Reprinted from JOURNAL OF ECONOMIC THEORY
All Rights Reserved by Academic Press, New York and London
Vol. 16, No. 2, December 1977
Printed in Belgium

Independence of Irrelevant Alternatives, and Solutions to Nash's Bargaining Problem

ALVIN E. ROTH

University of Illinois, Urbana, Illinois 61801

Received June 24, 1977

Several variants of the axiom of independence of irrelevant alternatives are considered, in the context of Nash's bargaining problem.

In this paper we will consider a class of solutions to the classical two person bargaining problem as formulated by Nash [4]. A bargaining problem consists of a compact convex subset S of the plane, and a point $d = (d_1, d_2)$ on the plane such that there exists at least one point s in S with $d < s$.

The interpretation usually given is that the plane represents the set of von Neumann–Morgenstern utility payments to the two players, the set S represents the set of feasible utility payoffs resulting from some agreement, and the point d represents the utility levels achieved by the two players if no agreement is reached. The point d is called the "disagreement point" or "status quo," and the condition that $(d_1, d_2) < (s_1, s_2)$ for some s in S simply says that there is at least one feasible agreement which gives both players an incentive to bargain.

A *solution* to the bargaining problem is a function which selects a feasible outcome for every bargaining problem. That is, if B is the set of all bargaining problems, a solution[1] is a function $f: B \to R^2$ such that, for every (S, d) in B, $f(S, d) \in S$. Such a function is often interpreted either as an arbitration procedure or as a model of the bargaining process.[2]

Nash proposed that a solution f should obey the following conditions.[3]

1. *Independence of positive linear transformations.* For any bargaining problem (S, d) and real numbers a, b, c, and e such that $a, b > 0$, let $T = \{(ax_1 + c, bx_2 + e) | (x_1, x_2) \in S\}$, and let $d' = (ad_1 + c, bd_2 + e)$. Then $f(T, d') = (af_1(S, d) + c, bf_2(S, d) + e)$.

[1] When no confusion will result, we sometimes refer to the point $f(S, d)$ as the solution to the bargaining problem (S, d).

[2] It has recently been shown [7] that under appropriate conditions such functions can also be interpreted as utility functions reflecting the expected reward of bargaining.

[3] See Luce and Raiffa [3] for a more complete discussion of the conditions.

ISSN 0022-0531

This condition requires that changes in the origin and scale of the utility functions of either player do not have any essential effect on the solution. This is a necessary requirement, since the representation of a von Neumann–Morgenstern utility function is defined only up to the arbitrary choice of origin and scale.

2. *Symmetry.* If (S, d) is a symmetric bargaining problem (i.e., if $(x_1, x_2) \in S$ implies $(x_2, x_1) \in S$ and if $d_1 = d_2$) then $f_1(S, d) = f_2(S, d)$.

This condition says that the labels of the players do not matter: if switching the labels of the players leaves the bargaining problem unchanged, then it should leave the solution unchanged.

3. *Pareto optimality.* The point $x = f(S, d)$ is Pareto optimal in S. That is, $x \in P(S) \equiv \{x \in S \mid y \in S \text{ and } y \neq x \text{ implies } y \not\geqslant x\}$.

This condition says that the solution f picks a point with the property that it is impossible to increase the payoff of one player without decreasing the payoff of the other player.[4]

4. *Independence of alternatives other than the disagreement point.* If (S, d) and (T, d') are bargaining problems such that $d = d'$ and T contains S, and if $f(T, d') \in S$, then $f(T, d') = f(S, d)$.

Condition 4, which Nash called "independence of irrelevant alternatives," says that the solution of a bargaining problem does not change as the set of feasible outcomes is reduced, so long as the disagreement point remains unchanged, and the point originally selected remains feasible. The condition says that the selection of a feasible point in an outcome set does not depend on any point except possibly the disagreement point.

Nash showed that there is a unique solution f obeying conditions 1–4. It is the function N which picks the point $N(S, d) = x$ such that $x > d$ and $(x_1 - d_1)(x_2 - d_2) \geqslant (y_1 - d_1)(y_2 - d_2)$ for all $y \in S$ such that $y > d$. Thus the Nash solution of a bargaining problem picks the point which maximizes the geometric average of the gains which the players receive from bargaining (as opposed to settling for their disagreement utilities). In view of condition 4, it is not surprising to note that the Nash solution is dependent on the disagreement point.

Another solution to the bargaining problem, first proposed by Raiffa [5][5] depends on both the disagreement point d and the *ideal point* $i(S)$ defined by $i_1(S) = \max\{x_1 \mid (x_1, x_2) \in S\}$, $i_2(S) = \max\{x_2 \mid (x_1, x_2) \in S\}$. Note that if $i(S)$ is feasible then it is the unique Pareto optimal point in S. The solution Raiffa proposed is the function R which picks the point $R(S, d) = x$ such that x is

[4] It has recently been shown [6] that condition 3 can be replaced by the requirement of strict individual rationality, i.e., the requirement that $f(S, d) > d$, without changing Nash's result.

