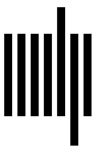
The Handbook of Rationality

Edited by: Markus Knauff, Wolfgang Spohn

Citation:

The Handbook of Rationality
Edited by: Markus Knauff, Wolfgang Spohn
DOI: 10.7551/mitpress/11252.001.0001
ISBN (electronic): 9780262366175
Publisher: The MIT Press
Published: 2021

Funding for the open access edition was provided by the MIT Libraries Open Monograph Fund.



The MIT Press

8.1 Preferences and Utility Functions

Till Grüne-Yanoff

Summary

This chapter discusses the notions of preference and utility in relation to individual judgment and decision making. The first two sections sketch the formal properties of preferences—in particular, completeness and transitivity—and of utility functions—in particular, continuity, differentiability, diminishing marginal utility, and risk attitudes. Whether these properties should be considered as conditions of practical rationality is also discussed in section 2. The third section focuses on the dominant interpretations of the formal preference and utility notions, in particular, distinguishing between mental and behavioral interpretations. The fourth section discusses relations between total and partial preferences, and the last section presents some models of rational preference change.

1. Preference Relations

The most common way to represent preferences is as a binary relation between two relata. This is how preference is typically conceptualized both in the social sciences and in philosophy. The properties that this relation is assumed to have are commonly described in a formalized language.[1] The preferences studied in preference logic are usually the preferences of rational individuals, but preference logic is also used in psychology and behavioral economics, where the emphasis is on actual preferences as revealed in behavior.

1.1 Concepts and Notation

The basic formal preference concept consists of a binary relation $\succcurlyeq$ on a set of alternatives $\boldsymbol{A}$. This relation is most commonly used to denote the objectives of a decision maker as a *comparative evaluation*. That is, $x \succcurlyeq y$ compares two members x and y of $\boldsymbol{A}$, describing x to be *at least as good as* y. This comparison might represent subjective desires of an individual but also her evaluative judgments. Furthermore, it might also represent evaluative judgments of social agents (e.g., groups, companies, government institutions) or intersubjective evaluations that do not coincide with the subjective attitudes of any particular agent (e.g., certain types of moral judgments). Finally, preference relations are also used to represent choice patterns.

The logical properties of this comparative relation do not seem to differ between the cases where they correspond to what we usually call "preferences" and the cases where they do not. The term "preference logic" is therefore standardly used to analyze the basic properties of these relations irrespective of their interpretations.

From the relation $\succcurlyeq$, two further relations on $\boldsymbol{A}$ can be derived. The *strict preference* relation $\succ$ is defined by

$$x \succ y \quad \text{if and only if} \quad x \succcurlyeq y \text{ and not } y \succcurlyeq x \tag{1}$$

and reads as "x is strictly preferred to y," "x is better than y," or "y is worse than x." The *indifference* relation is defined by

$$x \sim y \quad \text{if and only if} \quad x \succcurlyeq y \text{ and } y \succcurlyeq x \tag{2}$$

and reads as "x is indifferent to y" or "x is equal in value to y" (von Wright, 1963).

The above-described relations relate members of $\boldsymbol{A}$, the set of alternatives. Most broadly, one can characterize $\boldsymbol{A}$ as a set of propositions (Jeffrey, 1965/1983). Economists typically consider $\boldsymbol{A}$ to consist of vectors of consumption goods (Debreu, 1959). In most applications, members of $\boldsymbol{A}$ are assumed to be mutually exclusive (i.e., none of them is compatible with, or included in, any of the others). Preferences over a set of mutually exclusive relata are referred to as *exclusionary preferences* (S. O. Hansson, 2001).

The properties discussed in this section all concern such exclusionary preferences. In practice, people also have preferences between relata that are *not* mutually exclusive. These are called *combinative preferences* (S. O. Hansson, 2001).

