Individual and Social Welfare: A Bayesian Perspective*

David Pearce

March, 2021

Abstract

Ragnar Frisch, one of the greatest proponents of scientific methodology in economics, warned that one cannot unreflectively apply the methodology of natural science to economics. I will argue that the unreflective application of logical positivism to welfare economics in the mid-twentieth century did great harm to that discipline. It led Arrow to impose the independence of irrelevant alternatives on any ethical method of preference aggregation, which I join many others in considering an unfortunate idea. It further caused Arrow and most of the profession to adopt an unnecessary and unnatural interpretation of his Impossibility Theorem. Viewing that theorem as a disturbing paradox encouraged many in the profession to consider theoretical work on social welfare and social choice an unpromising dead end. Samuelson's positivist-inspired criterion for a meaningful concept has left us without a fully developed and widely accepted language for discussing utility information richer than that which is generated by standard revealed preference. The second half of the paper wades into that forbidden territory, examining different ways that measurability, intensity, cardinality and comparability of utility are used in the literature and searching for alternative ways of understanding them.

[*] I would like to thank numerous people for discussing utility and welfare with me, and for their comments on this work. I had many valuable conversations with academic friends at NYU, including Dilip Abreu, Sylvain Chassang, Laurent Mathevet, Yaw Nyarko, Misha Panov, Diego Perez, Debraj Ray, Mauricio Ribeiro, Ariel Rubinstein, Andy Schotter, Ennio Stacchetti and Basil Williams. Further afield, I would like to thank Orazio Attanasio, D. C. Cymbalista, Juan Dubra, Piotr Dworczak, Julian Jamison, Sam Jindani, Mark Machina, Pepe Montiel Olea, Dilip Mookherjee, Ben Polak, Burkhard Schipper, Ludwig Sinander, Hugo Sonnenschein and Peter Wakker. When I wrote the paper in 1995 upon which this one builds, Amartya Sen gave me several pages of generous suggestions. Gian Luca Carniglia has been a remarkable research assistant on the current paper, finding relevant veins of literature of which I was not even aware. Some of the positions I take herein are unconventional, and none of the persons above is responsible for my errors or eccentricities. Faruk Gul especially must be absolved of responsibility: although he is in print taking positions quite different from mine, he spent hours discussing much of this material with me and giving valuable help. Finally, special thanks to two early mentors who introduced me to some crucial ideas in welfare economics: John Burbidge, who gave me a directed reading course at McMaster in my undergraduate days, and the late Dan Usher, who supervised my master's research at Queen's University at Kingston.

1 Introduction

Of all the major contributions of Ragnar Frisch to economics, perhaps the most decisive was his influence on economic methodology: how we do economics. Attending public lectures by Albert Einstein on recent developments in physics, at the University of Oslo in 1919, Frisch was struck by the complementary roles of inspired theoretical reasoning and ingenious empirical testing of conjectures[1]. Frisch believed that economics would benefit from the development and practice of a scientific methodology. This he impressed on the profession by the example of his own applied theoretical work, by his lectures on economic methodology[2] and by the founding of institutions with that point of view. A critical figure in the creation of the Econometric Society[3] and of its journal *Econometrica*, Frisch served as the journal's Chief Editor for its first twenty-two years!

To his endorsement of a scientific approach to economic inquiry, Frisch attached the following caveat:

> "... the methods of natural science cannot unreflectively be copied for use in economics."[4]

In my view, this qualification has not received the attention it deserves. This lecture will focus on what I consider the unreflective, and harmful, application of ideas popular among philosophers of science in the 1920's, to individual welfare and social welfare.

The logical positivists, such as Schlick, Carnap, Reichenbach and Hempel, developed the idea that only statements verifiable through direct observation (perhaps augmented by deduction) are *meaningful*. While this seems to have influenced many economic writers in the 1930's and 1940's, it is given stark expression by Paul Samuelson.

> "It is clear that every assumption either places restrictions on our empirical data or is meaningless." (Samuelson, 1947, p. 171)

Here, Samuelson is speaking of consumer theory. But he makes it plain elsewhere that he believes nothing is added to welfare economics by going beyond ordinal information about individual tastes. See Samuelson (1938). Some light-hearted evidence suggests that he didn't revise this attitude. In an essay entitled "Utility and All That" (a reference to Yeatman and Sellar (1930)), Robertson (1951) argued that cardinal utility was a useful concept in economics. Eulogizing Robertson after his death, Samuelson (1963, p.518) praised his persuasive powers: "The man could almost make you believe in such absurd things as cardinal utility." It is clear what Samuelson thought of utility and all that.[5]

[1] See Bjerkholt (2008, pg. 10). There are many fine "history of thought" papers by Olav Bjerkholt, sometimes with coauthors, on the life and work of Ragnar Frisch. In the summer of 2019 I wrote to him asking for further details on that subject, and received help and encouragement. When I learned later that Olav Bjerkholt passed away in February 2020, I was sad to lose this new academic friend.

[2] See Frisch (1926a) and Bjerkholt and Qin (2010).

[3] No less a figure than Alfred Cowles (1960) pronounced that "Ragnar Frisch, more than any other one man, took the initiative in founding, and in establishing the broadly international character of, the Econometric Society and its journal, *Econometrica*".

[4] Frisch (1926a, pp. 302, 303)

[5] One could argue indefinitely about what is admissible "empirical data", in Samuelson's criterion above. Do we admit only actual choice experiments, or also experiments with hypothetical prizes? Are choice frequencies relevant? How about verbal reports, or neurological data? In their early work on social choice and welfare, Arrow and Samuelson both apparently had fairly severe standards in this regard, basically viewing a preference ordering as the most one could elicit from an individual (more on this in Section 2). It is in their severe sense that I will use the words meaningful and meaningless in

If we take Samuelson's criterion for meaningfulness literally, the question "Does this experimental subject have thoughts and feelings, or is it insentient and simply programmed to simulate human behavior?" must be considered meaningless.[6] This is an unpromising foundation on which to build welfare economics. To me, a satisfying welfare economics must involve the experiencing of human feelings. In contrast, a welfare economics that considers thoughts and feelings meaningless is like Seinfeld: a show about nothing.

Long ago, in a paper that had a stormy reception, I tried to argue this with greater solemnity:

> "... when one speaks of welfare economics and social choice, human perceptions and feelings are of the essence: without them, it is not clear what is being discussed. To trim experience from a model of social welfare in the name of Occam's razor is to kill the subject by cutting out its heart." (Pearce, 1995)

Section 2 suggests that traditional utility and welfare theory are based on assumptions that share a lot with solipsism. It argues, by example, that the utility information viewed by Arrow and Samuelson as being sufficient for doing welfare analysis is entirely inadequate. A sympathetic observer, if given the power to make resource allocations for a society, should value information beyond that specified by a preference profile.

What if the only reports received by the observer are individuals' respective preference orderings? Then we are in the domain of Arrow's Impossibility Theorem. An observer, or more generally a rule, must order social alternatives, using only information about the preference profile. Section 3 considers the implications of rationality on the part of a sympathetic observer in this situation. We will see that *rationality implies not that the observer conform to IIA (the independence of irrelevant alternatives), but that she violate it.* The Bayesian perspective combined with the power of Arrow's Theorem has implications about social choice in this modest informational environment that are entirely different from the way the profession has long interpreted Arrow's Theorem. The results here draw heavily on the core material from Pearce (1995).

Citations of the various editions of Arrow's *Social Choice and Individual Values* passed twenty thousand long ago. This paper is not a survey. It cannot even scratch the surface of this immense literature. But each Section presents some of the thinking of the leading participants in the literature. Section 4 takes advantage of the detective work of Fleurbaey and Mongin (2005) and Igersheim (2019) to understand how the popular interpretation of Arrow's Impossibility Theorem hurt the reputation of welfare economics, heeding the changing views of Arrow and Samuelson on that subject as well as the proponents of single-profile impossibility theorems. Section 3's reinterpretation of the Theorem casts those debates in a different light.

this paper. Of course one need not take a "binary" approach to these terms, nor to the term "revealed preferred": it all depends on what data is considered admissible, and what "bridging assumptions" are entertained to interpret that data. The profession has opened up a lot in these respects. In his Presidential Address to this World Congress, for example, Orazio Attanasio includes a discussion of cutting edge survey techniques for complementing "revealed preference" data with data on stated preferences and stated expectations.

[6] Reading the thoughtful account of decision theory by Gilboa (2009), one quickly encounters a discussion of free will. I was momentarily dismayed to see the statement (pg. 7): "... I know that I have free will, but whether anyone else has free will is a meaningless question." But I decided I didn't believe Tzachi thought it was meaningless in the everyday sense of the word; he was just using standard (and I would say, unfortunate) positivist terminology. Later I was delighted to see his lucid posting (Gilboa, 2014) on the Decision Theory Forum criticizing narrow positivism and what he calls the MRV: the "mere representation view".

Over many decades Amartya Sen has been a constant, if sometimes apologetic, champion of expanding the informational basis of social choice beyond what Arrow originally envisaged (see, for example, Sen (2011) and many of the references therein). While I argue in Section 3 that, even restricting attention to reports of ordinal information, the case for relaxing IIA can be pressed much more aggressively than Sen has done, Section 5 goes beyond ordinalism, trying to grapple with some of the foundational issues involved in the measurement of utilities. Although it has a strong point of view, it raises more questions than it answers. By way of conclusion, Section 6 mentions a few developments I would welcome in the economics literature concerning individual and social welfare. They culminate in hoping for the development of a language that is missing from economics due to the unreflective application of a "natural science" definition of "meaninglessness" to individual and social welfare.

2 Information about Individual Utility

How much can a person know about the well-being of others? For that matter, how much can he know about anything beyond his own mind? This solipsistic question was of concern to Descartes[7], who reassured himself as to his own existence by his famous declaration: "I think, therefore I am." While the Latin roots of "solipsism" suggest that Roman philosophers struggled with these ideas, the Greeks beat them to it. The early itinerant sophist Gorgias wrote a tract entitled "On Nonexistence". Fittingly, it does not exist, having been lost. But it was not lost before Roman times, and it particularly won the favor of a skeptic philosopher, appropriately named Sextus Empiricus. Thanks to scholars such as McComiskey (1997), it is possible to give a casual translation of Gorgias' three tenets of nonexistence as follows:

> First, nothing exists.
> Secondly, if it did, you wouldn't know.
> Finally, if you happened to know, you couldn't communicate it.

Join me in imagining Arrow and Samuelson as itinerant early philosophers. As the two of them arrive at an intellectually curious ancient Greek village, a crowd gathers. The two philosophers have been chatting between themselves about how one might elicit an individual's preference ordering, so they tell the crowd about choice and utility, as one does. The villagers are intrigued. One of them shouts: "Could you explain it again?" They do. Another villager calls out: "Tell us of the *further* utility information, beyond the preference ordering." A murmur runs through the crowd. "Yes, the *further* information. Tell us about *that*!" Arrow and Samuelson reply: Utility information beyond the preference ordering?

> First, it doesn't exist.
> Secondly, if it did, you couldn't access it.
> Finally, if you did, you would find it meaningless.

Is this precisely how they would have answered? I can't say. But see how Professor Arrow answers exactly that question, if in a different setting:

[7] Philosophers sometimes use the term "Cartesian anxiety", introduced in Bernstein (1983).

> "It is further assumed that utility is not measurable in any sense relevant to welfare economics, so that the tastes of an individual are completely described by a suitable preference pattern or indifference map." (Arrow, 1950, pg. 331)

Those of us trained in the neoclassical tradition have heard words to this effect countless times. Still, I can't make sense of them. For my own part, I would urgently want more information than just preference orderings when making social choice decisions. Perhaps an example will help.

I have been put in charge of making a social choice allocation. The people involved are five kindergarten children, and there are only two possible alternatives, x and y. The five kids have reported strict preferences as follows:

1	2	3	4	5
x	x	x	x	y
y	y	y	y	x

As the majority, even a substantial supermajority, strictly prefer x to y, so do I.

Now I am told, by the messenger who is relaying information: "This probably won't interest you, but the first four children prefer x, in which they each get to choose $1,001$ toys, as opposed to only $1,000$, under y. The fifth child has a fatal illness. Under x, she will die a long, terrifying death; under y, she will be treated and be fine. "

I will be falling over myself to change my ordering to y strictly preferred to x. This despite the fact that I did not learn that any child's strict preference had changed. I care about a lot more than the preference orderings. Ideally, I would like to understand what each alternative would be *like* for each of the five children. The preference orderings just give me a hint about that, leaving a tremendous amount unspecified. When the messenger gives me the extra information, it changes my guess (subjective prior) about what the alternatives would be like for each of the five year olds.[8]

Some readers may object that this example is misleading: it was supposed to give additional information about preferences, whereas in fact, it gives information about alternatives. I agree, although the *reason* the new information about alternatives matters to me, is that it changes my estimate of what the alternatives will be like for each child. My changed response reflects the fact that my guesses about the welfare of each child under the two alternatives have changed, even though my beliefs about their preference orderings have not. My priorities in social choice are not remotely consistent with Arrow's assumption stated earlier. Nonetheless, one can revise the example to focus on information about preferences rather than about alternatives. Modify the original example so that the messenger's extra information is: "When I asked the kids their preferences, the first four didn't seem very interested, and said x would be better, I guess — can we go back to playing now? But the fifth child said that although y would be ok, if x were to happen, she would be plunged into such a deep personal crisis that she would probably never emerge."

What would I say to myself then? First, I'd say that's quite a mouthful for a five year old. Secondly, I would feel quite unsure that x is worth it, even though it is relatively popular. The choice

[8] Notice that even the extra information does not tell us everything we would like to know. For example, it might be the case that the fifth child is so deeply depressive that she scarcely prefers life to death. But probably NOT! The new information changes my guess about how different the two alternatives would be for the fifth child (let's not argue about what units we should use to quantify this, but accept that it would plausibly change my guess). Social choice, like most choice in life, is usually about guessing. We would like those guesses to be as intelligent as possible, and additional information will often be relevant.

of alternatives doesn't seem to be a big issue for the first four kids, and it seems to be devastatingly important for the fifth. This *proves* nothing, but as a Bayesian, I would be discouraged about alternative x and more inclined to rank y ahead of x. Again, for me, information that didn't change any of the five preference orderings is crucial for my social choice decision.

