes.) But the only way of

[30] And demonstrably so. To see this, we construct a Kripke model for S5. At every world, every γ is true, each C, B, and F have a total extension, and $<$ behaves as usual. Each K_i is simply treated as $\Box$. $\rightarrow$ can be taken to be either a strict or a material conditional. It is then easy to check that β is true at every world, as is $K_{1n}\beta$.

[31] Though if we add a little more unobjectionable information to β, and treat the conditional as a material conditional, $K_{1n}\beta$ is inconsistent, at least given J1. The information in question is $\neg(B(t_1) \wedge B(o_1))$, which says that a player cannot choose both one and two dollars at the first move. Let $K_{1n}\beta$ be δ. Since $\delta \vdash \beta \vdash \neg(B(t_1) \wedge B(o_1))$, and $\delta \vdash B(t_1)$, $\delta \vdash \neg B(o_1)$. By properties of the material conditional, and Closure:

$\delta \vdash B(o_1) \rightarrow F(a, 100)$
$J_1\delta \vdash J_1(B(o_1) \rightarrow F(a, 100))$

But now, as in the regular argument:

$J_1\beta \vdash J_1(B(t_1) \rightarrow F(a, 2))$

and since η and ρ are conjuncts of β:

$\beta \vdash 100 > 2$
$\beta, J_1\beta, J_1\delta \vdash B(o_1)$

Hence:

$K_1\delta, K_1\beta \vdash B(o_1)$

But, $\delta \vdash \beta$. Hence, $K_1\delta \vdash K_1\beta$. Thus:

$K_1\delta \vdash B(o_1)$

But $K_1\delta \vdash \delta \vdash \neg B(o_1)$. Hence, $K_1\delta$ is inconsistent. Given J1, it follows that $K_{1n}\beta$ is also inconsistent. This idea is to be found, one way or another, in Bicchieri (1989), Bonanno (1991) and Vilks (1997).

https://doi.org/10.1017/S0266267100000250 Published online by Cambridge University Press

inferring $K_{1n}\beta$ from this would appear to be by using J1 and J2. And the lesson of the surprise exam paradox, as we have just seen, is that there is no sense in which one can know β, and in which J1 and J2 hold. The β in question, it is true, is a different β in this case; but there seems to me to be no reason why the evidential status of the two βs must differ. In any case, as we saw, J2 must fail on any reasonable interpretation of J. In particular, even if we have $J_1\beta$, if we arrive at move two at all, then *a* did not make the 'rational' move at move one; our ground for believing β (and *a fortiori*, believing rationally and knowing it) has disappeared,[32] so $J_2\beta$ is false. Hence, $K_1\beta$ does not entail the paradoxical result.[33]

This failure of the argument is hidden in many standard accounts of the centipede game. For in these, knowledge is not taken to be temporally indexed. In particular, a single K operator is used, rather than a family. One can therefore not even distinguish between what is known at different times. Thus, if one goes through the formal arguments above, and simply deletes all the suffixes on J and K, everything goes through happily. The contentious J2 collapses into the trivial validity $J\delta \vdash J\delta$. In other words, where there is but a single K operator, the persistence of knowledge is packed into the very notation. This is so, for example, in the well known account of Aumann (1995).[34] In Aumann (1998, p. 98), a time-indexed notion of knowledge is employed, but persistence is still packed in explicitly, as the assumption that knowledge is never lost.

A restricted form of persistence is built into the formalization of Rabinowicz (1998). He assumes that knowledge persists, not in all possible situations, but in all actual situations. This, in the end, is just as problematic. For example, it falls to the example about the black swans and travelling to the equator, that we had at the end of the last section: knowledge is lost as we *actually* view darker and darker swans. In fact,

[32] The need for the knowledge to persist if the argument is to work is pointed out by Reny (1992, p. 110). And a number of writers have, in effect, noted that it does not, e.g., Bicchieri (1989), Pettit and Sugden (1989). Sorensen (1988, pp. 355–61) also argues the point. For him, the non-monotonicity of justifiable belief (and knowledge) under the addition of information provides the uniform resolution of both the sorites paradox and the game-theoretic 'paradoxes'.