An important kind of preferences are *preferences over certain outcomes*, where the relata are conjunctions of propositions that are interpreted as describing specific states of the world. Another kind are *preferences over lotteries*, where

the relata are mutually exclusive and jointly exhaustive disjuncts of propositions that are interpreted as alternative states of the world, each of which might be realized with a certain probability. Many relata of actual preferences do not satisfy either of these two conditions—in particular, the relevant disjuncts might be only partly identified, thus making it difficult to describe the relata as well-formed formulas. Such relata feature in decision making under *deep uncertainty* (S. O. Hansson, 1996).

1.2 Completeness

In most applications of preference logic, it is taken for granted that the following property, called "completeness" or "connectedness," should be satisfied:

$$x \succcurlyeq y \vee y \succcurlyeq x, \quad \text{for all } x, y \in \boldsymbol{A}, \tag{3}$$

that is, $\succcurlyeq$ is assumed to connect every member of $\boldsymbol{A}$ to every other member.

Completeness is assumed in many applications, not least in economics. Bayesian decision theory is a case in point. The Bayesian decision maker is assumed to make her choices in accordance with a complete preference ordering over the available options (Savage, 1954/1972, p. 18). However, in many everyday cases, we do not have, and do not need, complete preferences. Consider a person who has to choose between three objects *a*, *b*, and *c*. If she knows that she prefers *a* to the others, she does not have to make up her mind about the relative ranking among *b* and *c*. In such cases, preference *incompleteness* might be rationally justifiable.

In many practical contexts, people might exhibit incomplete preferences, even if not rationally justified. This raises the question how incompleteness can be resolved, which might happen in different ways. First, if the reason for preference incompleteness is lack of knowledge or reflection of one's own desires or judgments, then through observation, introspection, logical inference, or some other means of discovery, the missing preference relations might be discovered. In that case, the incompleteness is resolved in exactly one way, by arriving at exactly one complete preference relation. Second, incompleteness may be resolvable in several different ways. In this case, one's own desires or judgments on these issues are genuinely undetermined, so that the missing preferences cannot be discovered but must be *constructed* (Slovic, 1995). Finally, incompleteness may be irresolvable. Even with all the necessary resources available for preference discovery or construction, a person might be unable to say which she prefers. This last case of preference incompleteness is often also discussed as "preference *incommensurability*" (Chang, 1997).

1.3 Transitivity

By far the most discussed logical property of preferences is *transitivity*:

$$x \succcurlyeq y \wedge y \succcurlyeq z \rightarrow x \succcurlyeq z, \quad \text{for all } x, y, z \in X. \tag{4}$$

That is, if *x* is weakly preferred to *y*, and *y* weakly preferred to *z*, then *x* is weakly preferred to *z*. Experiments suggest that *in*transitive preferences are quite common (e.g., Tversky, 1969). Nevertheless, there is a strong tradition, not least in economic applications, to regard $\succcurlyeq$-transitivity as a necessary prerequisite of rationality.

The most famous argument in favor of preference transitivity is the *money-pump argument*. Its basic idea is that transitivity ensures against practical losses (Davidson, McKinsey, & Suppes, 1955; Ramsey, 1928/1950). Consider a situation where a stamp collector has cyclic preferences with respect to three stamps, denoted *x*, *y*, and *z*. He prefers *x* to *y*, *y* to *z*, and *z* to *x*. He is willing to pay 10 cents every time he exchanges a stamp for one he prefers. Now he enters a stamp shop with stamp *x*. The stamp dealer offers him to trade in *x* for *z* if he pays 10 cents. He accepts the deal. Next, the dealer offers him to trade in *z* for *y*, again for a 10-cent fee, which he accepts. Finally, the dealer offers him to trade in *y* for *x*, again for a 10-cent fee, which he accepts. The collector now has obtained the same stamp with which he entered the shop but is 30 cents poorer. Unless he revises his preferences such that they satisfy (4), he will accept further trades from the dealer until he runs out of funds. The money-pump argument thus claims to show that transitivity is a necessary condition in order to avoid negative practical results.