Returning for a moment to Samuelson's definition of meaningless, one might say that it is, after all, only a definition.[9] This is how we use the term. But no sooner have you called something meaningless, than you want to say we should not care about it. How can we care about a distinction that is meaningless? And so, after Samuelson's 1947 definition, we see Arrow in 1950 saying that for the purposes of social choice, nothing beyond a person's preference ordering is relevant. Things go downhill from there.

Nonetheless, the next two Sections will focus on situations in which preference orderings are the only information available (postponing considerations of richer informational settings to Section 5.) Traditional interpretations of Arrow's Theorem and the independence of irrelevant alternatives (IIA) will be challenged. Notice, though, that in the kindergarten example, there are only two alternatives, so IIA is satisfied vacuously. In other words, my problems with neoclassical social choice and welfare begin long before the imposition of IIA.

3 Arrow's World

If a social choice decision were entrusted to me, I would like to know what each social alternative (or allocation of resources) would be like for each person affected. How a particular person feels about, or experiences, each respective alternative, determines how he or she ranks them: the ranking is just a shadow cast by what the alternatives are like for the individual. It's the nature of the experiences that I care about, not the shadow. But the former is more elusive than the latter, which can be elicited through choice experiments. If the shadow is all the information I have, I value it for the hints it gives me about the nature of the experiences that different alternatives would generate.

Let us stipulate that in some cases of importance, preference profiles are essentially all the information that is available for the purposes of social choice (not all the information that is *of interest* for social choice, but all the information that is *available* in those cases). Then a rule that ranks allocations will have to be a function only of the preference orderings of individuals. This means we are in the world of Arrow (1950, 1951, 1963), even though, unlike Arrow (1950), we might have found further information about tastes important, had it been available. I quickly review the definitions and notation needed to state a simple version of the celebrated Impossibility Theorem.

Consider a set of N citizens, N finite, and a finite set of social allocations A, $|A| \geq 3$. An Arrovian social choice function (scf) maps preference profiles (over A) into a "social ranking" of A. Some think of this scf as a voting rule. This is appropriate only if citizens are viewed as revealing their preference orderings honestly: Arrow explicitly rules out any strategic behavior. If preferences are known, an scf is a way of aggregating them.

[9] Shakespeare's Juliet, in a flight of optimism, says:

> "What's in a name? That which we call a rose
> By any other name would smell as sweet..."

This convinced Romeo. But not all their beliefs were validated.

Arrow proposes four conditions that any acceptable scf should satisfy:

- $\mathcal{U}$ (Universal domain): The domain of the scf is the set of all logically possible preference profiles.
- $\mathcal{P}$ (Pareto principle): If all N individuals strictly prefer any x to any y, x is socially strictly preferred to y.
- $\mathcal{D}$ (Nondictatorship): There is no person i such that whenever i strictly prefers any x to any y, then x is strictly preferred to y, no matter what others prefer.
- IIA (Independence of Irrelevant Alternatives): The social ranking of any pair of alternatives x and y depends only on the individual rankings of x and y.

Theorem (Arrow's Impossibility Theorem[10])**.** *For N finite and $|A| \geq 3$, there is no scf satisfying $\mathcal{U}$, $\mathcal{P}$, $\mathcal{D}$ and IIA.*

One of the most famous results of the twentieth century, the Impossibility Theorem sent shock waves through economics, political science and philosophy. Often described as "disturbing" and "troubling", the theorem was considered a paradox and even a challenge to democracy.

Much of this was in line with Arrow's own views. Arrow (1950) reviews his axioms and concludes that, as one can't do without $\mathcal{P}$, $\mathcal{D}$ or IIA, it is necessary to explore restricted domains. An extensive literature resulted. See for example Gaertner (2002). Still, many questioned the wisdom of imposing IIA. Might ranking information beyond the pair of allocations x and y not hold clues to the *intensity* of an individual's preferences for x over y? But for years these questions never made it past the dual guardians of the positivist gate: *meaninglessness* and *immeasurability.*

Even Amartya Sen sounds apologetic as he tentatively raises the issue:

> "The rationale of positional rules relates to attaching importance to the placing of intermediate alternatives in individual preferences, which can be taken as suggesting that the gap between the two must be, other things given, larger. This argument is not entirely convincing. Many intermediate alternatives can be placed in a small interval, while large intervals may happen to be empty because of the contingent fact that there happens to be no other alternative that fits in just there. On the other hand, if information is thought to be extremely hard to get in social choice (a view that was certainly taken by Borda, 1781), then it is not entirely unreasonable to attach some significance to the fact that the placing of intermediate alternatives might be indicative of something". (Sen, 1987, pg. 389)

I claim that a decision-theoretic approach helps *reinterpret* Arrow's Impossibility Theorem as a result on information, one that gives a resounding, unqualified affirmative answer to Sen's question.[11] To do so with maximal clarity, I adopt one of Arrow's interpretations of his social choice function: it reflects the rankings of an external observer (or perhaps of one of the N individuals, charged with taking others' rankings into account)[12].

[10] Arrow (1951, 1963)

[11] What is involved is a *framing exercise.* There is no new decision theory or new social choice theory here, and the conclusion drawn will rely entirely on the power of Arrow's original theorem.

[12] See Arrow's discussion of "Primus" in Arrow (1984, p.56).

The discussion in Arrow (1950, pp. 342, 343) leaves the impression that a rule (or observer) violating IIA is erroneously trying to make interpersonal comparisons where none can be meaningful. I wish to argue that every sympathetic observer will find the information banned by IIA strictly valuable. Hence, rationality requires **not** the satisfaction of IIA, but its violation!

Instead of assuming that IIA must be imposed to guard against irrationality, I propose to model rationality directly. What would we normally ask of the observer, to view her as rational? Speaking informally, we would like her to have a coherent view of the world, and to have clarity about what she knows and what she doesn't know. Ideally she would have clearly formulated preferences, and beliefs that get updated sensibly when new information arrives. In other words, we have in mind a rational Bayesian.

How best to translate this to a formal assumption? The profession's "gold standard" for rational behavior is given by Savage (1954) or (often easier to work with) Anscombe and Aumann (1963). They would ask for complete, transitive preferences over a set of acts mapping states to prizes, and associated probabilistically sophisticated beliefs.

Accordingly, my sympathetic external observer is a Bayesian subjectivist in the tradition of Savage. Initially, she knows little about the N individuals' tastes regarding allocations in A; she doesn't even know how they rank them. She formulates a state space S that captures this uncertainty. A particular $s \in S$ represents in the observer's mind, one possibility regarding what each of the allocations would be like for each of the N members of society.

For such an $s \in S$, there are individual preference orderings $\succsim_i^s$, $i = 1, ..., N$, assumed complete and transitive. Acknowledging that ex ante, any preference profile is possible, the observer has at least one state for each preference profile. In principle there might be many states corresponding to a single preference profile. Indeed, if I were the observer, I would certainly have many states for each profile. A preference profile is consistent with many ways of experiencing different allocations in A.

For each profile ρ of citizen preferences over A, define the set

$$S_\rho = \{s \in S \mid \text{the preference profile at } s \text{ is } \rho\}$$

An *act* f associates with each $s \in S$ an allocation $f(s) \in A$ (or in Anscombe and Aumann (1963), a lottery over A). The observer's preferences $\succsim_{\mathcal{O}}$ are assumed to satisfy:

- $\succsim_{\mathcal{O}}$ is complete and transitive
- all states in S are non-null (all states are considered possible)
- $\succsim_{\mathcal{O}}$ satisfies the Anscombe and Aumann independence and continuity axioms (see Kreps, 1988, chap. 7)

This is enough to yield an additively separable utility representation for the external observer's preferences over acts (Kreps, 1988, chap. 7). Here, we are certainly dealing with *state dependent preferences*. The ex ante uncertainty is all about how individuals feel about, and rank, the allocations in A.[13] As usual (unlike state independent preferences), it is not possible to identify "as if beliefs"

[13] I am **not** assuming that the observer can rank different (state, allocation) pairs. (Maybe she feels she can, but we don't need this, and assume nothing one way or the other.) Similarly, I am **not** assuming the observer believes that different individuals' utilities are comparable to one another. (Again, maybe she feels she can compare these, maybe not.)

separately from utilities, instead writing the utility of an act f as:

$$U(f) = \sum_{s \in S} u^s(f(s))$$

Our assumption on $\succsim_{\mathcal{O}}$ also yields, for each reported preference profile ρ, conditional preferences $\succsim_{\mathcal{O}}|_\rho$ (see Kreps, 1988, chap. 10) which induce conditional preferences over A that can be represented[14] as

$$V(a \mid \rho) = \sum_{s \in S_\rho} u^s(a)$$

The role of the observer is to reconcile differences in reported preferences of citizens: when some prefer x and others prefer y, her conditional preferences provide a social order of x and y nonetheless. If instead there is unanimity, there are no differences to adjudicate, and it is natural to assume the observer's preferences reflect the unanimity.

Definition 1. The observer's preference ordering $\succsim_{\mathcal{O}}$ satisfies *nonpaternalistic sympathy* if, for each non-null $s \in S$, $x, y \in A$ and acts f and g, if

(i) $x \succ_i^s y$, $i = 1, ..., N$

(ii) acts f and g differ only on s

(iii) $f(s) = x$ and $g(s) = y$

then $f \succ_{\mathcal{O}} g$.

Notice that, given the representation of $\succsim_{\mathcal{O}}$, for state s and alternatives x and y satisfying the maintained hypotheses of the definition, we have

$$u^s(x) > u^s(y) \tag{1}$$

We have seen that for each reported profile ρ, $\succsim_{\mathcal{O}}$ induces an ordering of A. In other words, $\succsim_{\mathcal{O}}$ generates a social choice function, call it $\succsim$. It has universal domain, by the "no null states" assumption. And it satisfies the Pareto condition: for any $x, y \in A$ and profile ρ with $x \succ_i y$, $i = 1, ..., N$,

$$V(x \mid \rho) - V(y \mid \rho) = \sum_{s \in S_\rho} \{u^s(x) - u^s(y)\}$$

By (1), each term on the right hand side is strictly positive, so $\succsim$ ranks x strictly above y under ρ.[15]

Just to make a point very clearly, I will make one more assumption to be relaxed in a moment.

- Suppose that individuals are numbered randomly before their preferences are reported to the observer, and that she therefore views them symmetrically.

[14] The states inconsistent with the profile ρ are, naturally, missing from the sum in the representation. One might expect the surviving terms to have increased coefficients (as with conventional Bayesian updating), but here in our "state dependent payoff" world, we are not working with absolute probabilities . The formula for $V(a \mid \rho)$ ranks allocations appropriately (and we are not concerned with absolute levels).

[15] To establish this ranking, one only needs a criterion that respects state-by-state dominance. That is, the same ranking applies under much weaker hypotheses than additive separability or maximization of expected utility.

Then the observer's scf also satisfies nondictatorship: if any i were dictator, then each j would have to be dictator as well, to be treated symmetrically, and this is impossible.

It is easy to write down state spaces and preferences satisfying all the assumptions made on our rational observer. As we have seen, they imply that she satisfies Arrow's axioms $\mathcal{U}$, $\mathcal{D}$ and $\mathcal{P}$. Hence by Arrow's Theorem, our rational observer's scf **must violate** IIA.

What does this mean? It says that a perfectly consistent rational Bayesian observer can treat citizens symmetrically and respect unanimity, but to do so, she **must** have some of the (ordinal) information that IIA would deny her. In other words, when Sen asks if the excluded information might not possibly have signified something useful, the answer is: yes, the excluded information is *always* useful. Without it, the sympathetic rational observer could not have ordered allocations the way she would like to do. Rather than satisfaction of IIA being a badge of rationality, it is evidence of *irrationality*. To cast this in Shakespearean terms:

> IIA would be more honoured in the breach than in the observance.

Some readers will ask why we need any decision theory to come to these conclusions. There is nothing wrong with someone treating people the same and respecting their unanimous preferences. If that is inconsistent with IIA, it means the information excluded by IIA is strictly valuable to *every* such observer, and to impose IIA would be destructive.

I am inclined to agree; it should be obvious.[16] But it was not obvious to Arrow and Samuelson in 1951 and for many years after. And in Section 4 I shall recount how the traditional dark interpretation of Arrow's Impossibility Theorem had a profound negative effect on attitudes toward welfare economics and social choice in the profession at large, lasting to this day. My hope is that modelling the sympathetic observer explicitly as an ideal example of Bayesian rationality will help readers take a perspective from which Arrow's Impossibility Theorem is evidently a powerful theorem about information, not a disturbing paradox.[17]

I promised to relax the assumption that the observer treated all individuals symmetrically. Let's do that now. What was the purpose of imposing symmetry? It was simply to rule out a dictatorial solution. Logic is never going to rule out such a solution. Perhaps the observer believes that individual 17 is God, and therefore She *should* be made dictator. Or maybe $N = 5$, and the fifth person is such a hypersensitive child that it is best always to give precedence to her preferences. There being no absolute reason for a nondictatorial solution, we just have to consider two cases:

Case 1: It's ok if your observer appoints a dictator. Then you're all set, and Arrow's Theorem is no

[16] I hasten to add that the universal importance of the information banned by IIA is obvious only because of Arrow's Theorem! If I didn't know that result, I would imagine I could make up crude enough observer preferences that she could always rank x and y without looking at allocations other than x and y. The fact that one can never construct such an example is a surprise, until one knows Arrow's result.

