# Reeling and A-Reasoning: Surprise Examinations and Newcomb's Tale

PETER CAVE

When considering in which equities to invest, which road to take or where best to spend a good night out, bad experiences from decisions past might suggest doing the opposite of what we think. What constitutes an opposition is frequently far from clear, but even when we have to choose between A and B alone, the advice is useless if taken *simpliciter*: were we thinking A and then, as a result of the injunction, switch to B, well, now we are thinking B, so, applying the injunction, we switch to A; and endlessly the switching continues. No appeal to rationality will tell us what to do; no one else will be able to predict, save by luck, what we shall do or rationally should do solely on the information given. Liken us to Buridan's ass, seeking the better bundle of hay when there is no better. Of course, such endless reeling in the reasoning need not threaten: an implicit assumption might be that 'do the opposite' should be deployed only from an *initial* thought of what is to be done. A stopper on the reasoning is now built within the injunction, a threatening endlessness ended.

Miss Jones cannot make up her mind: should she wear the revealing geranium red dress or the demure delphinium blue? She wants to appear confident rather than timid; this points to the red. She wants to look modest rather than a tart, suggestive of blue. Struck by the desire for confidence, she starts dressing in red; yet focusing on the colour, a tart taking shape in her full-length mirror, she switches to the blue; but attending more closely, she cannot escape its blue timidity, leading to a redly return—and so forth. With geranium and delphinium choices having overall equal appeal, Miss Jones's reasoning is unstopped; but a stopper is easily introduced. She must be out by noon, dressed and not bare. On the basis solely of the stated desires, no one—not even Miss Jones—can establish what she will, or should, wear; but with the temporal stopper in place, a predictor, aware that reasoning takes time, might (unlikely, 'tis true) calculate Miss Jones's toing and froing, working out which dress it will be, when she steps forth to some bells' chiming the noon.

doi:10.1017/S0031819104000476

Whatever constitutes the abstraction whereby Miss Jones's reasoning goes *endlessly* from red to blue—whereby doing the opposite rocks us from A to B back to A and forward again, and so on—reasoning on earth needs a beginning and, once begun, reaches an end. It is no surprise that following certain procedures in abstract heavens (were such to be possible) comes to no end; it would surprise, were our earthly procedures to be but the same and to be without end. Certain paradoxes seek to force upon us just such surprising earthly endlessnesses. Certain paradoxes lead us to think that what surely can be done—encountering surprise examinations, making rational choices and the like—cannot be done. The paradoxes, it is here suggested, equivocate between a metaphorical reasoning (in heaven) and a literal reasoning (on earth). The resolution is to recognize that obviously—and without paradox—endless reasonings fix no ends; for ends to be fixed, the reasonings need to be actual, with starters and stoppers. The paradoxical puzzles, at heart, are no more problematical than tales of the ass's eyes swivelling from bundle to bundle; than our always choosing the opposite of what we provisionally choose; and than Miss Jones's repetitive re-dressings in geranium red, then delphinium blue.

Before turning to two of the paradoxes, namely Newcomb's tale and the Surprise Examination[1], versions of which manifest the contradictory demands mentioned, let us introduce the 'Minimal Biconditional Surprise Examination'. The teacher's announcement, given today to a reflective rational pupil—one who seemingly has good reason to believe the teacher truthful—is that the aforementioned pupil will undergo an examination tomorrow if and only if he does not believe that there will be an examination tomorrow. In most cases it is possible that a pupil's belief that *p,* concerning what a teacher might do, could lead a teacher to make that *p* false (though it would be an unusual, even deviant teacher, of course). Indeed, a teacher's seeming to bring about that *p*, or even bringing about that *p*, could lead a pupil not to believe that *p*. However, a pupil who believes a teacher to be bringing about that *p* (and not merely trying to bring about that *p*) cannot rationally, at the same time, be brought not to believe that *p* is on the horizon—save (it seems) for such apparent paradoxes as the Biconditional Surprise.