Various arguments have been raised against the money-pump justification of transitivity. First, it has been argued that it relies on particular conditions, which must be in place for intransitive preferences to be exploited in this way. In particular, it must be assumed that the thus-pumped individual cannot take recourse to some strategy that avoids his being pumped—for example, through precommitment or resolution (McClennen, 1990). Second, it has been argued that the money-pump argument itself must make relatively strong consistency assumptions, which undermine its generality (Cubitt & Sugden, 2001).

Another argument for the normative appropriateness of preference transitivity suggests that transitivity is constitutive of the meaning of preference. Drawing an analogy to length measurement, Davidson (1976/1980) asks, "If length is not transitive, what does it mean to use a number to measure length at all? We could find or invent an answer, but unless or until we do, we must strive to interpret 'longer than' so that it comes out transitive. Similarly for 'preferred to'" (p. 273). Violating transitivity,

Davidson claims, thus undermines the very meaning of preferring one option over others.

Yet another argument rests on the importance of preferences for choice. When agents choose simultaneously from all the elements of an alternative set, then preferences should be choice-guiding. They should have such a structure that they can be used to guide their choice among the elements of that set. But when choosing, for example, from $\{x, y, z\}$, a preference relation $\succ$ such that $x \succ y \succ z \succ x$ does not guide choice at all: any or none of the alternatives should be chosen according to $\succ$. The transitivity of preference, it is therefore suggested, is a necessary condition for a meaningful connection between preferences and choice. A critic, however, can point out that preferences are important even when they cannot guide choices. Take, for example, preferences over lottery outcomes: these are real preferences, regardless of the fact that one cannot *choose* between lottery outcomes. Furthermore, the indifference relation does not satisfy choice guidance either. That does not make it irrational to be indifferent between alternatives. Finally, the necessary criteria for choice guidance are much weaker than transitivity (S. O. Hansson, 2001, pp. 23–25). Thus, choice guidance can be an argument for the normative appropriateness of transitivity only under certain restrictions, if at all (for further discussion, see Anand, 1993).

All of these arguments for transitivity remain controversial. In addition, examples have been offered that suggest intransitive preferences to be rational at least sometimes. One important class of such counterexamples is derived from the classic *sorites paradox*. Perhaps the most famous example applied to preference transitivity is Quinn's self-torturer. Here a device has been implanted into the body of a person. The device has 1,001 settings, from 0 (off) to 1,000. Each increase leads to a negligible increase in pain. Each week, the self-torturer "has only two options—to stay put or to advance the dial one setting. But he may advance only one step each week, and he may never retreat. At each advance he gets $10,000." In this way he may "eventually reach settings that will be so painful that he would then gladly relinquish his fortune and return to 0" (Quinn, 1990, p. 79). The self-torturer thus reveals intransitive preferences. But Quinn argued that these preferences are *not* irrational, because the self-torturer has good reasons to turn the dial, yet also has good reasons to want to revert to the original position. Much depends here on what might constitute good reasons to turn the dial. Some have argued that the self-torturer's inability to distinguish between two adjacent dial settings might constitute such a reason (Luce, 1956). Others have sought to explain the intransitive preferences as the result of a heuristic decision procedure (Voorhoeve & Binmore, 2006). It therefore seems that, ultimately, the rationality judgment depends on the process through which these preferences are generated, and here considerable ambiguity remains.

Another important class of counterexamples to transitivity is derived from the classic *Condorcet paradox*. A simple example here is Schumm's Christmas ornaments (Schumm, 1987), involving boxes containing three balls, colored red, blue, and green, respectively; they are represented by the vectors ⟨R1, G1, B1⟩, ⟨R2, G2, B2⟩, and ⟨R3, G3, B3⟩. A chooser might strictly prefer box 1 to box 2, since they contain (to her) equally attractive blue and green balls, but the red ball of box 1 is more attractive than that of box 2. She prefers box 2 to box 3, since they are equal but for the green ball of box 2, which is more attractive than that of box 3. And finally, she prefers box 3 to box 1, since they are equal but for the blue ball of box 3, which is more attractive than that of box 1. Thus,

