WHY THE EMPTY SHELLS WERE NOT FIRED:
A SEMI-BIBLIOGRAPHICAL NOTE[*]

by

Itzhak Gilboa[**]

February 1992, revised December 2010

Abstract

This note documents Aumann's reason for omitting the "empty shells" argument for the common prior assumption from the final version of "Correlated Equilibrium as an Expression of Bayesian Rationality." It then continues to discuss the argument and concludes that rational entities cannot learn their own identity; if they do not know it a priori, they never will.

[*] I thank Robert Aumann, Cristina Bicchieri and Paul Weirich for comments and references. NSF Grant SES-9113108, ISF Grant 396/10 and ERC Grant 269754 are gratefully acknowledged.

[**] Tel Aviv University, and HEC, Paris. tzachigilboa@gmail.com

1. Motivation

In the working paper version of "Correlated Equilibrium as an Expression of Bayesian Reality" (Aumann (1985)), Professor Aumann defended the common prior assumption by an "empty shell" argument (among others). This argument is not mentioned in the final version of the paper (Aumann (1987)).

In a personal conversation, Professor Aumann explained the reason for the omission, which was basically that he discovered that the original argument was flawed. The flaw is, in my view, more subtle and profound than the original argument.

The goal of this short note is to document this point, as well as the ensuing discussion that took place between Profs. Aumann, Schmeidler, and me.

The rest of this note is organized as follows: Section 2 discusses the original argument. Section 3 explains the problem. Section 4 describes some additional implications that were raised in the discussion that followed. Finally, Section 5 concludes with some remarks.

2. The Argument

The common prior assumption (CPA) states that all players have the same prior on the (measurable) space of states of the world. As explained in Aumann (1985, 1987), the informal assumptions that the information partitions as well as the priors are common knowledge are actually results rather than assumptions, if one takes Savage's (1954) idea of a state of the world "resolving all uncertainty" to the extreme. (See also Aumann (1976).) (For various formalizations of this idea, see Mertens and Zamir (1985), Brandenburger and Dekel (1987), Fagin, Halpern, Moses, and Vardi (1995), Kaneko and Nagashima (1996,1997), and Aumann (1999a,b).)

The CPA, however, does not seem to follow from a similar reasoning, yet it is certainly needed to distinguish correlated from rationalizable equilibria (see Aumann (1985, 1987) and Brandenburger and Dekel (1987)).

The "empty shell" argument runs, roughly, as follows: players may have different beliefs (priors) due to different information they acquired during their lives. Theoretically, one may try to model all this learning as simple Bayes' update of a prior one has at birth. Unrealistic as this may sound (or, indeed, be), this story could still be (a part of) a viable model, and a similar type of argument is to be found in Savage (1954) regarding the reduction of all decision problems in one's life to one "grand decision."

Yet people surely differ even at birth. For instance, they have different genes which may determine both their utilities and priors. Therefore, the argument goes one step further and considers the "players" before they acquired the information contained in their genes. Thus, we are asked to think of some intelligent entity capable of logico-mathematical reasoning but which does not yet know what actual player it will materialize in. At the moment of birth (or conception, or even much earlier, depending on the reader's faith and social policy preferences), this intelligent entity--the empty shell--learns the genes it got, updates its prior and becomes a "regular" player with a utility function and beliefs that are now the posterior.

However, the "empty shell" argument concludes, before learning the genes, there is no reason to distinguish between these empty shells. They are all identical, since any distinction among them is assumed to be learned later on. In particular, they all have the same prior.

This type of argument, although unpalatable to most economists, appears in the philosophical literature. (Aumann (1985) mentions Rawls and Harsanyi as well as Rousseau.) If nothing else, it is an interesting exercise that may also help us delineate the scope of the Bayesian approach.

Before presenting the main problem with this argument, it is worth mentioning two points, which are not made explicit in Aumann (1985).

First, the conclusion of the argument draws on some "insufficient reason" principle. In Aumann's words (1985, p. 18), "There cannot be any reason to distinguish between the empty shells; they must have the same priors, and we are back at the CPA."

