Counterfactuals in Wonderland

Dov Samet

March, 1997

In a recent paper, [1] Professor Robert J. Aumann has given a formal presentation and a proof of the claim that rational players, in a game with perfect information, whose rationality is common knowledge among them, must play according to backward induction strategies.

What keeps the players in Aumann's model from deviating from the backward induction path is the following. Each player knows that had he deviated from the backward induction path, then all players would have followed the backward induction path that starts at the node that follows the deviation. This new backward induction path is, by definition, worse for the would be deviating player, than the original backward induction path, which is precisely the reason why the player does not deviate.

Why would the players have followed the new backward induction path, after a deviation which, of course, violates the common knowledge of rationality assumption? Because common knowledge of rationality implies that players know (indeed commonly know) that even had they deviated from the backward induction path, common knowledge of rationality would still have been preserved for the rest of the game.

In his paper Aumann tirelessly emphasizes the unavoidable role of counterfactuals in game theory and decision theory. The whole argument hinges on what players know would have happened if things that wouldn't and couldn't have happened, happened.

This little note tries to trace the literary and cultural sources of the main ideas in Aumann's article. There is more than just a hunch, there is textual

[1] Backward Induction and Common Knowledge of Rationality, *Games and Economic Behavior,* Vol. 8, 1995.

evidence, that Professor Aumann wanted, at least subliminally, to lead us to his sources, by planting a very broad hint in his paper.

One of the most striking examples in Aumann's paper is a two-player game played by Ann and Bob. These names were chosen on purpose in alphabetical order to belittle their significance, as it were. But we are not deceived. Restricting names to those starting with A and B is a minor imposition that still leaves a lot of latitude.

There is no difficulty understanding why Aumann chose the name Bob. It is practice of many artists to give themselves a role, usually a minor one, in their pieces. Hitchcock, for example, can be seen in almost every one of his movies in some insignificant role. We are not surprised then to find Bob Aumann playing the role of a player in a plot he has so carefully designed.

The other name, Ann, is a harder nut to crack. Why Ann? This name is suspiciously conspicuous when we note that in numerous articles and presentations the names of the players are *Alice* and Bob. By changing the name of his female protagonist, which is so common in this context and therefore passes unnoticed, Aumann is trying to draw our attention to the omitted name Alice. Who is this Alice, whose name Aumann is using to play hide and seek with us? Assuming that he was not trying to be too obscure, we must look for a fairly familiar character, which leaves us with very few choices, indeed one, the heroine of the writer-logician the Rev. Charles Ludwidge Dodgson, better known by his pen name Lewis Carroll—Alice in Wonderland. It is in this book that we should look for a key to understanding Aumann's paper.

Of the two part book we should concentrate, for our purposes, on the more game theoretic one, *Through the Looking-Glass and What Alice Found There*, the plot of which takes place, from start to end, on a chess board. We zoom in on the middle of this book, the sixth chapter which is titled: Humpty Dumpty.

This is the first impression Alice has of the protagonist of this chapter.

> Humpty Dumpty was sitting, with his legs crossed like a Turk, on the top of a high wall—such a narrow one that Alice quite wondered how he could keep his balance...

The story of Humpty Dumpty's life, his very existence, is a story of striking the right balance, one of staying in *equilibrium.* Humpty Dumpty *is*

the equilibrium or at least a metaphor of it, ever so delicate and fragile, yet sophisticated and bold and at the same time proud and vain.

Equilibrium, by definition, is a choice of one alternative, precluding many others. Entertaining the other alternatives, those that do not materialize, is the essence of *counterfactuals*. Humpty Dumpty, with his keen conceptual analysis, gives us a crisp understanding of the simple logical structure of counterfactuals.

> Why, if ever I *did* fall off—which there's no chance of—but *if* I did—" Here he pursed up is lips, and looked so solemn and grand that Alice could hardly help laughing. "*if* I *did* fall," he went on, "*the King has promised me*—ah, you may turn pale, if you like! You didn't think I was going to say that, did you? *The King has promised me—with his own mouth*—to—to—"
>
> "To send all his horses and all his men," Alice interrupted, rather unwisely.

Here we skip a couple of lines, to which we return later, to the completion of Humpty Dumpty's answer.