$$\mathrm{R1} \succ \mathrm{R2} \sim \mathrm{R3} \sim \mathrm{R1},$$

$$\mathrm{G1} \sim \mathrm{G2} \succ \mathrm{G3} \sim \mathrm{G1},$$

$$\mathrm{B1} \sim \mathrm{B2} \sim \mathrm{B3} \succ \mathrm{B1}; \text{ and}$$

$$\langle \mathrm{R1}, \mathrm{G1}, \mathrm{B1} \rangle \succ \langle \mathrm{R2}, \mathrm{G2}, \mathrm{B2} \rangle \succ \langle \mathrm{R3}, \mathrm{G3}, \mathrm{B3} \rangle \succ \langle \mathrm{R1}, \mathrm{G1}, \mathrm{B1} \rangle.$$

The described situation yields a preference cycle, which contradicts transitivity of strict preference—yet arguably, each of these preferences is rational.

1.4 Order Typology

The completeness and transitivity of $\succcurlyeq$, in combination with the definitions of $\succ$ and $\sim$, imply the following properties (Mas-Colell, Whinston, & Green, 1995, p. 7):

$\succ$ is *irreflexive*: $x \succ x$ never holds. (5)

$\succ$ is *asymmetric*: $x \succ y \rightarrow \neg(z \succ x)$ for all $x, y \in \boldsymbol{A}$. (6)

$\succ$ is *transitive*: $x \succ y \wedge y \succ z \rightarrow x \succ z$ for all $x, y, z \in \boldsymbol{A}$. (7)

$\succ$ is *acyclical*: there is no series $x_1, \ldots, x_n$ of alternatives such that $x_1 \succ \ldots \succ x_n \succ x_1$. (8)

$\sim$ is *reflexive*: $x \sim x$ for all $x \in \boldsymbol{A}$. (9)

$\sim$ is *symmetric*: $x \sim y \rightarrow y \sim x$ for all $x, y \in \boldsymbol{A}$. (10)

$\sim$ is *transitive*: $x \sim y \wedge y \sim z \rightarrow x \sim z$ for all $x, y, z \in \boldsymbol{A}$. (11)

IP-transitivity: $x \sim y \wedge y \succ z \rightarrow x \succ z$ for all $x, y, z \in \boldsymbol{A}$. (12)

PI-transitivity: $x \succ y \wedge y \sim z \rightarrow x \succ z$ for all $x, y, z \in \boldsymbol{A}$. (13)

Weak connectivity of $\succcurlyeq$: if $x \neq y$, then $x \succcurlyeq y \vee y \succcurlyeq x$ for all $x, y, z \in \boldsymbol{A}$. (14)

All of these properties individually or in conjunction constitute a weakening of the completeness and transitivity of $\succcurlyeq$ (with the exception of those conjunctions that imply completeness or transitivity). Given the various counterarguments discussed above, some authors have suggested replacing completeness and transitivity with some of their weaker implications.

Sometimes, the following property is added to the above list, even though it is not implied by the completeness and transitivity of $\succcurlyeq$:

Antisymmetry of $\succcurlyeq$: $x \succcurlyeq y \wedge y \succcurlyeq x \rightarrow x = y$. (15)

Preference orderings of varying strength can be characterized by combinations of the above properties. Some of their common names are given in table 8.1.1 (Debreu, 1959; Sen, 1970).

2. Utility Functions

Another way to represent value comparisons or choice patterns is via a real-valued function u: $\boldsymbol{A} \rightarrow \mathbb{R}$, which maps the members of $\boldsymbol{A}$ into the real numbers. Of two alternatives x and y, we say "x is strictly preferred to y," "x is better than y," or "y is worse than x" if and only if $u(x) > u(y)$, and "x is indifferent to y" or "x is equal in value to y" if and only if $u(x) = u(y)$.

2.1 Ordinal Scale Utility Functions

Under this interpretation, there is an obvious connection to the relational representation of preferences. Specifically,

$x \succcurlyeq y$ if and only if $u(x) \geq u(y)$ (Ordinal Representation). (16)

Any function u that assigns a larger number to x than to y will work as such a representation. Consequently, the function u can be replaced with any function u' as long as u' is a *positive monotone transformation* of u. As this transformation property is the defining characteristic of ordinal scales, we call this an *ordinal preference representation*.

