A Direct Proof of Arrow's Theorem

Author(s): Julian H. Blau

Source: *Econometrica*, Jan., 1972, Vol. 40, No. 1 (Jan., 1972), pp. 61-67

Published by: The Econometric Society

Stable URL: https://www.jstor.org/stable/1909721

Econometrica, Vol. 40, No. 1 (January, 1972)

A DIRECT PROOF OF ARROW'S THEOREM

BY JULIAN H. BLAU

1. INTRODUCTION

IN HIS WELL-KNOWN book [**1**], Arrow created a mathematical structure embodying the concept of social welfare function, and then translated widely held beliefs and normative judgements into precise mathematical statements. The book's central thesis, known as Arrow's Theorem, was that these beliefs and judgements were inconsistent.

The statement and proof have undergone a number of improvements, the latest [**3**] being by far the best. Yet it too is encumbered by unnecessary structure, for which I accept a share of the responsibility [**4**]. Because of the great interest in the theorem itself, and in the problem created by the theorem, a new proof may be useful. To be in fact useful, it should be direct, and contribute to the clarification of issues. I attempt such a proof here. The principal novelty lies in the treatment of neutrality.

In Section 4 Arrow's theorem is proved under the strengthened hypothesis that the number of alternatives is at least five, which is true of all significant economic models. In Section 5 the proof of Arrow's theorem itself is completed, and the theorem is extended slightly in Section 6.

I assume that the reader has some familiarity with the subject, permitting the concise review in the next two sections. For background and details, see [**3**, **4**, and **6**].

2. REVIEW OF DEFINITIONS

(i) A is the set of *alternatives.*

(ii) A relation R on A is a *weak ordering* of A if it is reflexive, transitive and connected (complete).

(iii) If R is a weak ordering of A, then xPy means that yRx is false, while xIy means that xRy and yRx are true. (Interpretations: P is preference, I is indifference, R is preference or indifference.) Evidently xPy, $yRz \Rightarrow xPz$; also xRy, $yPz \Rightarrow xPz$ [**1**].

(iv) $\mathscr{R}$ denotes the set of all weak orderings of A [**6**].

(v) N is the set of *people.*

(vi) A *profile* p is a function mapping N into $\mathscr{R}$. Thus, for each i in N, p assigns a weak ordering R_i [**6**].

(vii) A *social welfare function* (SWF) is a mapping of a set $\mathscr{D}$ of profiles into $\mathscr{R}$.

(viii) *Notation and terminology.* When convenient, denote a SWF by F, so that $F(p) \in \mathscr{R}$. $F(p)$ is called the social ordering corresponding to the profile p of individual orderings. Sometimes it is more convenient to write, for example, $p \rightarrow aRb$ to denote $aF(p)b$. Several hypotheses to follow, mainly independence of irrelevant alternatives (IIA), make it possible to infer social information from *partial* information about p. (ix) would be an extreme example of this.

(ix) Let $i \in N$. Then i is a *dictator* if, for each profile, and each ordered pair (x, y), $xP_iy \rightarrow xPy$.

3. CONDITIONS THAT A SOCIAL WELFARE FUNCTION MAY SATISFY

Universal domain: $\mathscr{D}$ is the set of *all* profiles.

(UPP) *Unanimity for P implies social P*: For each profile, and for each ordered pair (x, y), xP_iy for each $i \in N$ implies xPy (Arrow's Condition P [3]).

(UPR) *Unanimity for P implies social R.*

The notational scheme introduced here is useful for expressing various types of unanimity succinctly. UPP is what I called the Unanimity Rule for Preference (URP) in [4]. (There is little chance of confusion, for in the new notation, URP is self-contradictory, since it implies UIP.)

(IIA) *Independence of Irrelevant Alternatives*: For each subset B of A, and each pair p, p' of profiles, $p = p'$ on B implies $F(p) = F(p')$ on B (equivalent to Arrow's Condition 3 [1]).

A typical application of the powerful condition IIA: Let $\{E, F, G\}$ be a partition of N. Let all the members of E prefer x to y, those in F prefer y to x, and those in G be indifferent. This partial information (which may be loosely called a profile on $\{x, y\}$) suffices to determine which of xPy, xIy, yPx is the social result, since any two ways of expanding it to a full profile are subject to IIA with $B = \{x, y\}$. This being so, one may freely insert "irrelevant" alternatives without altering that result. Indeed the purpose of such insertion may be to determine, from other information, what in fact that result is.