[17] Piotr Dworczak suggested to me (private communication, August 2020) that the Judgement of Solomon illustrates the relevance of information excluded by IIA. "Solomon must decide which of the two possible mothers, woman 1 and woman 2, should have the right to the baby. The two salient alternatives are: I) baby is given to woman 1, II) baby is given to woman 2. To these Solomon adds alternative III) baby is cut in half. The social choice function Solomon adopts in his solution has the property that if woman 1 has a ranking $I > II > III$ while woman 2 has ranking $II > III > I$, then woman 1 gets the baby, whereas if the rankings are $I > III > II$ and $II > I > III$, respectively, then woman 2 gets the baby. But this obviously violates IIA! According to IIA, Solomon should rank I and II the same in these two profiles." The Biblical story is one of incentive compatibility and information revelation, but even from the Arrovian perspective of known preferences, this is a telling example. Sam Jindani sent me an example illustrating how the desire to compromise in multidimensional decisions may militate in favor of violating IIA.

obstacle.
Case 2: It's bad if your observer appoints a dictator. Then if you don't let her violate IIA, bad things will happen: either she appoints a dictator, or she hurts everyone (rejects strict Pareto improvements).

Taking Case 2 to be the one usually considered relevant, we have an alternative interpretation of Arrow's Theorem:

Reinterpretation of Arrow's Theorem. *If you insist on throwing away critical ordinal information, bad things will happen.*

To me, this is altogether a more natural interpretation of the theorem than the doom and gloom of the traditional reaction (disturbing, paradoxical, a challenge to democracy). The interpretation presented here is the view taken in Pearce (1995), and independently, I discovered recently when researching this Frisch Memorial Lecture, by Saari (1995).[18]

> The key fact is that IIA restricts what type of information can be used about each pair... . Thus, we arrive at a new interpretation: Arrow's theorem asserts that the ignored information is vital... . Observe that this informational perspective toward Arrow's theorem is a far more benign explanation than some of the draconian interpretations found in the literature. (Saari, 1995, p. 196)

Exactly! This is music to my ears. The question is, how did the standard interpretation ever take root? It depends upon a commitment to IIA as an implication of rationality. That comes from a misplaced positivist belief to the effect that"nothing could possibly be learned about someone's preference for x over y, from considering how each compares to other alternatives, because it is meaningless to ask the nature of the preference for x over y." How x is preferred to y belongs to the"further utility information" banned by Samuelson and Arrow.

There are other arguments made for IIA, one of them so frequently that it should not be neglected here: the *dead candidate argument.*

> "Suppose that an election is held, with a certain number of candidates in the field, each individual filing his list of preferences, and then one of the candidates dies. Surely the social choice should be made by taking each of the individual preference lists, blotting out completely the dead candidate's name, and considering only the orderings of the remaining names in going through the procedure of determining the winner. That is, the choice to be made among the set S of surviving candidates should be independent of the preferences of individuals for candidates not in S." (Arrow, 1951, p.26)

I do not think the procedure described by Arrow is desirable. Suppose X wins the election, and then dies. A Bayesian approach would take the information learned from the election to be useful, and use it to choose the best remaining candidate. Arrow's procedure instead ignores the information concerning how X was ranked against all other candidates. But although X is no longer available, it is not wise to throw away possibly useful ordinal information. Candidate X is now irrelevant as a candidate, but not informationally irrelevant. On the other hand, suppose X was elected and Y dies. Then Arrow's procedure, with an IIA voting ruling, ensures X remains the victor. Had X

[18] Saari did not consider an observer nor a decision-theoretic framework. He instead used geometric and algebraic observations to arrive at the same information-theoretic interpretation of Arrow's result.

been elected by a Bayesian procedure ignoring IIA, then following Y's death, Bayesian reasoning would also maintain X as victor

Hildreth (1953) was an early critic of IIA. Arrow objected to the use of preference profile information to make rankings based on interpersonal comparisons, and therefore imposed IIA (Arrow, 1950, pg. 342). Hildreth pointed out that if you write down a nondictatorial social choice function, you have already used preference profile information to make rankings based on interpersonal comparisons; if this is forbidden, there can be no acceptable nondictatorial social welfare functions, and one doesn't need Arrow's Theorem to prove it. For other critical perspectives on IIA, see for example Rothenberg (1961), Gibbard (1968/2014), Hansson (1973), Mayston (1974), Bailey (1979), Pazner (1979), Lehtinen (2007), Fleurbaey and Maniquet (2008) and Coakley (2016). Many of these, notably including Rothenberg (1961), propose weakenings of IIA, as do Young (1976) and Maskin (2020). Not all of them are aware of the others' work. Lehtinen (2007) is more concerned with strategic issues, but his title is on target: "Farewell to IIA". Arrow himself gradually softened his insistence on IIA: see his remarks in Arrow (1967, pg. 19).

Do the informational objections to Arrow's IIA apply equally to the independence of irrelevant alternatives imposed by Nash (1950) on solutions to his two-person bargaining problem? I would say no. In Arrow's world, relaxing IIA yields valuable ordinal information about pairs x and y under consideration. In Nash's world, each alternative is presented as a pair of von Neumann-Morgenstern utils, so one has rich information about any two alternatives x and y without reference to any other points.

It may seem remiss to model a sympathetic observer as a rational Bayesian decision maker without mentioning John Harsanyi, who certainly took a Bayesian approach to social choice. But his treatments explicitly use "further utility information", in ancient Greek village terminology, and hence do not belong to Arrow's world. His remarkable ideas will be discussed in Section 5.

4 Impossibility and Welfare Economics

As the 1940s drew to a close, the question of a social welfare maximum was widely considered "underdetermined": honest men and women might disagree about how different allocations of resources "ought" to be ranked. With the "new welfare economics" less conclusive than originally hoped[19], the profession turned mainly to the social welfare theory of Bergson (1938) and Samuelson (1947) for guidance. Those authors clearly asserted that a value judgement from outside of economic analysis is needed to resolve these questions.

> "Without inquiring into its origins, we take as a starting point for our discussion a function of all the economic magnitudes of a system which is supposed to characterize some ethical belief - that of a benevolent despot, or a complete egotist, or 'all men of good will', a misanthrope, the state, race, or group mind, God, etc." (Samuelson, 1947, p.221)[20]

[19] de Scitovsky (1941) showed that the Kaldor (1939) and Hicks (1939) criteria for a welfare-improving policy change could exhibit embarrassing cycles. Moreover, it was hard to argue that a potential compensating payment, say from the rich to the poor, not actually paid, was a good substitute for actual payment

[20] It takes some confidence to end a list with "God, etc."

Initial resources, tastes and technology determine a utility possibility frontier, and then the function referred to in the quote above, now known as a Bergson-Samuelson social welfare function (swf), could choose a particular Pareto efficient point. But different swf's would choose different respective solutions (see the solid and dotted families of level sets in Figure 1, and their respective optimal solutions). With no particular authority for any one swf, the problem still seemed essentially underdetermined.

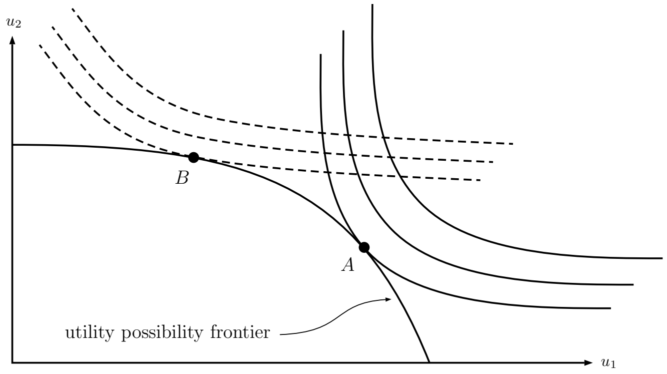

Figure 1

By the early 1950s, with the publication of Arrow (1950, 1951), the dominant view in the profession abruptly shifted: the social choice problem was *overdetermined*. **No** solution could meet even minimal criteria (our old friends $\mathcal{U}$, $\mathcal{D}$, $\mathcal{P}$ and IIA from Section 3). Arrow asserted early (1950) and repeatedly that his impossibility theorem also showed the impossibility of a Bergson-Samuelson social welfare function.[21]

> "Hence the Possibility Theorem is applicable here; we cannot construct a Bergson social welfare function. " (Arrow, 1950, p.346)

Samuelson disagreed vehemently, holding that a society faces just *one* profile of preferences, and that there is no need for it to confront Arrow's multiprofile problem and his axioms $\mathcal{U}$ and IIA. In Figure 2, Arrow starts at the black dot on the left, awaiting the particular preference profile that nature or history throws up. He wants a good aggregation procedure (scf) to rank alternatives appropriately whatever profile arises, and proves that none can satisfy his basic axioms. Bergson and Samuelson want to start instead at the open circle in the center of the Figure, and study how resources will be allocated according to any particular criterion (swf). Arrow points out that whatever that criterion is, it will look bad "in the large", that is, beginning at the black dot. Samuelson says in effect that society does not start at the black dot, and neither do he and Bergson.

Samuelson's "one profile at a time" stance was not persuasive, and many wondered if Arrow's Theorem was the end of welfare economics. For documentation and deep discussion of Samuelson's evolving defensive positions on this question, see Fleurbaey and Mongin (2005). Their title reflects

[21] Bergson and Samuelson studied a standard economy, rather than an abstract social choice problem. In their world it is natural to impose certain restrictions, such as monotonicity of individual preferences. But adaptations of Arrow's axioms produce notable impossibility results in economic environments as well. For an excellent survey see Le Breton and Weymark (1996).

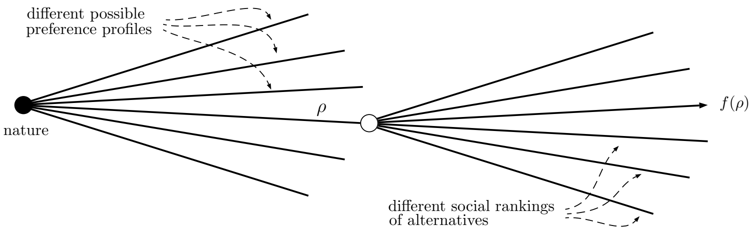

Figure 2

the gravity of the threat to welfare economics posed by Arrow's Theorem, chased by some hope: "The News of the Death of Welfare Economics is Greatly Exaggerated."

Having defended social welfare functions against Arrow's Theorem by adopting a "single profile" stance, Samuelson was particularly vulnerable when papers by Kemp and Ng (1976) and Parks (1976) claimed to have produced single-profile analogues of Arrow's Theorem. What could that mean? Arrow's IIA insists that each pair x and y of alternatives be ranked without reference to any person's preferences beyond x and y. So, comparing two different preference profiles for society, if each person ranks x and y the same way under the one profile as under the other, so must the social ordering. Clearly this says nothing if there is only one profile that will be considered. But these new papers impose a "neutrality" assumption combining IIA with a "what's good for the goose is good for the gander" rule: all pairs of social allocations must be adjudicated in the same way. For example, suppose x is ranked strictly ahead of y, socially, and each person ranks w relative to z the same way as she ranks x relative to y. Then just as society ranks x strictly ahead of y, it should rank w strictly ahead of z. At least, that's what is required by "neutrality" (treating each pair of alternatives the same).

Samuelson (1977) provided a clever example involving the division of 100 chocolates between two citizens, that showed how unconvincing neutrality is when obvious matters of distribution are concerned. One might reply that for cases that are not obviously matters of distribution (for example, the choice among a list of possible public goods), neutrality might be highly reasonable, and lead to a "single profile" impossibility theorem. In any event, the single-profile impossibility theorems[22] were taken by most to be a further blow to Bergson-Samuelson social welfare theory:

"The message got across to the non-specialists, and it became part of the official history of economics that a major refutation had taken place. If the official death of welfare economics were to be dated with some precision, the years 1976-79 would suggest themselves." (Fleurbaey and Mongin, 2005, pg. 382)

More generally, Sen writes that Arrow's Theorem

> "... generated further pessimism in an already gloomy assessment of the possibility of a reasoned and satisfactory welfare economics." (Sen, 2017, pg. 12)

Why didn't Samuelson just say, in 1950, that IIA was indefensible and therefore Arrow's Theorem posed no problem for Bergson-Samuelson welfare theory? My theory is that IIA looked to

[22] Advances to the single-profile literature include Pollak (1979), Roberts (1980b) and Rubinstein (1984). See also Feldman and Serrano (2008) and the references therein.

Samuelson like the intellectual child of one of his own tenets: the meaninglessness of the "further utility information" (ancient Greek village terminology.) Discussing Arrow's Theorem, Samuelson combines two of the axioms and then says:

> "All three Axioms seem reasonable. Arrow's great feat was to prove that no Constitution Function can satisfy them all. ... Which axioms should we reject? It is like asking which triplet one should put up for adoption.
>
> Many people think that the IIA is the one that should go. I cannot agree." (Samuelson, 1967, pg. 47)

There is no doubt that the impossibility theorems produced by the modern social choice literature did great harm to traditional welfare economics. But views in both camps drifted over time, in complicated ways. Even the principals of the two approaches, Arrow and Samuelson, changed their positions over the decades. On this matter, and on the debates between the two, Igersheim (2019) is especially enlightening. Apart from the softening of his views on IIA noted in Section 3, Arrow modified his position (but not entirely consistently) on the implications of his theorem for Bergson-Samuelson analysis; see the following passage referring to an earlier complaint of Samuelson's[23] about incorrect rumors:

> "... if there are rumors that 'Kenneth Arrow's Impossibility Theorem rendered Bergson's social welfare function somehow nonexistent or self-contradictory', they are indeed quite confused." (Arrow, 1983, pg. 21)

It is hard to reconcile that with the corresponding Arrow (1950) quote in Section 3 above saying that we cannot construct a Bergson social welfare function.

And after all of Samuelson's support of IIA, he grudgingly admits that he and Bergson are violating the axiom in some of their constructions. Of social choices, (Samuelson, 1987, pg. 170) says:

> "Once we agree that a choice *legitimately* can depend on what 'types' our persons are, and agree that defining people's types can depend on *more* than... binary choosings, then I must agree with Bergson's contention that, operationally we are explicitly (and reasonably) deciding to violate the axiom of Independence of Irrelevant Alternatives. Third states of the world seem to force themselves legitimately into our binary choices. Most ethical systems purport to define who is the deserving one by how the contemplated individuals react to a vast panoply of possible situations."