Table 8.1.1
An (incomplete) overview of ordering terminologies

	Properties	Name(s)
1.	Reflexive, transitive	*Preorder, quasi-order*
2.	Reflexive, transitive, antisymmetric	*Partial order*
3.	Irreflexive, transitive	*Strict partial order*
4.	Reflexive, transitive, complete	*Total preorder, complete quasi-ordering, weak ordering*
5.	Reflexive, transitive, complete, antisymmetric	*Chain, linear ordering, complete ordering*
6.	Asymmetric, transitive, weakly connected	*Strict total order, strong ordering*

However, a preference relation $\succcurlyeq$ can be ordinally represented by a utility function *only* if it is transitive and complete (von Neumann & Morgenstern, 1947). An *incomplete* preference ordering has an ordinal representation that satisfies

if $x \succcurlyeq y$ then $u(x) \geq u(y)$. (17)

The converse is obviously not true. However, incomplete preferences have been represented as *sets* of utility functions (Aumann, 1962).

Not every transitive and complete preference relation $\succcurlyeq$ can be ordinally represented by a utility function. Consider, for example, a lexicographic preference over goods bundles $(x, y) \in \mathbb{R} \times \mathbb{R}$ with

$(x_1, y_1) \succcurlyeq (x_2, y_2)$ iff $x_1 \geq x_2$ or $(x_1 = x_2$ and $y_1 \geq y_2)$.

Unless $x_i \neq x_j$ for all goods bundles (x_i, y_i), (x_j, y_j) with $i \neq j$, such a relation $\succcurlyeq$ cannot be represented by a utility function, although $\succcurlyeq$ is transitive and complete (Debreu, 1954).

A sufficient condition for a transitive and complete preference relation $\succcurlyeq$ to be ordinally representable is the continuity of $\succcurlyeq$:

> $\succcurlyeq$ is *continuous* iff for all $x \in X$, the *upper contour set of* x, $\{y \in X: y \succcurlyeq x\}$, and the *lower contour set of* x, $\{y \in X: x \succcurlyeq y\}$, are *closed*, that is, they include their boundaries.

Lexicographic preferences like the above example are not continuous. Every continuous, transitive, and complete preference relation $\succcurlyeq$ is ordinally representable by a utility function u (Mas-Colell et al., 1995, pp. 46–47).

In the social sciences, utility functions are typically analyzed with the tools of maximization under constraints. This requires that the thus-analyzed functions are *twice differentiable*. However, not every continuous, transitive, and complete preference relation $\succcurlyeq$ is representable by a differentiable utility function. Consider, for example, the preference $(x_1, y_1) \succcurlyeq (x_2, y_2)$ iff $\min\{x_1, y_1\} \geq \min\{x_2, y_2\}$ (Leontief, 1941).

Lexicographic and Leontief preferences appear at least sometimes to be fully rational. For example, it seems rational to lexicographically prefer subsistence goods to luxury goods when living at subsistence levels. Similarly, it seems rational to have Leontief preferences for perfectly complementary goods, like left and right shoes. Therefore, standard utility functions *cannot* ordinally represent *all* rational preferences.

2.2 Interval-Scale Utility Functions

Although utility functions cannot represent all preferences, an ordinal-scale functional representation of $\succcurlyeq$ represents *nothing but* information contained in $\succcurlyeq$. In other words, all information contained in an ordinal utility function can be represented by a (transitive and complete) ordering $\succcurlyeq$. This is not the case if we interpret utility functions *cardinally*—that is, either as an *interval scale* or as a *ratio scale*. A utility function as an interval scale allows for degrees of difference between preferences, such that, for example, the utility assignments $u(x) = 12$, $u(y) = 6$, and $u(z) = 3$ represent the information that the difference in preference intensity between x and y is twice as large as the difference between y and z. Ratio-scale utility functions are quite rare; they show up, for example, in prospect theory (see chapter 8.3 by Glöckner, this handbook) and in experienced-utility measures (see section 3.1).