The pupil, upon hearing the surprising biconditional announcement, applies what he hears to his current belief concerning the matter—the starter on earth. Had he already believed that there would be an examination on the morrow, he now realizes that there

[1] For these paradoxes and some references to the vast literature on them, see Michael Clark, *Paradoxes from A to Z* (London: Routledge, 2000).

will be none; but now, no longer believing one is due, he reasons that there will be one after all, but now…—and so forth. Wherever he starts, the reasoning moves on; and the reasoning lacks any reasoned place to rest. He might step back from this reeling, suspending belief about the relevant examination, but this meta-reflection—that no restful belief can rationally be reached—is still captured by the teacher's biconditional announcement. A reasoning pupil hence realizes that his uncertainty ensures an examination tomorrow—yet this new belief ensures that there will be none, but hence…—and so on. Were the teacher to track the pupil's changing beliefs, whether an examination be set would change accordingly and, it would seem, endlessly. The parties to the tale—and we—might think it paradoxical that the announcement determines no outcome. Could not teachers set examinations on such conditions, even if announced?

Any sense of paradox should be quickly dispelled. The Biconditional Surprise, as so far described, hovers in abstract heavens, fixing no time for the pupil's belief that determines whether the examination will come. Transform the tale. First, the teacher's announcement, given today at 11.00am, is that the pupil will undergo an examination tomorrow if and only if, at noon today, the pupil lacks belief that there will be such an examination. Secondly, reflect that reasoning (on earth) takes time and will start from the pupil's then belief or lack of belief concerning an examination tomorrow. We now see how, when the clock strikes noon, the outcome is determined by the pupil's belief at that time. With time on his hands, a pupil might have no idea what belief (or lack of belief) he will find himself with, when the noon bells sound; but those noon bells remain a reasoning stopper. An examination-adverse clever pupil might deploy a strategy to ensure that, with his reasoning's rocking and a-rolling, he ends up believing an examination will occur just as noon strikes; that is a practical matter—just as we should now see that it is a practical matter whether the teacher is justified in making the announcement and the pupil justified in believing that announcement true.

Let us turn to a version of the traditional Surprise Examination, one in which the teacher's announcement carries the implication that, even on the last possible day of the examination, rational reflective pupils will be surprised at the examination's occurrence—that is, that, having reached the last day, they will not believe the examination will occur on that last day. The pupils' reasoning on that last day might be thus (and they can, of course, reflect on this now): 'There is only today left for the examination, so the examination will be no surprise; hence, no examination will occur… Hold on, that's how the teacher said things would seem to us: we expect no examination; so now we see how the teacher was speaking the truth and an examination will occur after all. Ah, so we are expecting the examination; but hold on…'—and so forth. There is no firm conclusion, so long as the pupils keep revisiting the thought of the teacher as truth-teller; and, of course, if belief in the teacher as truth-teller is simply dropped from the tale, then there is no puzzle even to get started. The puzzle presents the teacher as truth-teller with the pupils rationally believing what she says, yet the teacher can have no good reason to think, either way, what rational pupils will conclude on the last morning, if the examination is left to the last—unless she knows the pupils' speed of reasoning. (We pretend here that there is a pupils' single reasoning *en masse*.) The pupils will have no good reason to expect an examination on any particular day prior to the last day; and, while the pupils' reasoning is left in abstract heavens, with no temporal duration, there remains a puzzle about the last day; but the puzzle remains only because there are no sufficiently determinate conditions for grounding the last day's reasoning.[2] (If an examination is set on the condition that it is a surprise, we edge closer to the Biconditional Surprise above.)

To ask what a rational pupil, in receipt of the traditional Surprise Examination announcement, would or should believe is as silly as to ask what a rational Miss Jones would or should wear or which particular bundle of hay a rational ass would or should choose, given the limited information available. The conditions set by the original puzzle determine no conclusion; so it is no paradox that even the most rational pupil can reach no firm conclusion by reason alone—and there is therefore no reason at all to believe that a teacher could be reliably right in predicting that an examination on the last day would be a surprise to rational pupils[3]. Thus it is that the conditions set set no conditions for resolution; the puzzle's demand for a resolution is unreasonable.