Nondictatorship: There is no dictator (Arrow's Condition 5 [1]).

Null: A SWF is *null* if for each profile p, $F(p)$ is the universal relation, i.e., universal social indifference.

Neutral (more accurately, neutral for profiles without indifference, for a SWF satisfying IIA): Let E be a set of people, E' its complement, and $x \neq y$. If p is a profile in which xP_iy for all $i \in E$, and yP_ix for all $i \in E'$, and $p \rightarrow xPy$, we say that E is *decisive* for (x, y) [2]. The SWF is *neutral* if each set E is decisive for *all* ordered pairs of distinct alternatives, or for *none* [4].

This is neutrality in the standard sense that the SWF does not discriminate among alternatives. It is special in that it applies only to functions satisfying IIA. It is weak in that it asserts nothing concerning contests in which some individual indifference occurs. The term has been used by several authors in ways somewhat different from this and from each other.

ARROW'S THEOREM: *If* $|A| \geqslant 3$ *and* N *is finite, then universal domain,* IIA, UPP, *and nondictatorship are inconsistent* [3].

The content of Arrow's theorem has varied, but always within the form above. Enlargement of the domain has been the essential variation. UPP replaced nonimposed (Condition 4 [1]) in [4], and replaced nonimposed and monotonicity (Condition 2 [1]) in [3].

I begin by weakening the unanimity rule from UPP to UPR. I do not argue that UPP is too strong an assumption, but rather investigate, at no cost in complexity, the extent to which UPR serves as well. The null function shows that UPR, unlike UPP, is consistent with the remaining conditions. The question is whether there are any other examples.

Henceforth I assume universal domain, IIA, *and* UPR.

4. PROOF OF ARROW'S THEOREM FOR FIVE OR MORE ALTERNATIVES

Let E be a set of people, fixed until further notice. We write xDy to mean that E is decisive for (x, y), i.e., D is a relation on A defined as follows: (i) D is irreflexive; (ii) xDy means

$$\begin{array}{cc} E & E' \\ x & y \\ y & x \end{array} \quad \rightarrow \quad xPy.$$

PROPOSITION 1: *If* a, b, x, y *are distinct, then* $aDb \Rightarrow xDy$.

PROOF: The independence of irrelevant alternatives is applied repeatedly, without comment. Using UPR, aDb, and UPR, in that order, the indicated social preferences are obtained from the profile below.

$$\begin{array}{cc} E & E' \\ x & b \\ a & y \\ b & x \\ y & a \end{array} \quad \rightarrow \quad xRa, aPb, bRy \Rightarrow xPy.$$

REMARK: The proof above brings out with unusual clarity the nature of the independence condition IIA and its role in the theory. The hypothesis aDb is used in what may be considered, perhaps naively, a strong form, and the conclusion xDy obtained in a weak form. These terms would have meaning only if there were interpersonal comparison of utility, which IIA was designed to exclude.

PROPOSITION 2: *If* $|A| \geqslant 5$, *and* $x \neq y$, *then* $aDb \Rightarrow xDy$.

PROOF: If a, b, x, y are distinct, then Proposition 1 applies. If they are not distinct, then there are at most three among them. In that case there are alternatives s, t distinct from a, b, x, y and each other. Then $aDb \Rightarrow sDt \Rightarrow xDy$ by Proposition 1.

We now know, for each set E of people, that whether E prevails against its complement E' depends solely upon E, and not upon the particular ordered pair of alternatives at issue. This property of the SWF is what we have called *neutrality* (more accurately, neutrality for profiles without indifference, for a SWF satisfying IIA). If E does prevail, we call it a *winning* set. The class of all winning sets is denoted by $\mathscr{W}$, borrowed from game theory. The sets which do not win constitute the class $\mathscr{L}$ of *losing* sets. Evidently a losing set is never decisive, for if it were so once, it would be so always, by Proposition 2, and would therefore be a winning set. In this context UPR may be expressed thus: the empty set is in $\mathscr{L}$.

For future reference, we note that the remainder of the proof is valid under the weaker hypothesis $|A| \geqslant 3$.

One is tempted to assert that the nondictatorship condition implies that each singleton is in $\mathscr{L}$, but this would be premature. For $\{i\}$ to be in $\mathscr{W}$, i must prevail against unanimous opposition, while to be a dictator he must in addition prevail when he receives some support and/or indifference. To infer the second from the first requires some degree of monotonicity. Arrow [**3**] was able to deduce this monotonicity from the other conditions, including UPP. He wove this into the neutrality proof, but I shall isolate it for emphasis, and state it in slightly more general form. The proof is Arrow's.

