GAMES AND ECONOMIC BEHAVIOR **20,** 25–30 (1997)
ARTICLE NO. GA970508

# A Comment on the Absent-Minded Driver Paradox*

Itzhak Gilboa

*MEDS / KGSM, Northwestern University, Evanston, Illinois 60201*

Piccione and Rubinstein (1996) present the ‘‘absent-minded driver paradox,’’ which shows that if a decision maker’s information set is allowed to intersect a decision tree path in more than one node, inconsistencies may arise. Specifically, the decision maker may wish to change her choice without any ‘‘intrinsic’’ change in preferences and without receiving any new information, apart from the mere fact that she was called upon to act. I argue that decision problems can and should be formulated in such a way that information sets do not contain more than one decision node on each path. Such formulations will not be subject to the paradox. More importantly, they follow from the classical lore of decision theory. Differently put, the absent-minded driver paradox is a result of decision modeling which violates some of the basic, though often implicit, foundations of decision theory.

*The Example*. Piccione and Rubinstein’s simplest paradoxical example involves a driver who may take one of two exits from a highway. However, at any decision node she would not be able to tell which exit she is facing. They model the decision problem by the tree shown in Fig. 1.

The paradox will not be presented here. Rather, the focus of this comment is on the very structure of this decision tree.

*An Alternative Formulation*. One may model the problem as follows: Imagine two identical agents of the decision maker, called *a* and *b*. Nature would decide—say, with equal probabilities—whether *a* or *b* will be called upon to act at the first decision node, and then it would evoke the second agent to make a decision at the second node. Each of the two agents thinks of herself, when called to make a decision, as ‘‘the self,’’ and of her twin as ‘‘the other.’’ That is, while making the choice, each agent knows that she is the one who is currently deciding, and by this very fact she distinguishes herself from her twin. She does not know whether she acts first or second, but she has a way to define her self, and thus she considers

*I thank an anonymous referee for comments and corrections. Current address: Department of Economics, University of Pennsylvania.

0899-8256/97 $25.00

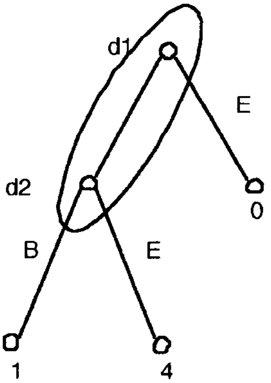

FIGURE 1

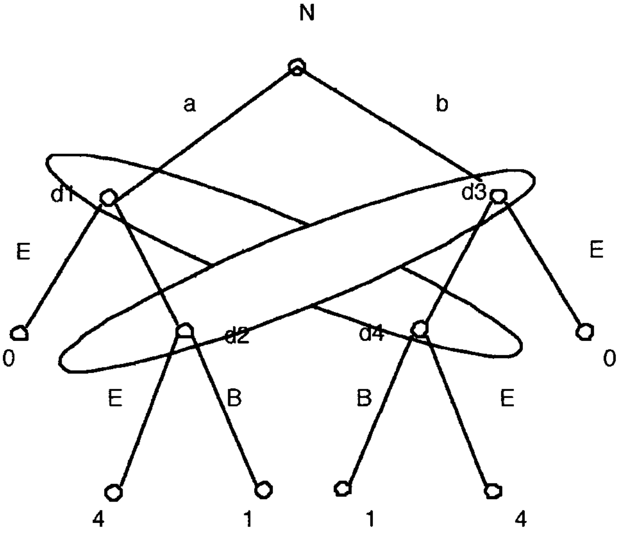

FIG. 2. The payoffs of the agents are identical.

the twin's decision as independent of hers. This approach suggests the game shown in Fig. 2.

Obviously, this is a game of perfect recall, and its analysis poses no special difficulties. One may compute Nash equilibria of the agents game, and these would correspond to optimal choices given their beliefs about each other. In particular, a symmetric Nash equilibrium would allow each agent to believe her twin is identical to herself in terms of strategies, beliefs, and so forth. However, when each agent is contemplating a ‘‘deviation’’ to an alternative strategy, that is, when she tests the optimality of a ‘‘recommended’’ strategy, her beliefs over her information set do not change as a result of her contemplated move. Thus this formulation makes each agent’s beliefs dependent only on *other* agents’ strategies (via the equilibrium concept), and beliefs can again be assumed given when an agent is making her choice.