> "Yes, all his horses and all his men," Humpty Dumpty went on. "They'd pick me up again in a minute, *they* would"

We can hardly fail to hear Humpty Dumpty's words reverberating in Aumann's paper. Why, if ever the players, who have common knowledge of rationality, *did* deviated from the backward induction path—which there is no chance of—but *if* they *did*... Well, the restoration of common knowledge of rationality, requires neither the King's horses, nor his men, nor indeed any artificial, *deux ex machina,* outside intervention to put back and restore the failing common knowledge of rationality. It is common knowledge of rationality *itself* that guarantees, in a breath-taking logical feat, that no matter how many times it is shattered it resurrects itself, like the legendary Phoenix, spreads its wide wings and soars to heavenly (im)possible worlds, untouched by cruel reality and facts. [2]

[2]The magical power of common knowledge of rationality is the result of Aumann's particular definition of rationality. I explain it in detail in a paper titled Rationality, Counterfactuals, and No-matter-what Theories, which I would have published had I only been rational. Aumann did not really have much choice in defining rationality in the

Now let us go back and listen to Humpty Dumpty's response to Alice's unwise interruption.

> "Now I declare that's too bad!" Humpty Dumpty cried, breaking into a sudden passion. "You've been listening at doors—and behind trees—and down chimneys—or you couldn't have known it!"
>
> "I haven't, indeed!" Alice said very gently. "It's in a book."

What a marvelous piece of interactive epistemology! The issue at hand is the epistemological status of this counterfactual. Humpty Dumpty, with some rudeness, accuses Alice of eavesdropping. That is, he claims that Alice has gained her knowledge of what he knows, concerning the said counterfactual, by listening behind doors, preventing him, this way, from knowing that she knows that. Equipped with modern day logic of epistemology, which Carroll, the logician, did not unfortunately have at the time, we can succinctly summarize Humpty Dumpty's claim by the formula $K \quad {}_2K_1C\&\neg K_1K_2K_1C$, where Humpty Dumpty is 1 and Alice is 2.

Alice's answer "It's in a book" is a beautiful informal way Carroll found to express the idea of common knowledge. What is in a book is known by any literate person, and moreover, every such person knows that it is read by every such person, and so on. [3]

What was the end of Humpty Dumpty? The last lines of Chapter VI describe Alice's departure from Humpty Dumpty, who no longer pays attention to her.

> "...but she couldn't help saying to herself, as she went, "Of all the unsatisfactory—" (she repeated this aloud, as it was a great comfort to have such a long word to say [4]) "of all the unsatisfactory people

special way he does. It was forced on him by the restricted power of expression of his model. I explain it in Hypothetical Knowledge and Games with Perfect Information, *Games and Economic Behavior,* Vol. 17, 1996, and show how a richer power of expression enables a definition of rationality in down to earth terms. See also footnote 6.

[3]Does this mean that Humpty Dumpty was illiterate? This bold conjecture deserves a separate study. The text seems to support the conjecture. When Alice gives Humpty Dumpty her memorandum-book, in which she worked out the difference $365 - 1 = 364$, Humpty Dumpty approves the result in a somewhat suspicious way—he holds the book upside down.

[4]No question Carroll would have preferred the word *counterfactual* to *unsatisfactory,* had he known it. The two words have the same number of letters and share *fact.*

> I *ever* met—" She never finished the sentence, for at this moment a heavy crash shook the forest from end to end.

Did the King's horses and men come to pick Humpty Dumpty up again? Carroll does not give us a clear answer. At the beginning of the chapter, he puts in Alice's mouth the wonderful words of the famous poem from *Mother Goose*:

> *Humpty Dumpty sat on a wall:*
> *Humpty Dumpty had a great fall.*
> *All the King's horses and all the King's men*
> *Couldn't put Humpty Dumpty in his place again.*

Yet it is not clear that this prophecy was realized completely. Assuming that the great fall was indeed irreparable, how many times was Humpty Dumpty proven wrong? Clearly he was wrong in being so sure there is no chance he would fall. A great scholar who dived into the unspeakable depths of the theory of counterfactuals, Professor Itzhak Gilboa, claims that Humpty Dumpty was also proven wrong after the fall; no horses and no men came to help him (if indeed this was the case).