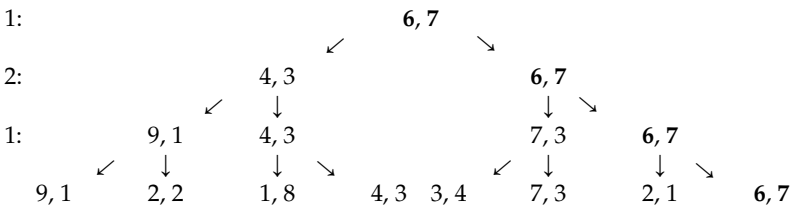
[33] Pettit and Sugden (1989), in effect, endorse this solution where J is rational believability, but not where it is knowledge, which they appear take to satisfy J1 and J2. For this, they say, there is no paradox. The paradox arises because of the conflict between the backwards induction, and the intuition that one might have done better. But in the case where J is knowledge, one *could not* have done better, as a matter of logical necessity. In particular, considering what would happen if the first player does not finish the game at move one is considering what to do in a logically impossible situation, given the defining conditions of the set-up. This explanation will not do. Doing better clearly *is* a logical possibility. It may be achieved by pursuing the game to its end.

[34] Aumann is aware that he is assuming the persistence of knowledge. He points this out himself (1995, p. 17). For other authors who use a non-temporally-indexed notion of knowledge, see Bicchieri (1989), Balkenborg and Winter (1997).

https://doi.org/10.1017/S0266267100000250 Published online by Cambridge University Press

in Broome and Rabinowicz (2000), a similar counter-example, due to Sliwinski, and formulated in terms of the game itself, is given. Their defence against the counter-example is to deduce the principle from something else, namely, that 'both players retain all their beliefs as long as they are consistent with their acquired beliefs'. Now, this strikes me as a somewhat curious strategy. For if γ entails δ, then any counter-example to δ is a counter-example to γ. And so it would seem in this case. In particular, in the swans example, the new evidence is quite consistent with there being no black swans; but it still becomes irrational to believe that there are none.

In game theory, backwards induction is used, not just in the centipede game, but quite generally.[35] What are the consequences of our investigations for this? The backwards induction concerns a certain algorithm for 'solving' certain games, that is, determining a path from the start of a game to a conclusion.[36] The algorithm is essentially as follows. We know the payoff at each terminal node. Working by recursion up the game-tree, we assign to each node the payoff at the node immediately below which assigns the maximum to whoever's move it is. (If there is more than one, just choose one.) We do this until we reach the root of the tree (the first move). The solution to the game is a branch from the root to a tip, all of whose payoffs are the same. The algorithm can be illustrated by the following annotated game tree. The players are 1 and 2. The player whose move it is, is indicated in the left hand column. In each payoff, the payoff for player 1 is given first. The solution is bolded.



[35] Not everything that is claimed to be a backwards induction is so, however. For example, Basu (1994) describes a game called the Traveler's dilemma. There are two players, A and B. Each, independently of the other, chooses a number, n, $2 \leq n \leq 100$. If both parties choose the same n, both receive n units. But if they choose n and m, where $n > m$, the chooser of m receives $m + 1$, whilst the chooser of n, receives $m - 1$. The standard solution to this game is for both players to choose 2; and Basu claims that this is determined by backwards induction, but it is not. $(2, 2)$ is, indeed, the only Nash equilibrium, but this is determined *directly*. If A knows that B will choose 2 then A will choose 2, and vice versa. Thus $(2, 2)$ is a Nash equilibrium. But if A knows that B will choose $n > 2$, A will choose $n - 1$; and if B knows that A will choose $n - 1$, B will not choose n. Hence for $n > 2$, we do not have a Nash equilibrium. No induction of any kind is necessary here.

[36] It seems to have been proposed first by Zermelo (1913).

https://doi.org/10.1017/S0266267100000250 Published online by Cambridge University Press

There is nothing problematic about this algorithm itself. But neither is it the backwards induction. The backwards induction is an argument to the effect that if each player is rational in a certain sense – namely wishes to maximize their payoff – (and is known to be so) the game will take the solution path. At a penultimate node, whoever's move it is will take whichever move it is that maximizes their payoff. This is therefore the payoff that both players may expect if the game reaches that node. Knowing this, at an antepenultimate move the other player will choose whichever move realizes the position which has the greatest expected payoff for them, as thus computed, and so on. The centipede game and its backwards induction are just, of course, a particular instance of this.