*Why Is It Nicer*? First, I find that the language of agents is appropriate whenever a dynamic game is concerned. The literature on changing preferences is almost unanimous in choosing the agent as the basic unit of analysis. (See Strotz, 1956; Peleg and Yaari, 1973; Hammond, 1976; and others.) Similarly, the debate regarding sequential rationality, common knowledge thereof, and backward induction also appears to call for splitting a player into agents in order to clarify what various assumptions mean. The agents language is rich enough to describe anything we may wish to say about players, by explicitly stating axioms about a player as a coalition of agents. (For instance, see Ferreira, Gilboa, and Maschler, 1992). On the other hand, treating all agents of a player as a single decision maker can be misleading, especially in the presence of potential changes in preference or revisions of belief.

An agent may have a non-trivial information set at which she is facing uncertainty about some ‘‘external’’ state space. As long as this set does not intersect any path more than once, it is possible to consider the agent’s decision problem, setting the clock to the time of that decision, and leaving any uncertainty to be reflected by a state space that is not under the agent’s control. By contrast, if an agent’s information set does contain two nodes on a path, part of the agent’s uncertainty is about her own decision. It is impossible to define the ‘‘time of decision,’’ since there are at least two candidates for the title, and none of them can correspond to both nodes. Furthermore, whichever node we choose to analyze, in the resulting decision problem the agent will have some control over the ‘‘states.’’

A related problem is that, in this case, the agent does not know who she is: the two ‘‘versions’’ of the agent cannot uniquely define themselves, nor can they tell when they are called to act even according to their subjective clocks. To clarify this point, note that an outside observer who considers the game tree in Fig. 1 can tell that d1 and d2 are two nodes that ‘‘objectively’’ can never occur at the same time. That is, an outside observer can clearly distinguish between ‘‘the agent of the driver facing d1’’ and ‘‘the agent of the driver facing d2.’’ But the game in Fig. 1 does not allow these two agents to have independent lives. They are lumped together in the same information set, unable to even conceive of choosing different strategies.

True, neither of these agents, when making a decision, can tell which agent she is in terms of the outside observer. But if she is smart enough to reason about the game tree, she should also be aware of the existence of two agents, and she should have some way of referring to herself as separate from her twin. (See Gilboa, 1992, for a related discussion.) Specifically, each agent can define herself at least at the time of action (which is her time of glory, the only time that she matters) by the statement, ‘‘I am the agent who is now playing. I, and no other, am right now being called upon to act.’’

As is readily seen, if we do allow the agents ‘‘knowledge of their selves,’’ no inconsistencies arise. The heart of the paradox is the cycle of belief-act-belief: in theories of ‘‘rational’’ decision, beliefs are supposedly determining acts. (Even when we deduce beliefs from observed acts, we typically believe that the causal relationship is reversed.) In some cases—as in the paradox at hand—we also get a causal relationship that leads from acts to beliefs. It is this second link which closes the cycle and yields a paradox. It is precisely this link which is severed in the game of Fig. 2.

Severing the act−belief causal link is not an ad-hoc solution to the driver’s paradox. Rather, it follows from the teachings of classical decision theory: beliefs are defined over states of the world, which are not for the decision maker to choose from, and the beliefs themselves are not under her control either. ‘‘Choosing what to believe’’ is considered irrational in the popular sense of ‘‘rational,’’ and it also stands in contradiction to the classical theory. Beliefs over acts, as well as beliefs induced by acts, are notorious for generating counter intuitive and paradoxical results, as in Newcomb’s paradox and the ‘‘twins paradox.’’ (See Gibbard and Harper, 1978 and Gilboa, 1993, where it is argued that beliefs over acts preclude a satisfactory model of ‘‘free choice.’’)

*The Twins Paradox.* Indeed, the absent-minded paradox is very close in nature to the ‘‘twins paradox.’’ In the latter, two identical players are playing a one-shot prisoner’s dilemma game, and it is argued that, since they are bound to eventually choose the same action, they should cooperate. While there is no denial that if there is a *causal* link between their actions cooperation will result, most game theorists would reject this argument for cooperation. A common view is that, whereas the players will end up choosing the same action, when they reason about it they can *conceive* of situations in which they choose differently. Specifically, when a player contemplates a deviation from the (dominating-strategies) Nash equilibrium, she holds her beliefs about her opponent’s behavior fixed.