This point of view seems to reflect a complete misunderstanding of the nature of counterfactual thinking. The counterfactual "If I *did* fall..." is meaningful *only* when Humpty Dumpty does *not* fall. It is meaningless, and therefore cannot be tested, after Humpty Dumpty does fall. This is the special beauty of counterfactuals. I hope that these words restore some of Humpty Dumpty's lost honor.

The Humpty Dumpty episode is deeply rooted in ancient cultural and religious traditions, and it hosts archetypal images that have accompanied humanity from time immemorial. It is a story of death and resurrection, of fall and restoration, of facts and counterfactuals, of reality and myth. It is dressed in many shapes and forms, and the names of the protagonists of this drama vary from culture to culture. But there are always the dying-falling male and the lamenting, giving rebirth, female companion. Dumuzi (Tammuz) and Ishtar in Mesopotamian mythology, Isis and Osiris in the Egyptian pantheon, Adonis and Aphrodite for the Greeks, Jesus and Mary [5]

[5] The trinity of Jesus the male, is mirror imaged by the three Marys, who are all related to birth-resurrection, Mary the mother, Mary Magdalene, and the other Mary, the mother of James.

in the Christian tradition, and last but not the least, Humpty Dumpty and Alice. The immortal lines of Alice's lamentation *"Humpty Dumpty sat on a wall:/Humpty Dumpty had a great fall."* do not fall short in poetic force and dramatic impact of the strongest verses in Ishtar's lamentation on Tammuz death, *"The wild bull who has lain down, lives no more,/ the wild bull who has lain down, lives no more,/ Dumuzi, the wild bull, who has lain down, lives no more."*

Some features of the more traditional stories are still preserved in the Carrollian story. The first and most obvious one is the phonetic resemblance Dumpty-Dumuzi. A second one is more subtle. The death-resurrection theme is understandably related to ancient fertility rites. One of these rites Easter (=Ishtar), is observed to this day. The trappings of Easter, bunnies and eggs, are, what else, ancient symbols of fertility. What could be more appropriate than to portray Dumpty-Dumuzi-Tammuz, the fertility God, as an egg? Sir John Tenniel's illustration of Humpty Dumpty, which has accompanied the book since its first publication, especially the beautiful belt, or cravat as Humpty Dumpty insists, he wears, strongly suggests a painted Easter egg.

The theory of common knowledge of rationality, as is shown here, is a highly intellectualized and abstract version of this old death-resurrection motif. It is a stylized story about the counterfactual resurrection of falling common knowledge of rationality. It draws directly, as I have proved, from Carroll's work, but it is influenced, no doubt, from much deeper currents of human thought and experience which are expressed in numerous stories and rites. No wonder the debate concerning this theory is so heated and emotional. On one hand, it touches a universally sensitive nerve, and as such it has tremendous appeal. On the other hand, in these post modern days, it raises almost automatic resistance, as an obsolete discourse of oppressing institutions and hegemonies.

The theory differs though in some important aspects form the old death-resurrection myths. The traditional functional male-female dualism is absent here. Alice and Bob are a pale reminder of this dualism. But the real protagonist of the theory is the abstract notion of common knowledge of rationality, which plays both roles, of male and female, in the best tradition of political correctness.

As a final remark and food for thought we should ask ourselves which protagonist in Lewis Carroll plays the role of rationality. After all, it is this

notion, as defined by Aumann, that gives common knowledge of rationality its out of this world [6] power of self resurrection. It is this notion, more than anything else, that seems to come directly from wonderland. I strongly believe that this should be looked for in another book of Lewis Carroll, the great epic *The Hunting of the Snark*. I suggest that Carroll is summarizing his quest for rationality in the unforgettable awesome bottom line of his epic:

> For the Snark *was* a Boojum, you see.

But that is another story.

[6] This phrase is meant literally, not literarily. By Aumann's definition, rationality of an agent is tested not only by the choices he makes in this world, but also by the ones he makes in counterfactual worlds—in situations which the agent is commonly known not to face. There is nothing wrong in Aumann's idiosyncratic use of the word rationality. He is entitled to hold Humpty Dumpty's view: "When *I* use a word it means just what I choose it to mean—neither more nor less."