[5] Also see Luce and Raiffa [3].

Pareto optimal in S, and lies on the line joining d to $i(S)$. Note that the function R obeys conditions 1–3, but does not obey condition 4, since it depends on the point $i(S)$. It does however, obey the following version of independence of irrelevant alternatives.

5. *Independence of alternatives other than the disagreement point and the ideal point.* If (S, d) and (T, d') are bargaining problems such that $d = d'$ and $i(S) = i(T)$, and T contains S, and if $f(T, d') \in S$, then $f(T, d') = f(S, d)$.

Of course the Nash solution also obeys condition 5, since it obeys condition 4. The function R has recently been characterized in an elegant way by Kalai and Smorodinsky [2] who show that R is the unique solution which obeys conditions 1–3 and also obeys a certain monotonicity requirement.

The relationship between conditions 4 and 5 suggests the consideration of solutions which depend only on the ideal point, i.e., solutions which obey the following version of independence of irrelevant alternatives.

6. *Independence of alternatives other than the ideal point.* If (S, d) and (T, d') are bargaining problems such that $i(S) = i(T)$, and T contains S, and if $f(T, d') \in S$, then $f(T, d') = f(S, d)$.

Two recent papers by Yu [8] and Freimer and Yu [1] consider a class of solutions which obey condition 6. For p greater than 1 they consider the functions Y_p such that[6] $Y_p(S, d) = x$ is the individually rational point in S which minimizes the 1_p-norm distance to the ideal point.

It is straightforward to verify that the functions Y_p satisfy conditions 2, 3, and 6, but they do *not* satisfy condition 1. That is, the agreement picked by the functions Y_p is dependent on the origins and scales chosen for the utility functions of the two players. It is therefore natural to ask whether *any* solution exists which satisfies conditions 1, 2, 3, and 6. The following theorem answers that question in the negative.

THEOREM 1. *There is no solution which satisfies conditions* 1, 2, 3, *and* 6.

Proof. Suppose that f is such a solution. Let T be the convex hull of the points $t1 = (0, -2)$, $t2 = (-2, 0)$, $t3 = (-15, 0)$, $t4 = (-15, -15)$, $t5 = (0, -15)$, and let $d' = t4$. Then (T, d') is a symmetric bargaining problem, whose Pareto set is the line segment joining $t1$ and $t2$. So, by conditions 2 and 3, $f(T, d') = (-1, -1)$.

Let S be the convex hull of the points $(0, -6)$, $(-\frac{1}{2}, -\frac{3}{2})$, $(-1, -1)$, $(-6, 0)$, and $(-6, -6)$. Then T contains S and $f(T, d') \in S$. Also, $i(S) = i(T) = (0, 0)$, so by condition 6, $f(S, d) = f(T, d') = (-1, -1)$ for any disagreement point d.

[6] Note that we continue to denote a bargaining problem by (S, d), even though any solution is now required to be independent of d. The problem would not be well defined if we failed to specify the outcome in the event that no agreement is reached.

Now let $S' = \{(2x_1, \frac{2}{3}x_2) | (x_1, x_2) \in S\}$, and let $d'' = (2d_1, \frac{2}{3}d_2)$. Then condition 1 implies $f(S, d'') = (-2, -\frac{2}{3})$. But T contains S', which is the convex hull of the points $(0, -4)$, $(-1, -1)$, $(-2, -\frac{2}{3})$, $(-12, 0)$, and $(-12, -4)$. Furthermore, $i(S') = i(T) = (0, 0)$, and $f(T, d') \in S'$. So condition 6 implies that $f(S', d'') = f(T, d') = (-1, -1)$. This supplies the contradiction needed to complete the proof.

Thus any solution which is Pareto optimal, symmetric, and independent of outcomes other than the ideal point, must be dependent on the (arbitrary) origin and scale chosen to represent each player's utility function.

Now the ideal point of a set S is determined by the boundary points of the Pareto optimal set $P(S)$. In particular, let y and z be the Pareto optimal points[7] most preferred by players 1 and 2, respectively. That is, y and z are the points in $P(S)$ such that if x is in $P(S)$ and $x \neq y$, then $y_1 > x_1$ (and $y_2 < x_2$), and if $x \neq z$ then $z_2 > x_2$ (and $z_1 < x_1$). So y and z determine the ideal point, since $i(S) = (y_1, z_2)$.