Samuelson's choice of words speaks eloquently of his discomfort at what he is saying: allocations beyond a pair being ranked "seem to force themselves legitimately..." — it is clear that they are unexpected and unwelcome, and yet cannot be denied. And little wonder he is unhappy with what he is writing: Paul Samuelson seems to be recanting IIA: the triplet he had defended so earnestly is being put up for adoption after all. Heaven and earth should have been shaking that day. But I'm not sure who was paying attention. Much of the profession had lost faith in welfare and social choice, and had abandoned the scene.

[23] (Samuelson, 1981, pg. 223) had complained of "the quite confused rumors that Kenneth Arrow's Impossibility Theorem rendered Bergson's 'social welfare function' somehow nonexistent or self-contradictory".

How long lasting has the damage done to welfare economics been? Sir Anthony Atkinson spoke to what he called the strange disappearance of welfare economics:

> "Welfare economics should be a central part of the discipline. But it is not. While welfare economics was a subject of importance half a century ago, today it has largely disappeared from the mainstream... ." Atkinson (2009)

In my view, logical positivism and the axioms it produced (IIA and, in the single profile case, *neutrality*) were largely responsible, via the impossibility theorems, for the decline in the confidence most of the profession placed in welfare economics and, for that matter, in social choice. Do the negative social choice results of the 1950s and 1970s continue to have that influence on the profession's attitudes to welfare economics? Although, as I mentioned earlier, there is now much greater diversity of opinions among specialists about independence axioms and the supposed meaninglessness of information beyond a preference profile, it is my impression that traditional, and what I would call unduly conservative, views of individual and social welfare, are still highly influential.

Consider what students are told about Arrow's Theorem. At the introductory level, the celebrated text by Greg Mankiw states:

> "Arrow's Impossibility Theorem is a deep and disturbing result... ." (Mankiw, 2015, pg. 495)

I think the result tells us that imposing IIA is always a bad idea, but I find that natural, rather than disturbing.

What message do doctoral students get? One of the most respected texts says:

> "The result of Arrow's Impossibility Theorem is somewhat disturbing... . What it shows is ... that we should not expect a collectivity of Individuals to behave with the kind of coherence that we may hope from an individual." (Mas-Colell, Whinston, and Green, 1995, pg. 799)

I don't read the theorem that way. Section 3 argued that it is irrational for an *individual* observer to satisfy IIA. The restriction is a bad idea for an individual or a group.

Another distinguished graduate text says:

> "Modern social choice theory begins, and in some senses ends, with a remarkable result variously known as Arrow's Possibility Theorem and Arrow's Impossibility Theorem." (Kreps, 2013, pg. 166)

Modern social choice theory certainly begins there. And it is a large part of the problem that many see it as ending there. Accepting the traditional interpretation of the theorem, they see reasoning about social welfare as hopeless.

What word do we get from the specialists? Many social choice scholars have followed Sen in going far beyond ordinal information in analyzing social welfare. Yet in some quarters, IIA retains much of its allure. One of the most decorated of all contributors to social choice remarks that, while no aggregation method can satisfy Arrow's four axioms all the time, we can still ask:

> "... which rule satisfies them **most often**? In other words, if we can't achieve the ideal, which voting rule gets us closest to that ideal and maximizes the chance that the properties we want are satisfied?" (Maskin, 2014, pg. 52)

This is very much in the traditional manner of interpreting Arrow's Theorem and IIA, and not a posture with which I agree. But Maskin takes an eclectic approach, elsewhere exploring ways of relaxing IIA (again, see Maskin (2020)).

There continues to be a great deal of serious work done on welfare economics, both theoretical and applied. I agree with Atkinson: I would like it to regain its place in the mainstream of the profession, and to see less methodological suspicion of informationally adventurous work. With this in mind, I hope that we can work at updating the broad profession's interpretation of the classic impossibility results, and of axioms such as IIA and neutrality. More than that, let us venture beyond Arrow's world to what might be called Sen's world. That is the subject of the next Section, which concerns itself with richer utility information than preference orderings afford.

As we prepare for that transition, I should say that I regard most of the content of Section 5 as positive questions, not normative. Most treatments of the distinction between the two begin with the assertion that, roughly speaking, positive economics concerns the way things are and how things work, whereas normative economics examines how things "ought" to be. The latter involves value judgements; the former does not. According to that classification criterion, I would say that what my welfare is under a particular allocation is a positive matter. It is a fact of life; there is nothing normative about it.[24] Once I have made the best of my circumstances, my life is a certain way, and "ought" does not enter into it. The fact that it is hard for others to understand precisely how I experience things does not make the issue normative; not all positive phenomena are transparent.

It must be said that many accounts of "positive versus normative" go on to remark that positive statements can be tested. This seems in contradiction to the criterion already laid down. If something can't be tested, why need it involve an "ought" statement? If you can't discern how I feel, how I feel remains a fact of life (known to me and not to you), and ought does not come into it. A possible resolution is to create an "unexcluded middle", as it were: there is a third category, namely, meaningless. This leaves Samuelson room to say that, although how David feels is not an "ought" statement, neither is it positive; it is meaningless. I take no personal offence at this. But I have suggested, in Section 2, illustrated in Section 3, and chronicled here in Section 4, that this use of words has proved destructive in welfare economics.

If I understand him correctly, Milton Friedman uses these terms in the same way as Samuelson. In his entertaining essay "What All Is Utility?", Friedman (1955) begins with the epigraph:

> "When I use a word," Humpty Dumpty said in a rather scornful tone,
> "it means just what I choose it to mean — neither more nor less."
> "The question is," said Alice, "whether you *can* make words mean so many different things."
> "The question is," said Humpty Dumpty, "which is to be master – that's all."

Friedman takes Dennis Robertson to task for his use of introspective utility[25] in "Utility and All That" (Robertson, 1951). As I will argue in Section 5, introspective welfare may be difficult to measure,

[24] I fully recognize that my current welfare is endogenous, and profoundly socially conditioned. One may think my feeling of contentment unforgivable, or that social pressures on me are immoral. Those are different matters. Whether I am riding a bicycle is a question of "is", even if you think the bike "should" have been allocated to someone else. Whether I am feeling proud is a question of "is", even if you find my pride contemptible, or the product of a warped society.

[25] See Kahneman, Wakker, and Sarin (1997) for a more inclusive attitude toward Benthamite utility.

but it refers to a particular thing. One can argue about how best to think of its units, but the welfare of a particular person in a specific circumstance has absolute reality. It may be inscrutable, but it is a positive concept, even if not a *positivist* concept (that is, a claim having Samuelsonian meaning – see Section 2). There is a place for introspective utility in scientific inquiries into welfare, even if not within narrowly positivist science.[26]

Friedman closes by saying:

> Science is science and ethics is ethics; it takes both to make a whole man; but only confusion, misunderstanding and discord can come from not keeping them separate and distinct, from trying to impose the absolutes of ethics on the relatives of science. To put it in Humpty Dumpty's words again: "It is a — MOST — PROVOKING — thing," he said at last, "when a person doesn't know a cravat from a belt!"

And, I would add, when a person conflates science and logical positivism.

5 The Further Utility Information

As we have seen, the great pioneers of mid-twentieth century social choice and welfare warned against thinking of utility as having intensity, being measurable, or having essentially cardinal meaning. Are these three the same thing? I think that there is a natural usage in which the three terms refer to quite distinct things, each of them meaningful. Further, is utility interpersonally comparable, and how does this relate to social welfare? These are even more difficult matters; let us postpone them until later in this Section, starting instead with questions of individual utility.[27]

It should be uncontroversial when I say that some preferences, of a given individual, are more intense (greater, more important...) than some other preferences of the same individual. Consider some individual i who has complete, transitive preferences $\succsim_i$ over $\{w, x, y, z\}$ as follows: $w \succ_i \; x \succ_i \; y \succ_i \; z$

$$\left\{\begin{array}{l} w \\ \left.\begin{array}{l} x \\ y \end{array}\right\} \\ z \end{array}\right.$$

I think it is entirely reasonable to say that i's preference for w over z is greater (more intense, more important...) than his preference for x over y. Certainly every ordinal representation U of these preferences exhibits

$$U(w) - U(z) > U(x) - U(y)$$

For me, that might mean that as a sympathetic observer I would make another individual j sacrifice something so that i could get w rather than z, but NOT so that i could get x instead of y.

[26] Conceptions of science keep changing. When A.J. Ayer, the celebrated logical positivist and author of "Language, Truth and Logic" (1936) was asked by Bryan Magee about the fall of logical positivism, Ayer said that its principal defect "was that nearly all of it was false". See Ayer (1978).

[27] While many contemporary social choice theorists will agree with various nonempty subsets of the positions I sketch in this Section, I suspect there would be more resistance from the broader profession. After reading a particular paragraph, some readers will think yes, we have known that for a long time. Of the same paragraph, others will think no, this is meaningless rubbish. I hope that discussion of this general territory will eventually narrow some of these gaps.

What this means about interpersonal comparability I will again delay discussing until later in this Section. But in my mind, the fact that some differences can matter more to an individual than other differences certainly militates against IIA.

Even readers who agree that i's preference for w over z can be viewed as "revealed more intensely preferred" than his preference for x over y, may object that it is still meaningless to compare i's preference for w over x to his preference for y over z, for example. (That is, although one preference being "more intense than" another is a sensible concept, that may not be a complete ordering.) I agree that this is a more subtle matter, concerning whether or not the preferences are naturally viewed as cardinal. Professor Arrow has interesting things to say regarding the cardinality of utility:

> "The problem of measuring utility has frequently been compared with the problem of measuring temperature. This comparison is very apt. Operationally, the temperature of a body is the volume of a unit mass of a perfect gas placed in contact with it (provided the mass of the gas is small compared with the mass of the body). Why, it might be asked, was not the logarithm of the volume or perhaps the cube root of the volume of the gas used instead? The reason is simply that the general gas equation assumes a particularly simple form when temperature is defined in the way indicated. But there is no deeper significance. Does it make any sense to say that an increase of temperature from 0° to 1° is just as intense as an increase of temperature from 100° to 101°? No more can it be said that there is any meaning in comparing marginal utilities at different levels of well-being." (Arrow, 1951, pg. 10)

Arrow offers this as a reason for not considering utility "measurable". But is *temperature* not measurable? What else are thermometers for? I would like to use measurable in the everyday, practical sense of referring to something whose level or magnitude we can determine clearly and communicate unambiguously (like the speed of a car, the length of a building, or the temperature in a particular location, abstracting in all cases from non-Newtonian phenomena). That is not what Arrow is talking about in the above passage. He is arguing instead (and I agree) that there is no fundamental sense in "counting" successive degrees of temperature, as if each were "the same" increment, the way we count people arriving at a party. If there are twenty people at the party, or instead forty, the number attending increases by the same amount in either case when five new partiers arrive. In that sense, we say that the number in attendance is cardinal, whereas the temperature really is not.

For twenty years Bill Brainard and I were colleagues in the Department of Economics at Yale. One day on the way to lunch, I asked him if he thought of utility as being ordinal or cardinal. He said: "I always ask myself, cardinal for what *purpose*?" A wise answer. Interestingly, there is something subjective about whether something is cardinal. It is not like asking if addition is commutative or if Scotland is part of Austria. Rather, it asks whether, for the purpose at hand, it is *natural* to identify various increments as being "the same" in a relevant sense, and then to count them.

To contrast utility and temperature in this respect, consider the following two scenarios.

Scenario 1. Suppose you teach a class on expected utility theory, and assign a problem where an individual with a particular von Neumann-Morgenstern utility function has to choose how much insurance to buy. A student emails you asking: "The utility function you gave us is one of many equivalent ones. But which scale is being used here? For this representation, what is 50 utils, for

example?"
You reply: "It doesn't matter; you don't need any information of that kind." You make a note that this student may need some extra help.
Scenario 2. You wake up and say: "Siri, I'm in zip code 10012. What's the outdoor temperature?"
Siri: "I know where you are. The temperature is 30."
You: "Is that 30° Fahrenheit?"
Siri: "Possibly, or Celsius. The scales are equivalent. I no longer record that distinction."

You realize with horror that Siri has been scanning the economics literature. Having read that temperature is like utility, and that a utility number has no significance by itself, she has concluded that when asked for a temperature number, as opposed to a temperature difference, it would be pointless to identify the scale.[28] Not being human, she does not perceive that temperature has *absolute meaning.* Thirty degrees Fahrenheit has a set of physical properties. Water will freeze at thirty Fahrenheit. Vodka will not freeze at thirty Fahrenheit. Having agreed with Arrow that temperature is *less* than cardinal, in the sense he explains so well, we have to admit it is also in an important way *more* than cardinal: it has absolute meaning.

And what of utility? Classroom utility, as we teach it in intermediate microeconomics, for example, is virtually always ordinal, or, for the *purpose* of expected utility maximization, cardinal (and yet not so obviously composed of comparable, and therefore countable, units, that Arrow would want to call it essentially cardinal). But it is never absolute. We never find out if the agent buying insurance is happy or miserable; that is not even taken to mean anything! So, concerning Arrow's assertion that the comparison between temperature and utility is apt, I would rather say it is instructive, for their differences.

Let us move now to actual utility. What on earth do I mean by *that*? We don't have that in standard economics. Once I explain what I mean by it, you will see why we don't have it in traditional economics: it is an offense to positivism, so we don't have it. By actual utility I mean, at a first pass, a projection onto the real number line of the degrees of satisfaction a particular person has with different choices or allocations, *endowed with absolute significance.* Don't worry about how to choose the units; that discussion is coming soon. For now, I just want to insist that if John strictly prefers situation A to situation B, the former is assigned a higher actual utility, perhaps 80, than the latter. And when John says his actual utility in situation A is 80, that has absolute meaning to him: situation A is real, his experience of it, were it to arise, is real and specific, and that degree of satisfaction he calls 80. In this way, it is akin to temperature: thirty degrees Fahrenheit does not have meaning only in comparison to other temperatures, but unto itself. Similarly, satisfaction of 80 means a specific degree of satisfaction on John's part. The difference between the 80 on John's scale and the 30 on the Fahrenheit scale is that there are things to point to in the directly observable external world to communicate to others what 30 F means. You can teach them how to interpret the Fahrenheit scale, how to extract absolute information from it. It is much more difficult, if even possible, for John to explain to the world what his 80 means, not in comparison to situations other than A in which he might find himself, but in absolute terms: what is that level of satisfaction? More generally, what does his scale mean? What respective absolutes do those numbers try to convey? Whereas I would call temperature measurable, in a practical sense, I consider actual utility much

[28] You wonder if Siri has filed a memo noting that you may need more help than she had estimated.

less easily measurable (more detail on this later).