PROPOSITION 3: *If* $E \in \mathscr{W}$, *and if* aP_ib *for each* $i \in E$, *then* aPb.

PROOF: Let p be a profile on $\{a, b\}$ in which all members of E prefer a to b. Among people in E', all three opinions concerning $\{a, b\}$ may occur. Since $|A| \geqslant 3$, there is a third alternative t. Insert t between a and b for persons in E, and above a and b for persons in E'. Using $E \in \mathscr{W}$ and UPR, we have

$$\begin{array}{cc} E & E' \\ a & t \\ t & \\ b & \end{array} \qquad \rightarrow \qquad aPt, tRb \Rightarrow aPb.$$

As an immediate, in fact our only, application of Proposition 3, we have the following. Let $i \in N$. If $\{i\} \in \mathscr{W}$, then i is a dictator. Thus, if i is not a dictator, then $\{i\} \in \mathscr{L}$. To exploit this, we need

PROPOSITION 4: $\mathscr{L}$ *is disjunctively additive, i.e., the union of any finite number of pairwise disjoint losing sets is a losing set.*

PROOF: Let E and F be disjoint losing sets. Denote by G the complement of their union. Consider

$$\begin{array}{ccc} E & F & G \\ x & z & y \\ y & x & z \\ z & y & x \end{array}$$

In the social ordering resulting from this profile on $\{x, y, z\}$, zRx because $E \in \mathscr{L}$, and yRz because $F \in \mathscr{L}$. Hence yRx socially. This implies $E \cup F \in \mathscr{L}$.

Using mathematical induction, we obtain the stated result easily. (Note that the proof did not require UPR.)

We now bring in the nondictatorship condition and the finiteness of N, to be used once. The first implies that each singleton is a member of $\mathscr{L}$, and the second, with Proposition 4, implies that $N \in \mathscr{L}$. This contradicts UPP, proving the statement below.

THEOREM (*Arrow's Theorem for Five or More Alternatives*): *If* $|A| \geqslant 5$ *and* N *is finite, then universal domain,* IIA, UPP, *and nondictatorship are inconsistent.*

In Section 6, we resume the chain of reasoning at this point, to obtain a slight extension of Arrow's theorem.

5. PROOF OF NEUTRALITY FOR THREE OR MORE ALTERNATIVES

As in the proof for $|A| \geqslant 5$, let E be a fixed set of people. First we show that $aDb \Rightarrow aDx$ for all $x \neq a$. The method of proof is entirely due to Arrow. If $x = b$, then aDx is true by hypothesis. If $x \neq b$,

E	E'		
a	b		
b	x	$\rightarrow$	$aPb, bRx \Rightarrow aPx.$
x	a		

Similarly, $aDb \Rightarrow xDb$ for all $x \neq b$. The remainder of the proof will be made to depend upon a simple lemma concerning relations, with no further reference to SWF's. (This was the case also for $|A| \geqslant 5$, where Proposition 2 was the corresponding lemma.)

PROPOSITION 5: Let A be a set with at least three members, and D an irreflexive relation on A. Let D have the property that aDb implies aDx and xDb except where irreflexivity forbids. Then aDb implies xDy for all ordered pairs (x, y) with $x \neq y$.

PROOF: The proof is very simple when the geometry of the "plane" $A \times A$ is visualized. The hypothesis asserts that if the set D contains a point, then it contains also the entire horizontal line through that point, and the entire vertical line through that point, with the exception of points on the diagonal line $x = y$.

Let $x \neq y$. If $a \neq y$, then $aDb \Rightarrow aDy \Rightarrow xDy$. If $b \neq x$, then $aDb \Rightarrow xDb \Rightarrow xDy$. This leaves only the case $(x, y) = (b, a)$. Here we use the fact that A has a third member t distinct from a and b. $aDb \Rightarrow aDt \Rightarrow bDt \Rightarrow bDa$.

This completes the proof of Arrow's theorem.

6. ANOTHER FORM OF ARROW'S THEOREM

We continue the sequence of Section 4, again relying on UPR instead of the stronger UPP. From the fact that N is a losing set, we may prove that the SWF is null.