Although this is admittedly quite intuitive, it falls short of a proof. By comparison, the claim that the priors are common knowledge is a theorem if one assumes that a prior refers to the trivial partition of W and each w in W specifies all the relevant aspects of the model. Put differently, if one has a model in which these concepts are formalized, the quotation marks can be dropped from "proof" and "theorem" above. Yet, even in such a model, another explicit assumption needs to be made for the identity of the empty shells to follow. It does not follow from the notion of a state of the world as "resolving all uncertainty."

To further clarify this point, notice that the logic of the empty shells argument is biased towards identicality. It asks why people differ and attempts to go back to a point of time preceding the cause of the difference. Almost symmetrically, one may ask why different people should be identical in whatever respect and try to consider the times when they still differed. Often it is identicality rather than discrepancy that arouses scientific curiosity, which means that differentiality would be the more intuitive assumption. (For instance, the peculiar fact that people share beliefs may be prodding us to study learning procedures that may cause different priors to converge. See, for instance, Blackwell and Dubins (1962), Kalai and Lehrer (1993, 1994).) Since, in general, curiosity is aroused by a mix of regularity and irregularity, analogy and distinction, the question "Why should people differ?" may well be countered by "Why not?". Thus, the insufficient reason principle is a non-trivial assumption implicitly made by the empty shells argument.

As explained by Aumann (1992), the empty shells argument was partly motivated by the quest for a "contentless" framework. According to this approach, which may be dubbed "classical," the model of states-of-the-world and (if possible) a prior on them should be "unprejudiced" in the sense that, in and of itself, it says nothing. It is "form without content." Only when updated in face of additional information does the model acquire substance by restricting the possible utilities, priors, the game being played and so forth. At the outset, however, the model should be "tautological."

If such a model existed, different priors would indeed seem inconsistent with it, since the fact that different players (or empty shells) have different priors is not tautological by any stretch of imagination. But in line with the points made above, identical priors do not seem to be tautological either. Furthermore, even in a one-person (or one-shell) model, while one may formulate states of the world which presuppose nothing (but the language), it is not clear whether any choice of prior could be thought of as "tautological" or "contentless" in the same sense.

The second conceptual point is the meaningfulness of empty shells' priors. One extreme view of empty shells is that they are (almost) nothing but the logico-mathematical entity needed to "understand" the model and reason about it. According to this view, they are free from all that is mundane, and, in particular, have no preferences. Loosely, pure logico-mathematical entities simply don't care (about anything).

But if empty shells do not (yet) have preferences, one may not attempt to derive priors from them a la Ramsey, de Finetti and Savage, and the concept of "prior" becomes somewhat metaphysical.

Another (again, extreme) view (Aumann (1992)) is that empty shells are just like regular players in every respect, and, in particular, they have hopes and desires, fears and aversions--in short, preferences. In such a model, a "consequence" for an empty shell would probably specify the player it materializes in, the game played and the game's outcome. The empty shell's utility for such a consequence can then be identified with the player's utility for the game's outcome.

This view of empty shells is consistent and even "tautological" in the sense that the empty shells' utilities are naturally derived from the very definition of a given consequence. Yet it is still not entirely clear what do these utilities mean. After all, empty shells exist (in principle) only prior to any choice situation. Their preferences therefore have to be purely hypothetical. One may feel uneasy with such a "behavioral" foundation for preferences, and even argue that the empty shells cannot even answer hypothetical questions, since they cannot even perform speech-acts.

To sum, it appears that the notion of "empty shells" has some metaphysical elements whichever interpretation is chosen. This casts at least a pale shadow of doubt on the theoretical validity of the empty shells argument, at least within the framework of modern economic theory.