The above "first pass" at defining actual utility ignores a host of complications. Situations A and B are multidimensional, perhaps involving different degrees of stimulation, anxiety, satisfactions of different wants, intertemporal profiles and so on. How are we to aggregate all these to arrive at a labelling of A having higher actual utility than B? Classroom utility resolves this beautifully through "revealed preference": A has weakly higher utility than B if A is chosen when B is also available. This is not an assertion, but a definition. What a relief to have such a clean resolution of such a potential mess! We don't have to theorize about how to aggregate complicated competing considerations: the individual is the aggregator. "Better" just means "chosen".

At a "second pass", I could qualify the description of actual utility the same way. John decides which of A and B has higher actual utility. And if he is indifferent between choosing A and C, he assigns actual utility of 80 to each of them. For the purposes of this paper, I would be happy to leave it there, but I have to say that there are many reasons to complicate the description. We are all too familiar with human frailty and have made many unfortunate decisions. Distraction, arithmetic incompetence, slick advertising, limited ability to plan or imagine, and many other problems intervene between choice and maximal satisfaction. [29] That statement clearly requires one to have some notion of satisfaction that doesn't depend on choice. This is, again, a mess, but I think most of us have intuitive, introspective instincts about happiness and satisfaction, that are not simply derived from the concept of choice. For a sophisticated and systematic opposing perspective that argues for "welfare without happiness", I recommend Gul and Pesendorfer (2007).

Much of the social choice literature has followed Sen (1970) in exploring everything from ordinal classroom utility to something akin to what I have called actual utility. Sen uses the idea of "admissible transformations of a utility function" to extend to broader questions the classroom distinction between an ordinal utility function, which can safely be replaced by any strictly increasing transformation, and a von Neumann-Morgenstern utility function, which can be replaced by any strictly increasing *affine* transformation. What if instead the level of utility referred to (as in "actual utility") conveys absolute information? Then in the social choice literature, no transformation of the function is allowed. This raises an intriguing question. How can 80 utils be the only way to describe John's satisfaction in situation A? Going back to temperature, where absolute information is also being expressed, we would never say that the Fahrenheit scale is the only one that is appropriate when conveying absolute information about temperature. There is a Celsius equivalent, and many other scales could be used just as well. The key thing is that the listener must understand the scale. If Sen says that no transformations of a given function U are admissible, when talking about one individual in isolation, he means that if you change to some other V, you at least have to alert the listener and provide a means of translation,[30] just as a weather forecaster would if she abandoned one scale in favor of another. In a social choice context there is more to this issue, and that discussion is coming

[29] In a conference in 1952, Maurice Allais famously posed some choice problems to Leonard Savage, whose choices violated his own axioms. Savage (1954, pg. 103) reports that, on reflection, he decided to change his choices to conform to the axioms. One can ask whether, in interpreting "'better' means "chosen"', chosen should mean chosen after time to reflect, chosen after expert advice, chosen after hearing other points of view, and so on. See, for example, Schotter (2003) and Rubinstein (2007).

[30] U still might have advantages over V, if for example it has a cardinal interpretation, so that utility differences measured by U are taken to be comparable in some relevant sense.

later.

So far, we can understand temperature as being measurable, absolute, and yet less than cardinal (as per Arrow), classroom utility as being measurable, totally lacking in absolute content, and less than cardinal, again in Arrow's sense, and actual utility as hard to measure and absolute — and what about cardinal? 1950s Arrow would presumably say not cardinal (and not even meaningful!), while Brainard may say be careful what you mean. I would say that actual utility is at least vaguely cardinal, in the following sense. Suppose you offer to supplement John's annual income by $ 10,000. Compare two situations: in the first, his pre-supplement annual income was $1,000, and in the second, it was $20 million. If you were to ask John in which of those two situations your supplement would result in a greater increase in his welfare (satisfaction, actual utility,...), I would be surprised if he had any trouble in saying the increment would be greater in the first case, where it would be life-changing in a sense it would not be in the second.[31]

That leaves us a long way from being able to say what shape John's marginal actual utility of money should take. What would let us assert, for example, that an increment of annual income of $1,000, starting at $100,000, would enhance his actual utility by the same amount as would an increment of $2,370, starting at $200,000? We would like some "external sacrifice" to use as a measuring stick: if John is indifferent at $100,000 to getting the extra $1,000 in return for the external sacrifice x (say, losing his hearing two weeks earlier in his distant nineties) and also indifferent at $200,000 to getting the extra $2,370 in return for the same accelerated hearing loss, it is tempting to say that the two income gaps are equally important to him, and hence represent equivalent changes in actual utility. But we would want to know that the hearing loss did not interact in any way with current consumption utility: the sacrifice is a good measuring stick only if it doesn't change the length of what is being measured, and vice versa (shades of Heisenberg).

Does choice under risk present an opportunity to measure marginal utility? Vickrey (1945) and Harsanyi (1953) were sympathetic to the idea. Theirs were minority views.[32] After all, how can you measure something that is meaningless? Not viewing satisfaction as meaningless, but rather unknown, I find choice under risk a promising way to get strong hints about how much a person cares about different improvements in welfare.[33]

As Kreps (2013) explains, in consumer theory we usually don't expect separability across commodities: one sees substitutes and complements everywhere. But as states of the world are mutually exclusive, independence axioms seem rather natural. There is no physical interaction between con-

[31] Sen (1970) speaks of rough comparability across individuals, especially in Section 7.4, and one can ask analogous questions at the intrapersonal level.

[32] Nash (1950) uses von Neumann-Morgenstern payoffs to assign surplus in his bargaining game. But this does not suggest that he equates those payoffs with Benthamite utility of any kind. Rather, their relevance lies in their ability to describe behavior under the kind of risk that bargaining involves. The Nash demand game Nash (1953), his noncooperative implementation of his axiomatic solution, makes this explicit.

[33] This is an introspective concept which I believe is meaningful to many people, and which we are trying to measure by finding plausible "external sacrifices" to perform calibrations. What if I ask someone which of two nonoverlapping income differences matters more to him, and he asks: "For what purpose?" Then I admit I'm in messy territory, as is someone asking standard revealed preference questions and getting intransitive answers. Perhaps the different "purposes" can be viewed analogously to different framings, and treated after the fashion of Bernheim and Rangel (2007). I register sustained support from Frisch for cardinal utility Frisch (1926b, 1964) and from Fishburn (1970, pg. 82), who says: "Although our preference-difference comparisons may not be as precise as length comparisons made with precision instruments, I do not feel that this is sufficient reason to abandon the idea of such comparisons."

sumption in different states: all of that consumption is notional, except for the one state that finally arises.[34] That same idea (but I don't want to implicate Kreps in any of what follows) suggests that a given sacrifice in state 2, for example, could be used to calibrate marginal utilities (at various levels of money prize, say) in state 1. Because consumption in one state doesn't "cast a shadow" on consumption in another state[35], as consumption in state 1 rises, that given sacrifice in state 2 can be viewed as a constant external sacrifice, like a measuring stick.

We might find various consumption intervals in state 1, perhaps (20, 30), (30, 46) and (46, 71), each of which is "equally important" to this decision-maker. If he is an expected utility maximizer, all such constructions will give consistent answers, and the von Neumann-Morgenstern utility function is a natural candidate for "actual utility". Of course I don't expect anyone to be an exact expected utility maximizer: we know that even Savage failed that test. But I would still take the kind of intervals mentioned above to be of approximately equal importance to the decision-maker.

Of von Neumann and Morgenstern's theorem on expected utility, Arrow says:

> This theorem does not, as far as I can see, give any special ethical significance to the particular utility scale found. For instead of using the utility scale found by von Neumann and Morgenstern, we could use the square of that scale; then behavior is described by saying that the individual seeks to maximize the expected value of the square root of his utility. (Arrow, 1951, pg. 10)

Leave aside for a moment questions of ethics, focusing instead on understanding how much the individual cares about different income differences, for example. Recall that the argument I sketched earlier uses the mutually exclusive nature of different states, along with the same intuition that makes the von Neumann-Morgenstern independence assumption plausible (at least as an approximation), to suggest that equal increments in von Neumann-Morgenstern utils, at different income levels, "matter" roughly the same to the individual. There is no reason to think that equal increments in the square of that scale matter the same amount. So for me, as a guessing man, an expected utility maximizer's von Neumann-Morgenstern index reveals a lot about how much he cares about different increments in income.

To his discussion of the expected utility theorem, Arrow (1951, pg. 10) adds that it has nothing to do with welfare considerations because to say otherwise "would be to assert that the distribution of the social income is to be governed by the tastes of individuals for gambling." Again leaving ethics aside, does Arrow's remark suggest that the von Neumann-Morgenstern index is unhelpful with the calibration of utility for an individual in isolation, because measurement of the "amounts he cares about different income levels" is obscured by his attitudes toward risk? If John satisfies the independence axiom, the expected satisfaction he derives from one part of the state space seems to

[34] Despite this lack of physical interaction of consumption across states, there could sometimes be some psychological entanglement across states. Mark Machina has given appealing examples of this. I might like Cheerios for breakfast, but if I wake up and find that a nuclear holocaust that would have occurred in one state has been averted, I might want a champagne breakfast instead. Similarly, after hearing that an operation leaves my friend with a slight permanent limp, I take flowers to the hospital as a gesture of sympathy if the likely outcome of the operation had been no limp at all, but I take champagne and cigars to celebrate if instead the operation had had a serious chance of being fatal. See, for example, Machina (1981, pg. 172).

[35] Samuelson (1952, pp. 672) says: "Within the stochastic realm, independence has a legitimacy that it does not have in the nonstochastic realm..." because there is no reason choices respectively concerning mutually exclusive events should be "contaminated" by one another.

be unaffected (uncontaminated, to borrow Samuelson's word) by rewards elsewhere. Suppose, for example, that the "left half" and the "right half" of the state space are equally likely, and that under some particular act, John gets a constant payoff of \$100 on the right half. He expects that \$100 to contribute $.5U(100)$ to his total utility, regardless of what the act specifies about what happens in states in the left half of the state space. The latter might be very safe or riddled with dramatic risk; the right-half contribution remains $.5U(100)$. If we increase his payoff on the right half of the state space to \$150, the extra \$50 could be increasing the income variability over the entire state space (if he is getting a constant \$50 on the left half of the space, for example), or decreasing the income variability over the entire state space (if he is instead getting \$200 on the left half of the space). *Either way*, the extra \$50 on the right half increases his total utility for the act by $.5(U(150) - U(100))$. Where is his attitude toward risk, or attitude toward gambling?[36] He doesn't seem to evaluate things in those terms. For me, it is a natural guess to say that $U(x)$ reflects how much he cares about x, and he maximizes the mathematical expectation of that. There are no risk or gambling attitudes obscuring how much he likes x compared to y. Now another person, Jake, might actually enjoy or detest the simultaneous possibility of different realized rewards, finding that either stimulating or terrifying, and there is no reason to expect him to satisfy independence or to be an expected utility maximizer.

Earlier I said that I find individual welfare at least vaguely cardinal, for the simple reason that some matters seem decidedly minor compared to others. Choice under risk lets us make more precise guesses about the shape of a utility function, in a meaningful sense, especially for expected utility maximizers. Another method that found favor as early as Edgeworth (1881) is to note that a person is capable of only a certain fineness of gradation in his or her perception of anything: sound, taste, sight, and, one would suppose, satisfaction.[37] If, for example, John likes ice cream but can distinguish two quantities of ice cream if and only if they differ by at least one fortieth of an ounce, one might say of two commodity bundles differing only in that the first has one ounce more ice cream than the second, that "the utility of the first bundle for John is 40 utils higher than that of the second". And if Jake, another ice cream lover, is more finely perceptive and requires just a difference of one eightieth of an ounce to distinguish amounts of ice cream, and hence to get added pleasure, one might say that the first bundle specified above gives Jake 80 utils more than does the second. Can we proceed to say that Jake's welfare increases more than John's does, when they are both given an extra ounce of ice cream? No, because intuitively, Jake may not care as much about ice cream as John does, even though Jake has fine powers of discrimination. If you are in doubt about what "doesn't care as much" means (as your strict neoclassical forefathers would hope), especially when that's what this construction was supposed to be measuring, take the limit and see that Jake might be able to discriminate quantities of ice cream without liking ice cream *at all*. Just away from that limit, Jake hardly cares, but we would still be misleadingly saying that he prefers the first bundle to the second more than John does. Sadly, there are two dimensions at work, and we don't have

[36] Clearly I am not using these words in the usual sense; I understand that John may have a strictly concave expected utility function, and its curvature is conventionally seen as reflecting his attitude toward risk. I am saying that a person could conceivably have aversions to bearing risk, beyond the concavity of his utility function, but that is not the case with expected utility maximizers. They just care about utility, not its distribution.

[37] For other notable treatments of the "minimum sensible differences" approach, see Goodman and Markowitz (1952), Ng (1975), and Argenziano and Gilboa (2019). For broader discussions, see Abdellaouia, Barrios, and Wakker (2007).

identification of intensity of preference here (even for an intrapersonal argument, for that matter).

Yet another way to try to get a utility yardstick whose length is independent of what is being measured, is to make tradeoffs across distant time periods (as I was doing with the "earlier loss of hearing many decades hence" example above). Just as one finds plausible (even though not logically or even practically inevitable) a high degree of "noncontamination" of tastes across mutually exclusive events, the same might be true, with similar qualifications, for periods of consumption greatly separated in time. As unlikely as it might seem, Samuelson (1937) may have opened this discussion (but we don't know what other lost works of Gorgias there may have been). See the interesting work of Fryxell (2019) on the relationship between intertemporal consumption and utilitarianism.