PROPOSITION 6: *If $N \in \mathscr{L}$, then F is null.*

PROOF: Since both N and its complement are losing sets, the latter by UPR, N is what game theorists call a blocking set. In our terms, a unanimous vote for aPb produces social indifference. More succinctly, UPI holds. There are various paths from this to our conclusion, but the simplest is due to Hansson [**5**], which we adopt. (The assumptions made by Hansson in arriving at UPI, while quite appropriate to his purpose, are extremely strong from the point of view of Arrow's theorem.) Given any profile on $\{a, b\}$,

$$\begin{array}{ccc} E & F & G \\ a & b & \\ & & ab, \\ b & a & \end{array} \qquad \text{we have} \qquad \begin{array}{ccc} E & F & G \\ a & b & \\ b & a & \overset{ab}{t} \\ t & t & \end{array} \qquad \rightarrow \qquad aIt, tIb \Rightarrow aIb.$$

THEOREM: *If $|A| \geqslant 3$ and N is finite, then universal domain, IIA, and UPR imply that the function is dictatorial or null.*

This result can be stated also as an inconsistency theorem by replacing UPP in Arrow's theorem by UPR and non-null. Easily constructed examples show that the replacement is weaker than UPP. Thus this is in fact a slight strengthening of Arrow's theorem.

7. REMARKS ON THE DEVELOPMENT OF THE PROOF OF ARROW'S THEOREM

1. One path to Arrow's result emphasizes the individual i. If i dictates one decision against the opposition of all others (aD_ib), then he does the same for all decisions (xD_iy). Any proof of this can be converted at once into a proof of the same assertion for any *set* E of individuals (aD_Eb implies xD_Ey). I call the latter property neutrality [**4**]. Clearly neutrality contains the result for individuals as a special case. The technical problems of proving the two assertions are identical.

2. I take the view that it is desirable to separate distinct issues, and therefore I deal with neutrality and monotonicity separately (Propositions 2 and 3). Arrow has given an elegant simultaneous treatment of $aDb \Rightarrow a\bar{D}x$ etc. in [**3**]. The present route provides a slightly different perspective.

3. The new proofs of neutrality (Sections 4 and 5) adapt at once to Arrow's combined treatment. For example, $aDb \Rightarrow x\bar{D}y$ (Arrow's notation) because the conclusion of Proposition 1 follows equally well when members of E' have various opinions concerning $\{x, y\}$, but all prefer both to a and prefer b to both.

4. [**4**] had substantially the same hypothesis as [**3**], with monotonicity in addition. Arrow showed [**3**] that monotonicity was superfluous, a fact with both technical and substantive significance. (This innovation was already present in [**2**] for the case $|A| = 3$. The proof is difficult to understand. After knowing a proof, such as in [**3**], one can almost see how to correct the misprints and thereby reconstruct the proof, which I believe is indeed there. An unjustified assumption in this proof, corrected in [**3**], makes it valid only for $|N| = 2$.)

5. I have noticed, while writing this paper, that in [4] I appear to have stated that the hypothesis of [2] was the same as that of [1], and that consequently the theorem in [2] was false. In fact the hypothesis includes universal domain, and the theorem is true. In the original manuscript of [4], [2] appeared on a list of papers having the same hypothesis as [1]. It was removed in the revised version but unfortunately was printed by mischance.

6. In [3], Arrow makes a flat priority claim for [2], and a qualified one for [1] as against "any proof of which I am aware." The first is surely not meant to be taken literally. The two older proofs are valid only for $|A| = 3$. It is not that the possibility $|A| > 3$ was treated erroneously, but rather that it was not really treated at all. [4] and [3] required Arrow's result for the case $|A| = 3$ as an intermediate step, while the present paper proceeds directly to the general case. It is indisputable that Arrow created the entire subject. His remarkable proofs for $|A| = 3$ are important in themselves, and in providing a strategy for a general proof. I heartily endorse his claims in this spirit.

Antioch College

Manuscript received March, 1970; revision received July, 1970.

REFERENCES

[1] ARROW, K. J.: *Social Choice and Individual Values*. New York: Wiley, 1951.
[2] ———: "Le Principe de Rationalité dans les Décisions Collectives," *Économie Appliquée*, 5 (1952), 469–484.
[3] ———: *Social Choice and Individual Values*. 2nd ed. New York: Wiley, 1963.
[4] BLAU, J. H.: "The Existence of Social Welfare Functions," *Econometrica*, 25 (1957), 302–313.
[5] HANSSON, B.: "Group Preferences," *Econometrica*, 37 (1969), 50–54.
[6] LUCE, R. D., AND H. RAIFFA: *Games and Decisions*. New York: Wiley, 1957.