3. The Flaw

Yet none of the problems mentioned above is as illuminating nor as conclusive as the following argument, due to Aumann (1986): consider two empty shells, say a and b. It may be useful, though not necessary, to think of them as computer programs and/or lists of facts they know. Assume that these programs/lists are identical, which means that they contain precisely the same "lines" or "propositions" in whatever language is used. In other words, a and b are subjectively identical. However, they may well differ objectively. That is, after these proposition lists are translated (by substitution) to an "objective" language, they may no longer be identical. For instance, assume that both a's and b's proposition lists include the statement "I am smarter." Thus, a ascribes a prior 1 to the event "a is smarter than b" and b ascribes probability 1 to the event "b is smarter than a."

"But wait," cries the reader, "Weren't the empty shells supposed to be perfect logico-mathematical reasoners? How could they believe one is smarter than the other?" A fine point, indeed, but it does not seem to apply to the proposition "I am more likely to materialize as player 1 than the other empty shell is."

It seems, indeed, that a "real" empty shell should be free of such prejudice. Yet this is another application of the insufficient reason principle: not only do we need to assume that the shells are subjectively identical, we also need to further assume that they are objectively identical. Whether one is willing to make these assumptions is, as always, a question of personal taste (and therefore depends on one's genes). The crucial point is the distinction between subjective and objective identity relations.

4. The Fundamental Difficulty

It is quite obvious that the problem described above is related to representation. The same "objective" event "a is smarter than b" is represented in a's language as "I am smarter than the other shell," while in b's language it is written as "the other shell is smarter than me." Thus, it is tempting to let our empty shells contain only objective-representation propositions, such as "a is smarter than b," "b is smarter than a," and so forth.

But then one has to add to their "knowledge base" the statement "I am a," and "I am b," respectively, which allow for the rest of their knowledge to be deduced by substitution. One cannot avoid these identity statements for two related reasons. First, since any notion of an intelligent empty shell knows its own identity, the omission of these statements would not result in an intuitively equivalent (or equivalently intuitive) "knowledge." In other words, even though much of game theory may sidestep this issue, we cannot claim to have satisfactorily modeled "knowledge" of intelligent, introspective entities without capturing their knowledge of their own identity. Second, without these statements the empty shells will not know which player's utility they are supposed to be maximizing, even after their materialization in players.

However, the identity statements obviously make the empty shells non-identical even subjectively. Given the type of reasoning that brought empty shells into existence in the first place, one can hardly avoid trying to go yet another step backwards. Couldn't we think of even emptier shells, say g and d, which are going to become a and b (or b and a), yet do not yet know which would be which?

Indeed, we need not use two levels here. Consider two players, say Yisrael and David. When they come to play the game they know not only their identity but also their utilities and priors. However, before the game begins they are only empty shells a and b, neither of which knows its identity. That is, a complete description of their knowledge does not use the term "I".

Now, according to the empty shell story, God sends down an angel (say, Michael), who is going to whisper in a's ear: "You are Yisrael" and in b's ear: "You are David."

At first there may seem to be nothing wrong with these statements--"I am Yisrael" and "I am David"--popping up miraculously in a's and b's respective "knowledge base." Yet, upon closer inspection, such empty shells cannot be considered fully rational. For, if Michael whispers in a's ear that it is actually Yisrael, a should have known that it was attached to this ear from the outset. That is, a knew that "I am the empty shell attached to ear number 1."

In other words, a shell, empty or not, must have some notion of its self in order to arrive at one. At least the way by which it acquires information should have distinguished it from other shells.

To repeat the same argument from a different angle (and, if you will, with a different angel), suppose that a and b are two shells with no notion of their identity. They sit in a room and listen to a radio broadcast which is supposed to convey their identities to them. Being identical before the broadcast, and processing the same information in the same way (say, Bayes' update), they will naturally be identical when the broadcast is over. If they have absolutely no way to define themselves, the little angel in the radio cannot use statements like "The shell sitting by the window is David." (To be precise, such statements will simply not be informative enough since no shell knows that "I am the one sitting by the window.") To sum, if you do not know who you are, there is no way to tell you that.