It is time to embark on the "interpersonal comparability" leg of our road trip through further information land. Please ensure that your seatbelts are securely fastened. Terms on reflection may be more ambiguous than they appear. What does interpersonal comparison of utilities mean? Speaking of the set of Pareto efficient allocations in an economy, Samuelson (1947, pg. 244) states that "...without assumptions concerning interpersonal comparisons of utility it is impossible to decide which of these is best." I trust that he is *not* suggesting that a social welfare function resolves this by importing information that compares different persons' levels of utilities or marginal utilities: he has already pronounced those meaningless. Samuelson means that the observer (God, etcetera), whose allocational ethical preferences are captured in a particular social welfare function, prefers allocation A to allocation B, given a particular reported profile of citizen preference orderings.[38] Here, he is importing not information about utilities, but ethical judgements.

By contrast, when Sen (accompanied by much of the social choice literature that follows him[39]) speaks of degrees of interpersonal comparability, he means, explicitly, how much information is being "carried" by the sets of utility functions describing the respective individuals (recall the earlier discussion of admissible transformations). Entirely different from Samuelson, this in turn has two distinct interpretations. Consider the following statement that a social choice theorist might make: "We can determine an expected utility maximizer's cardinal utility function, but we cannot compare his utility to that of any other individual." The theorist might be saying that the comparative information across individuals is meaningful but not available (we have no way of accessing it), or she might be saying that she considers orderings across different individuals' utility levels to be meaningless.

My impression is that, although the "invariance transforms" approach is quite rich, it still has limitations. For example, suppose for some person, we have ideal information: unlike ordinal or cardinal utility, his utility function conveys exact information about his degree of satisfaction. **No** transformations of his function are allowed. The literature calls this "perfect measurability". Now suppose we happen to have this for each of the N individuals in society. Does that mean we also have interpersonal comparability? I don't believe so.

[38] As an aside, why would the observer care? The only thing that makes sense to me is that she cares about what each of the allocations is like for each citizen. She is comparing each citizen's experiences, introspectively perceived by each person, under A to those under B (if instead she were just comparing the physical allocations, she wouldn't need to condition on the preference profile). But this interpretation leaves the observer comparing one vector of things that Samuelson says are unobservable, and hence "meaningless", to another. What we are to make of preferences over pairs of meaningless things, I am not sure.

[39] See, for example, D'Aspremont and Gevers (1977), Roberts (1980a), Hammond (1991), and the references in Sen (2017) and in the superb survey by Fleurbaey and Hammond (2004).

But the literature assumes we **do**. See the masterful, patient survey paper: Bossert and Weymark (2004). I would suggest that, even if we have a firm grasp of each person's utility function, we do not **necessarily** think they are objectively comparable (and hence, for example, a Rawlsian maxmin solution might not be well-defined). For those who are unpersuaded, I offer an example.

Consider a society with two members. The first is Jane, a highly intelligent, deeply musical person who loves animals. The second is Jane's pet Dalmatian, Spot. A fan of chasing squirrels, he likes humans and other dogs, and loves food. We don't normally think we understand fully what it is like to be Jane, much less what it is like to be Spot, who is so cognitively distant. But suppose we are in the year 2525, and we can run a program that lets us understand exactly what it is like to be Jane, and to be Spot.[40] That is, we have perfect measurability. Does this mean we have comparability as well? At no time does Jane get as excited as Spot does when he is about to be fed his favorite dinner and races hysterically around the apartment accidentally knocking things over. On the other hand, he never experiences the transcendental feeling of being transported to another world that Jane does, when she listens to her favorite Bach Brandenburg Concerto. Would any two people agree about whether they would rather be Jane or Spot? Would Spot agree? What would Harsanyi say? I leave these questions as exercises for the reader.[41]

Suppose for a moment that we think we *do* find it meaningful to compare (even rank) the utilities of different members of society. For the sake of argument, let's say we have a social welfare functional F that maps any profile of absolute utility functions into a social ordering of alternatives. Now suppose that, although we find all of the above meaningful, we don't think we have the information we need to evaluate each person's absolute utility. Perhaps we can estimate each individual's marginal utility of income using her behavior under risk, as discussed earlier, but we have no way of calibrating utilities levels across people. What class of utility transformations should our social welfare functional permit, in determining equivalence classes of profiles of utility functions? I would say none, except the identity transform. Any transformation of any of the functions that are inputs to the social welfare functional moves things into a different region of the domain, where we might wish to make different tradeoffs. We need to retain the sensitivity of F, and then aggregate not just across individuals but across the different possible values of utilities that pertain in different states (which we can't identify); this aggregation uses our Bayesian beliefs and also our ethical judgements (Harsanyi and Rawls[42] would presumably not want to do it the same way).

Perhaps it would be useful to distinguish, in speaking of interpersonal comparability of utilities, between subjective and objective comparability.[43] Full or partial objective comparability would

[40] Apologies to Nagel (1974), who would probably be almost as skeptical about our ability to comprehend what it's like to be Spot as to know what it's like to be a bat. And what of my ability to compare how I feel to what Jane feels? Inspired by Hammond (1976) and Strasnick (1975), Arrow (1977) tentatively flirts with extended sympathy. But in conclusion, he humbly says (pg. 225), "In a way that I cannot articulate well and am none too sure about defending, the autonomy of individuals, an element of mutual incommensurability among people, seems denied by the possibility of such interpersonal comparisons. No doubt it is some such feeling as this that has made me so reluctant to shift from pure ordinalism, despite my desire to seek a basis for a theory of justice."

[41] Perhaps the "meaningful statements" framework of Bossert (1991) is a better way to address these questions than the admissible transformations approach.

[42] The 560 pages of Rawls (1971) contain much food for thought. But I find more enlightenment in the combined 15 pages of Harsanyi (1953, 1955).

[43] Here I use objective comparability in the context of a single observer's ability to compare utilities, whereas Roberts (1997) uses it in the context of "distilling conflicting interpersonal comparisons into a single set of interpersonal compar-

involve statements about utility comparisons with which any rational person should agree (analogous to statements like "The car that just overtook us was going faster than we were"), whereas subjective comparability would involve matters of opinion (analogous to statements such as "I like the car that just overtook us better than this one"). If a sympathetic observer violates IIA and makes her ranking of alternatives x and y dependent upon how many alternatives are between x and y in each citizen's ordering, for example, Arrow (1951) said that she must be making interpersonal comparisons of utilities (which at that time he considered meaningless). But she may not think that different citizens' utilities are objectively comparable. To clarify, let's consider a consumer choice example far removed from interpersonal comparisons and IIA. Suppose Pierre works in a supermarket. Less fruit is purchased during the week than expected, and employees are told they can each take a bag of apples or a bag of oranges home with them, without charge. Liking oranges better than apples, Pierre intends to take a bag of oranges. But a friendly coworker murmurs to him that she had tasted one of those apples and she loved it. So Pierre selects a bag of apples instead. Notice that Pierre doesn't think there is anything objective about preference for oranges (some people prefer apples, some prefer oranges), and the fact that his coworker's extra information influences Pierre in no way suggests to us that Pierre thinks the new information is *objective* or that apples and oranges are *objectively* comparable. Even if the sympathetic observer in the abstract social choice problem thinks different people's utilities are not objectively comparable, the extra information (that IIA would exclude) may reasonably change her rankings.

Harsanyi (1955) divides the problem of interpersonal comparisons into two. First, what can we say of two individuals who have identical preferences and verbal and nonverbal reactions to different allocations? This is the purest instance of what Harsanyi calls the metaphysical problem: might the two persons still have different susceptibilities to satisfaction, and therefore associate different utilities from each other to the same circumstance? Harsanyi (1955, pg. 217) says: "If two objects or human beings show similar behavior in *all* their relevant aspects open to observation, the assumption of some unobservable hidden difference between them must be regarded as a completely gratuitous hypothesis and one contrary to sound scientific method." On many matters Harsanyi convinces me, but not here. There is plenty of room for a Bayesian to say quite possibly the two are experientially identical, and plenty of room also to have weight on their being different in various ways. In any case, Harsanyi goes on to say in the real world, people typically differ in their preferences and reactions, and this poses a real challenge to comparing their utilities, admitting freely (pg. 320) that "... the practical need for reaching decisions on public policy will require us to formulate social welfare functions - explicitly or implicitly - even if we lack the factual information needed for placing interpersonal comparisons of utility on an objective basis." We need to guess.

Harsanyi (1953, 1955) has an extremely important thought exercise for us, connecting individual choice under risk to social choice and welfare.[44] One can think of many versions of it; consider two

isons." (pg. 79). This matter of terminology reminds me of an acknowledgement I always wanted to get in print. In 1982, when Doug Bernheim and I were working on rationalizable strategic behavior (Bernheim, 1984; Pearce, 1984), he was using the term "temporary equilibrium" and I was using the term "ex ante equilibrium". Doug told me that Kevin Roberts suggested that we both switch to the term rationalizability. I liked that suggestion, and we agreed that Doug would put a footnote in his paper explaining the origins of the terminology. But that got lost in editing and isn't in Doug's published version. Here, we thank Kevin Roberts belatedly for his suggestion.

[44] Harsanyi's ideas in this territory are widely discussed. The literature ranges from the critical view of Diamond (1967) to more appreciative analyses such as Weymark (1991) and Grant, Kajii, Polak, and Safra (2006, 2010). Section 7 of the

simplified ones that I will call Harsanyi plain and Harsanyi mystical. In the plain version, for any choice of how to organize economic resources in an N-person society, one could think of each person having an equal chance, ex ante (behind the veil of ignorance), of occupying any of the N positions in the income distribution. How would any particular person, Sara for example, like income to be distributed? In effect, we are asking her to choose what society she would like to be born into (in terms of income distribution). If Sara is a risk-averse expected utility maximizer, this will bias her in favor of a relatively egalitarian distribution of income (entirely egalitarian, if, unrealistically, all distributions of a fixed total income are available). This "plain" version of Harsanyi's question does not provide a definitive answer to the social choice problem, because if you ask Vicky instead of Sara, she will typically have a different von Neumann-Morgenstern utility index, and will choose differently (although again, if all distributions of a constant total are available, without incentive complications, Sara's and Vicky's choices will coincide if they are both risk averse to any degrees). Nonetheless, given the prevalence of risk aversion in real populations, this idea lends enormous support to the idea of substantial redistribution of income. Being able to buy insurance against disasters in health, driving, floods, fire and so on is immensely important to most of us. There is a huge missing market, in that we cannot buy insurance against being born into extreme poverty. Harsanyi's thought experiment illustrates that powerfully.

For the mystical version of Harsanyi, Sara imagines, behind the veil of ignorance, that she is equally likely to *be* any of the individuals in society. If she turns out to be Greta, she takes on Greta's predilection for adventure and her curious taste for awkward situations. If she instead turns out to be Harry, she will have his sense of humor, shyness, and love of artistic men. Now, what society would she like to be born into? This is a deep question. First, it presupposes that she understands what it would like to be Greta, with different disposable incomes or different public goods. Let's stipulate that she does. What does it *mean* for Sara to prefer a baseball stadium to a hospital, which Greta (and some others) would prefer, and Harry (and still others) would not? How does she trade off their utilities? And would Vicky trade them off the same way? Harsanyi (1955, pg. 20) says "... the more complete our factual information and the more completely individualistic our ethics, the more the different individuals' social welfare functions will converge toward the same objective quantity, namely, the unweighted sum (or rather the unweighted arithmetic mean) of all individual utilities." This ex ante similarity is part of what is sometimes informally called the Harsanyi Doctrine. If, indeed, Sara and Greta will make the same choice about what society to be born into, this comes close to solving the problem of social choice: unlike Harsanyi plain, the mystical cousin has everyone agreeing about how to organize society. Seductive as this is, I can't convince myself of its validity. It seems to conjure a solution out of nowhere, based on some universal preference I don't know how to derive. But even Harsanyi plain is a huge eye-opener.

What Harsanyi calls the metaphysical problem of trying to understand how someone else experiences a particular social allocation is akin to uncertainty John may feel about whether Jake sees the color blue the same way he does. In my graduate school days I somehow got the impression that the latter question was first studied by the English philosopher George Berkeley. But our ancient Greek friend Gorgias got there long before! "How can anyone communicate the idea of color by means of words when the ear does not hear colors but only sounds?" (See McComiskey, 1997) Beautifully

2006 working paper version has a nice treatment of the relationship of their results to comparable welfare).

posed, and rather devastating. *This* is what we are up against when we confront the *practical* problem of actually conceiving of what someone else is experiencing. Will neuroscience eventually "map the brain" for us and will that, subject to a metaphysical leap of faith, solve this mystery? I don't know. In the meantime, we need to have more modest goals, and in keeping with my Bayesian theme, that involves guessing.

The facts that societies are fairly cohesive, a lot of people like food, sex, praise and company, and many of us can make sense of the same movie, suggest that there may be a lot of commonalities in perceptions across individuals. Still, when John and Jake disagree (one likes blue, one doesn't), we have to wonder if blue looks similar to them but one just doesn't like that look, or if instead it actually has a different appearance. There are some cases where we *don't* have to wonder: we know there is a difference in perception. A well-known example is the fact that a sizeable minority of people perceive cilantro to have a soapy taste and smell, whereas the majority don't agree. This would not be possible if everyone perceived cilantro the same way: either everyone would then agree it was soapy, or everyone would agree it wasn't. In fact, those who find it soapy are likely to have an SNP (single nucleotide polymorphism), namely rs729210001, on the OR6A2 gene, in the olfactory-receptor region (see Eriksson, Wu, Do, Kiefer, Tung, Mountain, Hinds, and Francke, 2012). So mapping the brain is related to mapping the human genome.