[2] All this is, of course, perfectly compatible with the pupils and the teacher knowing the exclusive disjunction that either a surprise examination will occur on a day prior to the last possible day or an examination will occur on the last day which might or might not then be a surprise. For discussion of this and identification of a contradiction in one version of the teacher's announcement, see Ardon Lyon, 'The Prediction Paradox', *Mind*, **68** NS (1959), pp. 510–17.

[3] Curiously, this is little remarked upon—as is a similar point concerning Newcomb's predictor. See, for example, recent assumptions concerning reliability in these paradoxes made by Laurence Goldstein, 'Examining Boxing and Toxin', *Analysis*, **63** (2003), pp. 242–4.

A recent version of the Surprise Examination is provided by Timothy Williamson: his Conditionally Unexpected Examination is one in which the reliable teacher's announcement is to the effect that there will be no examination over the coming time period such that pupils will know that if there is an examination at all, it will be on that day.[4] This generates a reeling in rational pupils' reasoning, when considering what to think if the morning of the last possible examination day arrives with no examination yet set: in view of the teacher's putative reliability, they might reason, no examination will be set; but, reasoning further, that allows for there being an examination after all; but then—and so on.[5] Once again, the conditions set by the puzzle determine no conclusion; and it is unsurprising that rational pupils can reach no firm conclusion.

In Newcomb's tale, players have the choice of selecting either an opaque box alone or the aforementioned opaque box together with a transparent box clearly containing £10,000. The opaque box's concealed contents are determined by a highly reliable predictor's prediction of the players' selections: when the predictor predicts that players will take both boxes, he boxes the opaque box with nothing; when he predicts that players will take the opaque box alone, he boxes that box with £1,000,000. Given the aim to maximize winnings, what should a rational player choose? A big tripartite assumption—one to be accepted, but temporarily only—is that it is rational for such players to believe that the predictor is successful in predicting the choices of *rational* players, yet that rational players, making their choices based on reason (in some way), can win at least £1,000,000, and that there is one right answer to the question of what it is rational to do.

In view of the predictor's success, it seems that rational players should take the opaque box alone; it might seem that that is an easy end of the matter—but these matters lack such ease and such ends. Players, grasping the above, reason further. 'As the predictor will therefore have placed the £1,000,000 in the opaque box, let us take both boxes (securing the additional £10,000). Yet, given the predictor's success, he will have taken that reasoning into account,

[4] Timothy Williamson, *Knowledge and its Limits* (Oxford: Oxford University Press, 2000), pp. 143–6.

[5] Williamson (ibid. pp. 135–46) treats versions of the Surprise Examination as akin to his tales of pupils' glimpsing an examination date ringed somewhere round the middle of the calendar displaying all possible examination dates. The kinship is difficult to accept, given that some Gimpsing Tales firmly rule out an examination on the last day in contrast to typical versions of the Surprise Examination.

so, if we take both, there will be no £1,000,000; thus, as it is mutually known that we seek maximum winnings, it is better to select the one concealed box alone, for the predictor will have predicted that. Of course, if he has, then the £1,000,000 is already there…'—and so the players reel in their reasoning.

A quandary results from the question: can the predictor be successful at predicting, when there is such endless reasoning flux by the players? The answer is: *if* there is just the aforementioned reasoning to go on—and so, no stopper—the predictor can no more be reliable in his predictions than players can decide on rational grounds what to do, given the tripartite assumption. Of course, one could reject one element in the assumption, namely that rational players can win at least £1,000,000 when choosing on the basis of reason; one could tie rationality to the dominance of a causal principle—'what is in the opaque box has already been fixed'—and then, if the predictor is reliably right, the rational can win only £10,000 by rational choice. The continuing debate in the literature by the rational over what it is rational to do is some evidence against such a tight causal tie round rationality. In any case, there can be higher-level rationalities that permit lower-level irrationalities—that is, it can sometimes be rational to act irrationally—and such higher-level rationalities would ensure that the reeling continues.