In addition to determining the ideal point, the points y and z also determine the point $m(S)$ of *minimal expectations*, defined by $m_1(S) = \min\{x_1 \mid (x_1, x_2) \in P(S)\}$, and $m_2(S) = \min\{x_2 \mid (x_1, x_2) \in P(S)\}$. The payoff $m_k(S)$ is thus the minimum payoff that player k can achieve at a Pareto optimal outcome, and $m(S) = (z_1, y_2)$.

Motivated by condition 6, and by the fact that, like the ideal point, the point of minimal expectations is determined by the boundaries of the Pareto optimal set, we consider the following form of independence of irrelevant alternatives.

7. *Independence of alternatives other than the point of minimal expectations.* If (S, d) and (T, d') are bargaining problems such that $m(S) = m(T)$, and T contains S, and if $f(T, d') \in S$, then $f(T, d') = f(S, d)$.

In view of Theorem 1, we might expect that no solution obeys conditions 1, 2, 3, and 7, but this is not the case. In fact, we have the following result.

THEOREM 2. *There is a unique solution which obeys conditions* 1, 2, 3, *and* 7. *It is the function M defined for any bargaining problem* (S, d) *by*

$$M(S, d) = \begin{cases} m(S) & \text{if } m(S) \in P(S) \\ N(S, m(S)) & \text{otherwise,} \end{cases}$$

where N is the Nash solution.

Proof. First observe that M obeys conditions 1, 2, 3, and 7. To see this, observe that when $m(S) \in P(S)$, M obeys condition 3 by definition, and

[7] Of course y and z depend on S, but to keep the notation simple we have suppressed this.

conditions 1 and 2 since $m(S)$ does. When $m(S) \in P(S)$ it is the unique Pareto optimal point, and so condition 7 is also satisfied in this case.

When $m(S) \notin P(S)$, then $m(S)$ can be interpreted as a disagreement point (since there is an $s \in S$ such that $m(S) < s$). Thus $N(S, m(S))$ is well defined, and obeys conditions 1, 2, and 3, since N does. Also, N obeys condition 4 which is stated for arbitrary disagreement points, so it obeys condition 7 which is the special case $d = m(S)$ and $d' = m(T)$.

To see that M is the *unique* function satisfying conditions 1, 2, 3, and 7, first observe that, when $m(S) \in P(S)$, any solution obeying condition 3 must pick the point $m(S)$. When $m(S) \notin P(S)$, M coincides with N, which is the unique function defined on the set of all bargaining games which obeys conditions 1–4. If we show that N is the unique function which satisfies 1–4 on the subclass of bargaining problems $B' = \{(S, d) \in B \mid d = m(S)\}$ then we will have shown that M is the unique solution satisfying conditions 1, 2, 3, and 7.

Let $(S, m(S))$ be a bargaining problem normalized so that $m(S) = (0, 0)$ and $N(S, m(S)) = (1, 1)$. Nash observed that if $x \in S$, then $x_1 + x_2 \leqslant 2$. Consequently we can construct a symmetric set T containing S such that $(1, 1) \in P(T)$. Furthermore, since $m(S) = (0, 0)$, T can be constructed so that $P(T)$ is equal to the line segment joining the points $(0, 2)$ and $(2, 0)$, i.e., T can be constructed so that $m(T) = m(S)$.

So if f is any solution obeying conditions 2 and 3, then $f(T, m(T)) = (1, 1) = M(T, m(T))$. If f also obeys condition 7, then $f(S, m(S)) = f(T, m(T)) = M(S, m(S))$. Finally, if f obeys condition 1, then for any problem $(S', m(S'))$, f and M coincide, which completes the proof.

REFERENCES

1. M. FREIMER AND P. L. YU, Some new results on compromise solutions for group decision problems, *Manage. Sci.* **22** (1976).
2. E. KALAI AND M. SMORODINSKY, Other solutions to Nash's bargaining problem, *Econometrica* **43** (1975).
3. R. D. LUCE AND H. RAIFFA, "Games and Decisions," Wiley, New York, 1957.
4. J. F. NASH, The bargaining problem, *Econometrica* **18** (1950).
5. H. RAIFFA, Arbitration schemes for generalized two-person games, *in* "Annals of Mathematics Studies," (Kuhn and Tucker, Eds.), Vol. 28, Princeton Univ. Press, Princeton, N.J., 1953.
6. A. E. ROTH, Individual rationality and Nash's solution to the bargaining problem, *Math. Oper. Res.* **2** (1977).
7. A. E. ROTH, The Nash Solution and the Utility of Bargaining, *Econometrica*, in press.
8. P. L. YU, A class of solutions for group decision problems, *Manage. Sci.* **19** (1973).

Printed by the St Catherine Press Ltd., Tempelhof 37, Bruges, Belgium.