For the purposes of twenty-first century social choice, we will need to proceed at a much less detailed level. How might we achieve a rough calibration of satisfaction levels across individuals? A brutal start might be to say that we can identify the utility 0 for any person with an indifference between life and death: if life is scarcely worth living, your utility is 0. This is immediately complicated by differences in perception about experience after death: some expect heaven, some hell, and others a great nothingness. Moreover, willingness to commit suicide is not a good calibration, as some people feel it is immoral and will resist even if life is a true misery, while others won't. Conversational communication surely helps with calibration, and again, can't be perfect. How about different persons' ranges of experience? Is Sara's maximum welfare (experienced under her favorite allocation) the same as Vicky's? Again, verbal communication helps here, but differences in expressiveness and use of language still leave us guessing. If both of them are expected utility maximizers, and if we interpret those indices as expressions of introspective welfare, in principle each can convey to the other how her utility depends on the social allocation, up to the usual two degrees of freedom associated with representation of von Neumann-Morgenstern or Savage utility indices. The tremendous dimensional reduction associated with those indices is lost if Sara and Vicky are *not* expected utility maximizers. Either way, it is fascinating to think of how they might attempt, conversationally, to communicate information beyond what can be revealed by observable choice. Sara characterizes what something feels like to her, knowing that Vicky is likely misunderstanding what she means. Vicky, aware that she is likely misunderstanding, tries to check on what Sara meant, knowing Sara won't necessarily interpret the content of Vicky's question as Vicky intended. I would be interested in knowing what epistemic or linguistic models might tackle this communicational process, wherein the conversationalists don't share a commonly understood state space. For an interesting step down this path, see Chauvin (2020).

There are many aspects of welfare that are beyond the scope of this paper. I have not discussed liberty, equity, justice or capabilities, but these receive rich discussion in Sen (2017) and the copious references therein. There are topics at the intersection of economics and psychology[45] that are highly relevant. For example, can a well-informed act of altruism of person i toward person j leave i with lower welfare?[46] Does an intertemporally fragmented conception of individual welfare, such as the pathbreaking paper of Strotz (1955) make sense, or can we adequately model temptation and self-control *without* time inconsistency, as in the elegant decision theoretic treatment by Gul and Pesendorfer (2001)? And consider a final example: happiness studies, the subject of the 2005 Frisch Memorial Lecture by Daniel McFadden.[47] In an erudite essay, McCloskey (2012) criticizes that literature for its tendency to overreach and oversimplify. Armed with her skepticism, Bayesians can cautiously take into account some of the questions that happiness research raises, as we gather scraps of evidence about individual and social welfare and continue to guess. More generally, economists should stay engaged with other disciplines in the borderlands of our subject. We have a lot to learn, and a lot to contribute.

6 Developing Our Thoughts about Welfare

Having been critical about the 1950s foundations of social choice and welfare and the lack of consensus in the profession to this day, I could fairly be asked how I think these disciplines should be done. The novelist Somerset Maugham was sometimes asked for tips about how to write a novel. He liked to say: "There are three rules for writing a novel. Unfortunately, no one knows what they are." His three rules remind me of Gorgias' three principals of nonexistence (see Section 2). Apparently, they don't actually exist (in keeping with Gorgias' first principle). On closer inspection, Maugham allows that they may exist, but no one knows what they are (Gorgias' second principle comes to mind). But Maugham's own mastery of the form suggests that he knew the three rules, but he couldn't communicate them.

As elusive as the great novel, convincing social choice and welfare theory have to face the profound difficulties of observing another person's welfare and of understanding others' welfare in comparative terms. I don't do social choice or welfare theory and I certainly can't give you three rules for how to do them well. But given the esteem in which I hold those subjects, I will try to enunciate some things I would like to see happen in the literature.

For the purposes of welfare economics, I hope the profession at large can say **farewell to meaninglessness**. In this realm, meaninglessness has messed us up for eighty years.[48] I would argue that the attempt to build welfare on a foundation which, to me, is incoherent (see the discussion in Sec-

[45] In his famous tract on economic methodology, Robbins (1932, pp. 83-84) opined: "The borderlands of Economics are the happy hunting ground of the charlatan and the quack... ." As one of those quacks, I should feel insulted. But the charms of the cadences and the condescension are too delightful.

[46] In my first year in the Princeton economics doctoral program I showed Greg Mankiw, then a precocious senior doing PhD coursework, a note I had written on nonpaternalistic sympathy, where each generation cares about the *utility* (not just the consumption) of future generations. He said it almost seemed as if double-counting of utility were going on (my sacrifice of consumption for that of my daughter yields utility for her, and also for me, leaving both of us with higher utility). I wasn't sure I agreed, but I was certainly sympathetic to the question. The note grew into the final essay of my dissertation Pearce (1983). In the same territory, see Ray (1987), Galperti and Strulovici (2017) and Vasquez and Weretka (2019).

[47] For a taste of this large literature, see Kahneman, Diener, and Schwarz (1999) and Stevenson and Wolfers (2013).

[48] Let's not extend this to one hundred years of solipsism.

tion 2 and in footnote 38 of Section 5), inevitably led to much confusion in the ensuing debates over the decades. Fleurbaey and Mongin (2005) make it plain that often, the participants were talking entirely at cross purposes. Most remarkably, after a quarter of a century of discussion even Arrow and Samuelson proved incapable of understanding each other on these subjects: see the well-chosen epigraph of Igersheim (2019):

> My only point in writing is that I never find myself in disagreement with you on matters of logic. Sometimes my taste for, say, bounded utility, differs from yours; but I understand and respect your view. Your position on SWF baffles me.
>
> Samuelson to Arrow, April 30, 1976.

Could we all join Lehtinen (2007) and **say farewell to** IIA? The decision-theoretic perspective I proposed in Section 3 militates against imposing IIA in Arrow's World (where only the preference profile is reported to the observer). If instead the observer is given what she feels is fully informative "absolute" information about citizens' utilities (see Section 5), all she needs to do a "welfarist" comparison of alternatives x and y is to use the absolute utilities of each person under each of those two alternatives. In that extreme case, she will happily ignore information about other alternatives when ranking x and y. But otherwise, when information is not at such an unlikely polar ideal, she might feel that as a Bayesian she learns something about absolute utilities by looking at information involving other alternatives, and there is no clear reason to disallow it.[49] With meaninglessness and IIA no longer looming over it, could the profession at large embrace the view long held by a large part of the social choice and welfare community, that the social choice problem is underdetermined, not overdetermined (see Section 4), and there is ample scope for important and adventurous work to be done outside of Arrow's original strictures?

Many in the social choice literature have followed Sen (1970) in exploring the implications for social choice of different degrees of measurability and comparability of utilities (again, see the references in Fleurbaey and Hammond (2004) and in Sen (2017)). I value this work greatly. A further step is required: we never actually *have* this kind of partial information. We are never *sure* of utility levels, utility differences, difference ratios, and so on. Finally, then, we must **say farewell to certainty**. We need to embrace a messy Bayesian welfare economics, where evidence is culled from many quarters (the market, surveys, experiments, neuroscience and so on), and then careful consideration leads to a guess (a probability distribution, not a point estimate).[50]

If all the culling of information and intelligent guessing results in a probability distribution, where does said distribution live? What space does it inhabit? In the most general terms, we are guessing about the state of the world. But what do we write down for the state space? An element of the state space describes what each allocation would be like for each citizen. Again, what space is *that* in? Because information beyond a preference ordering was deemed meaningless, the profession

[49] In the same way as many writers raise doubts about the comparability of the utility of different individuals, one can ask if the utilities of one individual can be compared across states, as these states are different ways the world, in fact the individuals, might have been. We could expect that 1950s Arrow and Samuelson, as "comparability conservatives", might have said "no, such cross-state comparisons are meaningless". And then on what authority did they impose IIA, when an individual's preference for x over y may mean an entirely different thing in one state from another?

[50] Perhaps the closest thing we have to this culling and synthesizing, albeit in a relatively simple space, is cost-benefit analysis (see, for example, Little and Mirrlees (1974), Brent (2011) and Boadway (2016).

was discouraged from exploring the nature of such information, and from developing a language to describe or summarize it. I hope that future researchers will **develop a rich language to deal with information beyond a preference profile**.

When I expressed that hope in my talk to the Econometric Society World Congress in August 2020, some of my intellectual confidants expressed interest in this language of which I spoke, and said they were not entirely sure what it would *be*. That largely summarizes my own situation. At the less complicated end of things, it should make it easy to speak of interpersonal comparability or cardinality or measurability or actual utility and so on, without needing to introduce thermometers or Dalmatians or Siri to clarify one's intended meaning. More ambitiously, it might assist two people in organizing their attempts to share the nature of their welfare in different circumstances, as mentioned toward the end of Section 5. (That might include considerations of satisfaction or fulfilment or altruism as distinct from more hedonistic senses of happiness.) Professionally, it might help social scientists develop a distributional sense of how often people are "far" from the median in terms of how sensitive they are to circumstances, how vulnerable they are to suffering and what is their capacity for joy. That in turn could help economists favoring opposing economic policies discover whether their disagreement comes more from differing value judgements or from differing opinions about sensitivity dispersion and so on. Frank Knight said that "Values are established or validated or recognized through discussion...". (Knight, 1947). We don't have an ideal language for carrying on such discussions. I would like future Joan Robinsons and Paul Samuelsons to be able to debate not just capital theory and reswitching, but redistributional policies and social welfare, with as rich a language as we can develop for them.

What would Ragnar Frisch think of all this discussion? I would not be surprised if he took exception to much of what I have said. But at least I feel he would be pleased that we were focusing on *human welfare*. In his Nobel Prize Lecture, Frisch is as expansive as ever, discussing not just econometrics but antimatter, metagalaxies, chaos, evolution, and even our own influence on evolution. Reading it reveals the scope of his conception of economic science. I extract one brief passage from it, because it gives a taste of the *humanity* of that conception:

> "Understanding is not enough, you must have compassion... . I cannot be happy if I can't believe that in the end the results of our endeavours may be utilized in some way for the betterment of the little man's fate."
>
> Ragnar Frisch (1970, pg. 15)

References

Abdellaouia, M., C. Barrios, and P. P. Wakker (2007): "Reconciling introspective utility with revealed preference: Experimental arguments based on prospect theory," *Journal of Econometrics*, 138(1), 356–378, 50th Anniversary Econometric Institute.

Anscombe, F. J., and R. J. Aumann (1963): "A Definition of Subjective Probability," *Ann. Math. Statist.*, 34(1), 199–205.

Argenziano, R., and I. Gilboa (2019): "Perception-theoretic Foundations of Weighted Utilitarianism," *The Economic Journal*, 129(620), 1511–1528.

Arrow, K. J. (1950): "A Difficulty in the Concept of Social Welfare," *Journal of Political Economy*, 58(4), 328–346.

——— (1951): *Social Choice and Individual Values.* Cowles Comission Monograph No. 12, 1 edn.

——— (1963): *Social Choice and Individual Values.* Cowles Comission Monograph No. 12, 2 edn.

——— (1967): "Public and Private Values," in *Human Values and Economic Policy*, ed. by S. Hook. New York University Press, New York.

——— (1977): "Extended Sympathy and the Possibility of Social Choice," *The American Economic Review*, 67(1), 219–225.

——— (1983): "Contributions to Welfare Economics," in *Paul Samuelson and Modern Economic Theory*, ed. by E. C. Brown, and R. M. Solow, pp. 15–30. McGraw-Hill, New York.

——— (1984): "The Principle of Rationality in Collective Decisions," in *Collected Papers of Kenneth J. Arrow*, vol. 1: Social Choice and Justice. Harvard University Press.

Atkinson, A. B. (2009): "Economics as a Moral Science," *Economica*, 76(s1), 791–804.

Ayer, A. J. (1978): Interviewed by Bryan Magee on his television series *Men of Ideas.* British Broadcasting Corporation.

Bailey, M. J. (1979): "The Possibility of Rational Social Choice in an Economy," *Journal of Political Economy*, 87(1), 37–56.

Bergson, A. (1938): "A Reformulation of Certain Aspects of Welfare Economics," *The Quarterly Journal of Economics*, 52(2), 310–334.

Bernheim, B. D. (1984): "Rationalizable Strategic Behavior," *Econometrica*, 52(4), 1007–1028.

Bernheim, B. D., and A. Rangel (2007): "Toward Choice-Theoretic Foundations for Behavioral Welfare Economics," *The American Economic Review*, 97(2), 464–470.

Bernstein, R. J. (1983): *Beyond Objectivism and Relativism: Science, Hermeneutics, and Praxis.* University of Pennsylvania Press, Philadelphia.

Bjerkholt, O. (2008): "Ragnar Frisch on Scientific Economics," *working paper, University of Oslo.*

Bjerkholt, O., and D. Qin (2010): "Teaching Economics as a Science: the 1930 Yale Lectures of Ragnar Frisch," *Memorandum 05/2010, Oslo University, Department of Economics.*

Boadway, R. (2016): "Cost-Benefit Analysis," in *The Oxford Handbook of Well-being and Public Policy*, ed. by M. Adler, and M. Fleurbaey. Oxford Handbooks Online.

Bossert, W. (1991): "On Intra- and Interpersonal Utility Comparisons," *Social Choice and Welfare*, 8, 207—-219.

Bossert, W., and J. Weymark (2004): "Utility in Social Choice," in *Handbook of Utility vol 2*, ed. by S. Barbera, P. Hammond, and C. Seidl, chap. 20, p. 1122. Springer New York, Kluwer Boston.

Brent, R. (2011): *Handbook of Cost-Benefit Analysis.* Edward Elgar.

Chauvin, K. (2020): "Unacknowledged Heterogeneity in Communication," *Dept. of Economics, Harvard.*

Coakley, M. (2016): "Interpersonal Comparisons of the Good: Epistemic not Impossible," *Utilitas*, 3, 288–313.

Cowles, A. (1960): "Ragnar Frisch and the Founding of the Econometric Society," *Econometrica*, 28(2), 173–174.

D'Aspremont, C., and L. Gevers (1977): "Equity and the Informational Basis of Collective Choice," *The Review of Economic Studies*, 44(2), 199–209.

de Scitovsky, T. (1941): "A Note on Welfare Propositions in Economics," *The Review of Economic Studies*, 9(1), 77–88.