The formalization of the backwards induction for general games is more complex than the special case that we have looked at; but we do not need to go into this here. If the induction worked in the general case, it would work in the particular case of the centipede game – which, as we have seen, it does not. Hence, backwards induction does not tell us how rational players will play in a game of this kind. This poses the question of how they will play. But that is a whole new question.

7. CONCLUSION

Let me conclude by stating what we have learned about backwards induction. Given an α and β satisfying the conditions of the backwards induction scheme, the conclusion that $K_{1n}\beta \vdash \alpha_1$ follows. This is not problematic. What is problematic is whether, on any particular occasion, $K_{1n}\beta$ is true. Often, as we have seen, it is inconsistent, and so cannot be true. But even if it can be true, its truth is not itself evident; and there is, in general, no way of inferring it from $K_1\beta$ without invoking generally invalid principles of inference, notably J2, the persistence of knowledge. If this is the case, backwards inductions, in general, are rarely going to be of any use in establishing sound conclusions.[37]

REFERENCES

Aumann, R. 1995. Backwards induction and common knowledge of rationality. *Games and Economic Behavior*, 8:6–19

Aumann, R. 1996. Reply to Binmore. *Games and Economic Behavior*, 17:138–46

Aumann, R. 1998. On the centipede game. *Games and Economic Behavior*, 23:97–195

Balkenborg, D. and E. Winter. 1997. A necessary and sufficient condition for playing backward induction. *Journal of Mathematical Economics*, 27:325–45

Basu, K. 1994. The traveler's dilemma. *American Economic Review*, 84:391–95

Bicchieri, C. 1989. Self-refuting theories of strategic interaction: a paradox of common knowledge. *Erkenntnis*, 30:69–85

Binkley, R. 1969. The surprise examination in modal logic. *Journal of Philosophy*, 65:127–36

Binmore, K. 1987. Modeling rational players, Part 1. *Economics and Philosophy*, 13:179–214

[37] I would like to thank referees of this journal, and particularly one of its editors, Wlodek Rabinowicz, for many helpful comments on earlier drafts of this paper.

https://doi.org/10.1017/S0266267100000250 Published online by Cambridge University Press

Bonanno, G. 1991. The logic of rational play in games of perfect information. *Economics and Philosophy*, 7:37–65

Broome, J. and W. Rabinowicz. 2000. Backwards induction in the centipede game. *Analysis*, to appear.

Halpern, J. Y. and Y. Moses. 1986. Taken by surprise: the paradox of the surprise test revisited. *Journal of Philosophical Logic*, 15:281–304

Hintikka, J. 1962. *Knowledge and Belief*. Cornell University Press

Jackson, F. 1987. *Conditionals*. Blackwell

Kaplan, D. and R. Montague. 1960. A paradox regained. *Notre Dame Journal of Formal Logic*, 1:79–90

Luce, D. and H. Raiffa. 1957. *Games and Decisions*. Wiley

Pettit, P. and R. Sugden. 1989. The backward induction paradox. *Journal of Philosophy*, 86:169–82

Prior, A. 1971. *Objects of Thought*. Oxford University Press

Rabinowicz, W. 1998. Grappling with the centipede. Defence of backward induction for BI-terminating games. *Economics and Philosophy*, 14:95–126

Reny, P. 1992. Rationality in extensive-form games. *Journal of Economic Perspectives*, 6:105–18

Rosenthal, R. 1981. Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic Theory*, 25:92–100

Sorensen, R. 1988. *Blindspots*. Oxford University Press

Vilks, A. 1997. Analyzing games by sequences of metatheories. *Epistemic Logic and the Theory of Games and Decisions*. Ch. 12. M. O. L. Bacharach, L.-A. Gérard-Varet, P. Mongin and H. S. Shin (eds.). Kluwer Academic Publishers

Wright, C. and A. Sudbury. 1977. The paradox of the unexpected examination'. *Australasian Journal of Philosophy*, 55:41–58

E. Zermelo. 1913. Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels. In *Proceedings of the Fifth International Congress of Mathematicians, Cambridge, 1912*, Vol. II. Cambridge University Press

https://doi.org/10.1017/S0266267100000250 Published online by Cambridge University Press