With Newcomb's tale, as it is, there is a practical ending—a selection is made—and that occurs at a point in the reasoning, but it cannot occur as a settled conclusion on the basis of rational reasoning; this is because of the flux aforementioned. It occurs because players tire or because they are, for example, disposed to grab at two rather than one or because the selection must be made by noon. Once such stoppers are introduced—and they are not justified by the reasoning on the information given—it becomes possible (albeit unlikely) that a predictor could have successfully predicted the players' choices. Rational players, aware of such irrationalities, might, of course, build such factors for the predictor's success into their reasoning, but that just spins the reasoning off yet again; and, yet again, to ask what rational individuals (as players this time and not as pupils) would or should do, on the basis of their reasoning alone, is as silly as asking our Miss Jones to work out, by reason alone, which dress to wear, the rational ass which bundle to select, and ourselves what to conclude when doing the opposite of what we conclude.

W. E. Johnson[6] stressed that inferences are processes conducted

[6] See this suggested in his *Logic* Part II (Cambridge: Cambridge University Press, 1921), pp. 1–10.

by people—and let us stress, by people here on earth. The conditions of the surprise examinations and Newcomb's tale depend on participants engaging in such processes. Validity and soundness place constraints on what should be inferred, but they neither start nor stop the reasoning processes. For inferences to be made, we need—for a start—inference makers sufficiently motivated to make a start; and, for inferences to cease, makers need reason (which validity alone cannot supply) to go no further. From *p* together with *if p then q*, makers, relevantly motivated, might well conclude *q*; but they might not stop there: they might further conclude *p and q* and *q or r* and *q* yet again—unless resistant to repetition and futility. Solely from premises *p* and *if p then q*, with conclusion drawn *q*, it is, of course, irrational to wonder whether, if one repeats that argument, then *not q* might put in an appearance. So, money-seekers offered money for validly reaching (from both premises mentioned) a conclusion that cannot itself be undermined by the argument would typically be irrational, if they dallied for long before committing to *q* as a conclusion. With the surprise examinations and Newcomb's tale, things are otherwise: on the information given, reflective reasoners find that there is no good reason to stop at one conclusion rather than the opposing other, for any seeming conclusion is destablized by reflection on how it stands with regard to that information. Reasonings continue to reel—unless and until an empirical temporal stopper is imposed.

There is a kinship here with trying to satisfy explicitly inconsistent demands. Miss Jones receives a conjunctive instruction: the first conjunct tells her to wear the red dress only, so she dresses accordingly; the second demands that she wears only the blue, so off comes the red and on goes the blue. She checks the instruction and sees that she now offends the first element; so she steps out of the blue, and slips on the red—and so on. A wiser Miss Jones would not get started on this endless quest of satisfying a demand which, if grasped as a whole, is paralysing.[7]

If rational, usually we can rest with conclusions only when having no reason to think any further reasoning would undermine those conclusions. Often we can and should rest. Newcomb's tale and the surprise examinations highlight reelings in the reasonings—of abstract players and pupils, of predictors and teachers, alike. These puzzles puzzle because they start off such endless reasonings, yet

[7] For a discussion of some reelings and paralyses deployed by certain jokes and other paradoxes, see my 'Humour and Paradox Laid Bare', *The Monist*, **88.1** (2005).

inconsistently demand that examinations get set and boxes filled on the basis of the outcomes of such reasonings. Of course, surprise examinations can get set and boxes do get filled, but courtesy of the intrusive empirical world, not solely of the endless reasoning as unreasonably demanded. Clocks strike, eyelids close and choices get made by hunch or by whim. It is by way of such empirical intrusions that reasonings on earth—happily—do in fact reach some ends; and it is only by way of the anticipation of such empirical intrusions that the puzzling announcements of the paradoxes discussed could ever rationally be given.

Is it not paradoxical that rationality should set us endlessly reeling in the puzzles' puzzling ways? No more so than it should be paradoxical that by adding one to each number that we consider, our outcomes reel from evens to odds and odds to evens. A god of infinite power would reel thus as well; and a god of infinite power, as player or predictor, pupil or teacher—as a Miss Jones, as ourselves or even as ass—can no more reach a conclusion by reason alone, than can we mere finite beings that we are.

Had we looked even into God's mind, we should find no conclusion there.[8]

*Soho, London*

[8] My thanks to Michael Clark, Laurence Goldstein, Ardon Lyon and Jerry Valberg for stimulating discussions on paradoxes over the years.