Diamond, P. A. (1967): "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparison of Utility: Comment," *Journal of Political Economy*, 75(5), 765–766.

Edgeworth, F. Y. (1881): *Mathematical Psychics.* C. Kegan Paul & Co., London.

Eriksson, N., S. Wu, C. B. Do, A. K. Kiefer, J. Y. Tung, J. L. Mountain, D. A. Hinds, and U. Francke (2012): "A Genetic Variant Near Olfactory Receptor Genes Influences Cilantro Preference," *BMC Flavour*, 1(22).

Feldman, A. M., and R. Serrano (2008): "Arrow's Impossibility Theorem: Two Simple Single-Profile Versions," *Harvard College Mathematics Review*, 2(2), 46–57.

Fishburn, P. (1970): *Utility Theory for Decision Making.* Wiley.

Fleurbaey, M., and P. Hammond (2004): "Interpersonally Comparable Utility," in *Handbook of Utility vol 2*, ed. by S. Barbera, P. Hammond, and C. Seidl. Springer New York, Kluwer Boston.

Fleurbaey, M., and F. Maniquet (2008): "Utilitarianism versus Fairness in Welfare Economics," in *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, ed. by M. Fleurbaey, M. Salles, and J. A. Weymark, pp. 263–280. Cambridge University Press, Cambridge.

Fleurbaey, M., and P. Mongin (2005): "The News of the Death of Welfare Economics is Greatly Exaggerated," *Soc Choice Welfare*, 25(4), 381–418.

Friedman, M. (1955): "What All is Utility?," *The Economic Journal*, 65(259), 405–409.

Frisch, R. (1926a): "Kvantitativ Formulering av den Teoretiske Økonomikks Lover," *Statsøkonomisk Tidsskrift*, 40, 299–334.

——— (1926b): "Sur Un Problème d'Économie Pure," *Norsk Mathematisk Forenings Skrifter*, 1, 1–40.

——— (1964): "Dynamic Utility," *Econometrica*, 32(3), 418–424.

——— (1970): "From Utopian Theory to Practical Applications: The Case of Econometrics," *Lecture to the Memory of Alfred Nobel*, June 17.

Fryxell, L. (2019): "A Theory of Experienced Utility and Utilitarianism," *Northwestern University.*

Gaertner, W. (2002): "Domain restrictions," in *Handbook of Social Choice and Welfare*, ed. by K. J. Arrow, A. K. Sen, and K. Suzumura, vol. 1 of *Handbook of Social Choice and Welfare*, chap. 3, pp. 131–170. Elsevier.

Galperti, S., and B. Strulovici (2017): "A Theory of Intergenerational Altruism," *Econometrica*, 85(4), 1175–1218.

Gibbard, A. F. (1968/2014): "Social Choice and the Arrow Conditions," *Economics and Philosophy*, 30(3), 269–284, Printed from an unpublished manuscript written in 1968.

Gilboa, I. (2009): *Theory of Decision under Uncertainty.* Econometric Society Monograph, Cambridge University Press.

——— (2014): *On the Interpretation of Probabilities.* November 27, 2014 posting on the Decision Theory Forum.

Goodman, L. A., and H. Markowitz (1952): "Social Welfare Functions Based on Individual Rankings," *American Journal of Sociology*, 58(3), 257–262.

Grant, S., A. Kajii, B. Polak, and Z. Safra (2006): "Generalized Utilitarianism and Harsanyi's Impartial Observer Theorem," *Cowles Foundation Discussion Paper no. 1578, Yale University.*

——— (2010): "Generalized Utilitarianism and Harsanyi's Impartial Observer Theorem," *Econometrica*, 78(6), 1939–1971.

Gul, F., and W. Pesendorfer (2001): "Temptation and Self-Control," *Econometrica*, 69(6), 1403–1435.

——— (2007): "Welfare without Happiness," *American Economic Review*, 97(2), 471–476.

Hammond, P. J. (1976): "Equity, Arrow's Conditions, and Rawls' Difference Principle," *Econometrica*, 44(4), 793–804.

——— (1991): "Independence of irrelevant interpersonal comparisons," *Social Choice and Welfare*, 8(1), 1–19.

Hansson, B. (1973): "The independence condition in the theory of social choice," *Theor Decis*, 4, 25–49.

Harsanyi, J. C. (1953): "Cardinal Utility in Welfare Economics and in the Theory of Risk-taking," *Journal of Political Economy*, 61(5), 434–435.

——— (1955): "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," *Journal of Political Economy*, 63(4), 309–321.

Hicks, J. R. (1939): "The Foundations of Welfare Economics," *The Economic Journal*, 49(196), 696–712.

Hildreth, C. (1953): "Alternative Conditions for Social Orderings," *Econometrica*, 21(1), 81–94.

Igersheim, H. (2019): "The Death of Welfare Economics: History of a Controversy," *History of Political Economy*, 51(5), 827–865.

Kahneman, D., E. Diener, and N. Schwarz (1999): *Well-Being: Foundations of Hedonic Psychology*. Russell Sage Foundation.

Kahneman, D., P. P. Wakker, and R. Sarin (1997): "Back to Bentham? Explorations of Experienced Utility," *The Quarterly Journal of Economics*, 112(2), 375–405.

Kaldor, N. (1939): "Welfare Propositions of Economics and Interpersonal Comparisons of Utility," *The Economic Journal*, 49(195), 549–552.

Kemp, M. C., and Y.-K. Ng (1976): "On the Existence of Social Welfare Functions, Social Orderings and Social Decision Functions," *Economica*, 43(169), 59–66.

Knight, F. (1947): *Freedom and Reform: Essays in Economic and Social Philosophy*. Allan and Unwin.

Kreps, D. M. (1988): *Notes on the Theory of Choice*. Westview Press.

——— (2013): *Microeconomic Foundations I*. Princeton University Press.

Le Breton, M., and J. Weymark (1996): "An Introduction to Arrovian Social Welfare Functions on Economic and Political Domains," in *Collective Decision-Making: Social Choice and Political Economy*, ed. by N. Schofield, pp. 25–61. Springer Netherlands, Dordrecht.

Lehtinen, A. (2007): "A Farewell to IIA," *working paper, University of Helsinki.*

Little, I., and J. Mirrlees (1974): *Project Appraisal and Planning for Developing Countries.* Heinemann, London.

Machina, M. J. (1981): ""Rational" Decision Making versus "Rational" Decision Modelling?," *Journal of Mathematical Psychology*, 13(24), 163–175.

Mankiw, D. G. (2015): *Principles of Economics*. Cengage Learning, 7th edn.

Mas-Colell, A., M. D. Whinston, and J. R. Green (1995): *Microeconomic Theory*. Oxford University Press, Oxford.

MASKIN, E. (2014): "The Arrow Impossibility Theorem: Where Do We go from Here?," in *The Arrow Impossibility Theorem*, ed. by E. Maskin, and A. Sen, pp. 43–55. Columbia University Press, New York.

——— (2020): "A modified version of Arrow's IIA condition," *Social Choice and Welfare*, 54, 203–209.

MAYSTON, D. J. (1974): *The Idea of Social Choice*. Macmillan, London.

MCCLOSKEY, D. (2012): "Happyism: The Creepy New Economics of Pleasure," *The New Republic*.

MCCOMISKEY, B. (1997): "Gorgias, "On Non-Existence": Sextus Empiricus, "Against the Logicians" 1.65-87, Translated from the Greek Text in Hermann Diels's "Die Fragmente der Vorsokratiker"," *Philosophy & Rhetoric*, 30(1), 45–49.

NAGEL, T. (1974): "What Is It Like to Be a Bat?," *The Philosophical Review*, 83(4), 435–450.

NASH, J. F. (1950): "The Bargaining Problem," *Econometrica*, 18(2), 155–162.

——— (1953): "Two-Person Cooperative Games," *Econometrica*, 21(1), 128–140.

NG, Y.-K. (1975): "Bentham or Bergson? Finite Sensibility, Utility Functions and Social Welfare Functions," *The Review of Economic Studies*, 42(4), 545–569.

PARKS, R. P. (1976): "An Impossibility Theorem for Fixed Preferences: A Dictatorial Bergson-Samuelson Welfare Function," *The Review of Economic Studies*, 43(3), 447–450.

PAZNER, E. (1979): "Equity, Nonfeasible Alternatives and Social Choice: A Reconsideration of the Concept of Social Welfare," in *Aggregation and Revelation of Preferences*, ed. by J. J. Laffont. North-Holland, Amsterdam.

PEARCE, D. G. (1983): "Nonpaternalistic Sympathy and the Inefficiency of Consistent Intertemporal Plans," *Final Chapter of Princeton Doctoral Dissertation*, reprinted in Foundations in Economic Theory: A Volume in Honor of Hugo F. Sonnenschein, Ed. M. Jackson and A. McClennan, Springer, 2008.

——— (1984): "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica*, 52(4), 1029–1050.

——— (1995): "Arrow's Theorem on its Head: A Bayesian Perspective on Social Choice," *Yale University.*

POLLAK, R. A. (1979): "Bergson-Samuelson Social Welfare Functions and the Theory of Social Choice," *The Quarterly Journal of Economics*, 93(1), 73–90.

RAWLS, J. (1971): *A Theory of Justice*. Harvard University Press.

RAY, D. (1987): "Nonpaternalistic intergenerational altruism," *Journal of Economic Theory*, 41(1), 112 –132.

ROBBINS, L. (1932): *An Essay on the Nature and Significance of Economic Science.* Macmillan, London.

ROBERTS, K. W. S. (1980a): "Interpersonal Comparability and Social Choice Theory," *The Review of Economic Studies*, 47(2), 421–439.

——— (1980b): "Social Choice Theory: The Single-Profile and Multi-Profile Approaches," *The Review of Economic Studies*, 47(2), 441–450.

——— (1997): "Objective interpersonal comparisons of utility," *Social Choice and Welfare*, 14(1), 79–96.

ROBERTSON, D. H. (1951): "Utility and All That," *The Manchester School*, 19(2), 111–142.

ROTHENBERG, J. (1961): *The Measurement of Social Welfare.* Prentice-Hall.

RUBINSTEIN, A. (1984): "The Single Profile Analogues to Multi Profile Theorems: Mathematical Logic's Approach," *International Economic Review*, 25(3), 719–730.

——— (2007): "Instinctive and Cognitive Reasoning: A Study of Response Times," *The Economic Journal*, 117(523), 1243–1259.

SAARI, D. G. (1995): "Inner Consistency or Not Inner Consistency: A Reformulation is the answer," in *Social Choice, Welfare, and Ethics: Proceedings of the Eighth International Symposium in Economic Theory and Econometrics*, ed. by W. A. Barnett, H. Moulin, M. Salles, and N. J. Schofield, pp. 187–212. Cambridge University Press.

SAMUELSON, P. A. (1937): "A Note on Measurement of Utility," *The Review of Economic Studies*, 4(2), 155–161.

——— (1938): "The Numerical Representation of Ordered Classifications and the Concept of Utility," *The Review of Economic Studies*, 6(1), 65–70.

——— (1947): *Foundations of Economic Analysis.* Harvard University Press.

——— (1952): "Probability, Utility, and the Independence Axiom," *Econometrica*, 4(20), 670–678.

——— (1963): "D. H. Robertson," *Quarterly Journal of Economics*, 77(4), 517–536.

——— (1967): "Arrow's Mathematical Politics," in *Human Values and Economic Policy*, ed. by S. Hook. New York University Press, New York.

——— (1977): "Reaffirming the Existence of "Reasonable" Bergson-Samuelson Social Welfare Functions," *Economica*, 44(173), 81–88.

——— (1981): "Bergsonian Welfare Economics," in *Economic Welfare and the Economics of Soviet Socialism: Essays in Honor of Abram Bergson*, ed. by S. Rosefielde, pp. 223–266. Cambridge University Press, Cambridge.

——— (1987): "Sparks from Arrow's Anvil," in *Arrow and the Foundations of the Theory of Economic Policy*, ed. by G. R. Feiwel. Palgrave Macmillan UK.

Savage, L. J. (1954): *The Foundations of Statistics*. Dover.

Schotter, A. (2003): "Decision Making with Naive Advice," *American Economic Review*, 93(2), 196–201.

Sen, A. (1970): *Collective Choice and Social Welfare*. Holden-Day.

——— (1987): "Social Choice," in *The New Palgrave*, ed. by J. Eatwell, M. Milgate, and Newman, vol. 4, pp. 382–393. MacMillan, London.

——— (2011): "The Informational Basis of Social Choice," in *Handbook of Social Choice and Welfare*, ed. by K. J. Arrow, A. Sen, and K. Suzumura, vol. 2 of *Handbook of Social Choice and Welfare*, chap. 14, pp. 29–46. Elsevier.

——— (2017): *Collective Choice and Social Welfare*. Harvard University Press, Cambridge, Massachusetts, an expanded edition edn.

Stevenson, B., and J. Wolfers (2013): "Subjective Well-Being and Income: Is There Any Evidence of Satiation?," *American Economic Review*, 103(3), 598–604.

Strasnick, S. L. (1975): "Preference Priority and the Maximization of Social Welfare," *Doctoral Dissertation, Harvard University.*

Strotz, R. H. (1955): "Myopia and Inconsistency in Dynamic Utility Maximization," *The Review of Economic Studies*, 23(3), 165–180.

Vasquez, J., and M. Weretka (2019): "Mutual Empathy in Games," *working paper, University of Wisconsin.*

Vickrey, W. (1945): "Measuring Marginal Utility by Reactions to Risk," *Econometrica*, 13(4), 319–333.

Weymark, J. (1991): "A Reconsideration of the Harsanyi-Sen Debate on Utilitarianism," in *Interpersonal Comparisons of Well-being*, ed. by J. Elster, and J. Roemer. Cambridge University Press.

Yeatman, R. J., and W. C. Sellar (1930): *1066 and All That: A Memorable History of England.* Methuan Publishing.

Young, P. (1976): "Optimal Voting Rules," *Journal of Economic Perspectives*, 9(1), 51–64.