The similarity between the paradoxes is twofold: first, in both of them paradoxical results seem to follow from allowing beliefs to depend on actions. In the twins paradox, a player’s belief about her opponent is determined by her currently contemplated move. In the driver paradox, an agent's belief regarding the decision node she is at is determined by her choice of mixed strategy. Second, in both paradoxes the rules of the game do not allow the players to even conceive of the possibility of choosing different acts. In the twins paradox, whatever is a candidate for a decision, it is evaluated based on the assumption that it is the choice of both players. Similarly, lumping the two agents of the absent-minded driver in a single information set makes the game too restrictive to even describe different choices by the agents.

Should we have decision-theoretic foundations of game theory, the game structure should reflect the decision problem as perceived by each of the decision makers, namely, each of the agents involved. It is therefore essential that the game tree not allow beliefs to depend on acts, just as classical decision theory does not allow a decision maker's choice to alter her beliefs over states of the world. By a similar token, the game tree should not presuppose that several agents—or players—are identical in their choices. Identical choices may result from behavioral assumptions on individual decision makers. But if we wish to describe the agents' reasoning process, the model should be able to describe different choices as well.

*What Will Happen*? As mentioned above, the game in Fig. 2 has a symmetric Nash equilibrium. Furthermore, the (ex-ante) optimal strategy in the game in Fig. 1, when applied to both agents, is an equilibrium strategy.[1] The analysis of the game in Fig. 2 with symmetric Nash equilibria is basically identical to the analysis of the game in Fig. 1 with ‘‘modified multi-self consistent’’ strategies, as defined in Piccione and Rubinstein (in the revised version, this issue). They also show that optimal strategies are modified multi-self consistent.

This, however, is not the main point. My argument against the game in Fig. 1 is based on ideological grounds, and should not be viewed as a desperate attempt to retrieve optimality.[2]

In general, when we model a game with separate agents (as the one in Fig. 2), we may have asymmetric Nash equilibria as well. This should pose no difficulty. Since we have a compelling theoretical reason to focus on symmetric equilibria, we should simply do so. Still, the game itself should be rich enough to describe all eventualities that some agents may conceive of when they reason their way to a ‘‘rational’’ choice.

[1] I thank an anonymous referee for pointing out this fact to me. The earlier version of this comment contained an erroneous claim to the contrary. The referee also drew my attention to ‘‘modified multi-self consistent’’ strategies.

[2] While the earlier version of this comment reflects poor algebra, it may serve to prove my ideological commitment.

*Agents*? One may object to the concept of ‘‘agents.’’ In particular, some argue that nothing is known about ‘‘agents,’’ that their choices can never be observed a priori, and that the whole notion is somewhat metaphysical, where the latter epithet is not intended as a compliment. Without delving into this question here I argue that sometimes the best way to clear up our intuition may involve terms which can only be ‘‘observed’’ by introspection. (See Peleg and Yaari, 1973 and Ferreira *et al*., 1992, for further discussions.)

## REFERENCES

Ferreira, J.-L., Gilboa, I., and Maschler, M., (1992). ‘‘Credible Equilibria in Games with Changing Utility,’’ *Games and Econ*. *Behav*., in press.

Gibbard, A., and Harper, W. L., (1978). ‘‘Counterfactuals and Two Kinds of Expected Utility,’’ *Foundations and Appl*. *Decision Theory* **1**, 125−162.

Gilboa, I. (1992). ‘‘Why the Empty Shells Were Not Fired: A Semi-bibliographical Note,’’ discussion paper, Northwestern University.

Gilboa, I. (1993). ‘‘Can Free Choice Be Known?,’’ forthcoming in *The Logic of Decision*, C. Bicchieri, R. Jeffrey, and B. Skyrms (Eds.) Oxford Univ. Press.

Hammond, P. J. (1976). ‘‘Changing Tastes and Coherent Dynamical Choice,’’ *Rev*. *Econ*. *Studies* **43**, 159−173.

Peleg, B., and Yaari, M. E., (1973), ‘‘On the Existence of a Consistent Course of Action when Tastes Are Changing,’’ *Rev*. *Econ*. *Studies* **40**, 391−401.

Piccione, M., and Rubinstein, A., (1997). ‘‘On the Interpretation of Decision Problems with Imperfect Recall,’’ *Games and Econ*. *Behav*., **20**, 3−24.

Strotz, R. H. (1956). ‘‘Myopia and Inconsistency in Dynamic Utility Maximization,’’ *Rev*. *Econ*. *Studies* **23**, 165−180.