5. Concluding Remarks

a. The way people actually obtain knowledge of their identity and develop the notion of the self has obviously little (or nothing) to do with our discussion here. In fact, our arguments imply that one's identity indeed has to "pop up miraculously" in one's mind or, equivalently, that one cannot be fully rational when learning one's identity. For instance, an infant may learn its identity without being aware that it is the person attached to its ear, or that other people may be hearing different things.

By contrast, empty shells should be viewed as an exercise, studying how far one can go back with "rational" entities who represent information symbolically and process it according to some "rational" updating function. From this perspective it seems that one would have to concede that empty shells could never be in completely identical epistemic positions (even subjectively), or that they have to be less than perfectly rational in ignoring some way by which they could identify themselves.

b. The fact that (rational) empty shells cannot be in completely identical epistemic positions may have implications beyond the common prior assumption. As a matter of fact, almost every argument in moral or political philosophy that refers to an "original position" (Rawls (1971), Harsanyi (1953, 1955)) and that thereby treats individuals as empty shells makes an implicit or explicit assumption of identicality of the shells. Since in most cases they are still taken to be rational (and to have beliefs), such arguments are weakened by the difficulties pointed out above.[1]

c. A related discussion appears in Lewis (1983, Ch. 10), who argues for representation of beliefs by properties rather than propositions. There seem to be two major differences between his point and ours: first, as long as propositions may use a subjectively-defined term as the pronoun "I," we do not make any claim regarding the limitations of propositional beliefs. (This point may be crucial for discussions of artificial intelligence and its scope.) Second, our main point is not merely that the statement "I am a" contains non-trivial informaiton. Rather, it is that without some statement of this sort, none could be learnt, and that, therefore, empty shells cannot be identical.

[1] This comment is due to Cristina Bicchieri.

References

Aumann, R. J. (1976), "Agreeing to Disagree," Annals of Statistics, **4**: 1236-1239.

______ (1985), "Correlated Equilibrium as an Expression of Bayesian Rationality," working paper.

______ (1986), personal communication.

______ (1987), "Correlated Equilibrium as an Expression of Bayesian Rationality," Econometrica, **55**: 1-18.

______ (1999a), "Interactive Epistemology I: Knowledge," International Journal of Game Theory **28**: 263-300.

______(1999b) "Interactive Epistemology II: Probability," International Journal of Game Theory **28**: 301-314.

______ (1992), personal communication.

Blackwell, D. and L. Dubins (1962), "Merging of Opinions with Increasing Information," Annals of Mathematical Statistics, **38**: 882-886.

Brandenburger, A. and E. Dekel (1987), "Rationalizability and Correlated Equilibria," Econometrica, **55**: 1391-1402.

Fagin, R., J. Y. Halpern, Y. Moses, and M. Vardi (1995), *Reasoning About Knoweldge*. Cambridge: MIT Press.

Harsanyi, J. (1953) "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking", Journal of Political Economy, **61**: 434-435.

______ (1955), "Cardinal Welfare, Individualistic Ethics and Interpersonal Comparisons of Utility," Journal of Political Economy, **63**: 309-321.

Kalai, E. and E. Lehrer (1993), "Rational Learning Leads to Nash Equilibrium," *Econometrica*, **61**: 1019-1045.

______ (1994), "Weak and Strong Merging of Opinions," *Journal of Mathematical Economics*, **23**: 73-86.

Kaneko, M. and T. Nagashima, (1996), "Game logic and its applications I", Studia Logica, **57**: 325-354.

______ (1997) "Game logic and its applications II", Studia Logica, **58**: 273-303.

Lewis, D. (1983), Philosophical Papers, Vol. 1, New York and Oxford: Oxford University Press.

Mertens, J.-F., and S. Zamir (1985), "Formulation of Bayesian Analysis for Games with Incomplete Information", International Journal of Game Theory, **14**: 1-29.

Rawls, J. (1971), A Theory of Justice, Cambridge: Harvard University Press.

Savage, L. J. (1954), The Foundations of Statistics, New York: Wiley.