Interval differences are particularly important for analyzing preferences over lotteries under the *expected utility* model. Lotteries are ordered tuples of outcome–probability pairs that are mutually exclusive and jointly exhaustive. For example, $[x, p; y, q; z]$ is the lottery that gives outcome x with probability p, y with probability q, and z with probability $1 - p - q$. Under the expected utility model, the utility of a lottery is the sum of the utilities of its outcomes, weighted by their probabilities (von Neumann & Morgenstern, 1947):

$$u([x_1, p_1; x_2, p_2; \ldots; x_n, p_n]) = \sum_{i=1}^{n} p_i \times u(x_i). \tag{18}$$

In the example, then, $u([x, p; y, q; z]) = p \times u(x) + q \times u(y) + (1 - p - q) \times u(z)$. This is a very powerful tool for analyzing preferences over lotteries, making them comparable to preferences over certain outcomes. But it requires an interval utility function. For example, a comparison of lottery $[x,p;y,q;z]$ to certain outcome y depends not only on whether $x \succ y \succ z$ but also on whether the difference in preference intensity between x and y, weighted by p, is larger than the difference between y and z, weighted by $1 - p - q$. Such information cannot be provided by an ordinal scale but requires an interval-scale interpretation of u.

Interval-scale utilities are represented by sets of functions, where each member u of the set is a positive linear transformation of another member u': $u(x) = a \times u'(x) + b$, where $a > 0$. An interval-scale representation of a preference relation $\succcurlyeq$ is thus a proper subset of an ordinal-scale representation of $\succcurlyeq$. For such a representation to exist, and to be unique up to positive linear transformation, however, $\succcurlyeq$ must satisfy certain properties. Different approaches to these representations include Ramsey (1928), von Neumann and Morgenstern (1947), Savage (1954/1972), Jeffrey (1965/1983), and Fishburn (1970). There are substantial differences between these approaches and their respective assumptions, but the rationality of at least some of the required preference properties in each of these approaches is controversial. Thus, in addition to the above-discussed limitations of ordinal utility representations, the interval scale further restricts which rational preferences utility functions can represent. For more detail, see chapter 8.2 by Peterson on decision theory (in this handbook).

Interval utility scales represent differences in intensities of preference. This allows describing two abstract properties: the relation between utility and quantities of goods as well as the relation between utility and uncertainty.

When evaluating a quantifiable good or service, the *marginal utility* is the value added by the last increment of that good or service. For example, if the value of x sheep is $u(x) = 15$ and that of $x + 1$ sheep is $u(x + 1) = 17$, then the marginal utility of adding one sheep to the x previous ones is 2. The so-called *law of diminishing marginal utility* (also known as a "Gossen's first law"; Gossen, 1854/1983) states that marginal utilities decrease, ceteris paribus, as additional amounts of a good or service are added to available resources. One striking implication of this principle is that any good, however much valued in its first unit, will eventually be valued less than even the first unit of a good that is valued rather little. This insight solves the *paradox of value* (famously presented by Smith, 1776/1904): water commands a lower market price than diamonds, despite its being on the whole more useful, in terms of survival, than diamonds. Why? Because most people are well supplied with water and thus value an additional unit very little. By contrast, diamonds are scarce, and many people value the first unit of diamonds higher than the *n*th unit of water.

When evaluating lotteries, *risk attitudes* express evaluations of the respective uncertainties of the alternatives. For example, a lottery [\$100, 0.5; \$0] has an expected value of \$50; those being indifferent between the lottery and \$50 for certain are called *risk neutral*, those preferring the certain outcome *risk averse*, and those preferring the lottery *risk loving*. It is widely agreed that moderate risk aversion is perfectly rational at least in some cases—for example, when someone with few resources prefers a certain gain of \$100,000 to an even gamble between receiving \$200,000 or nothing (Arrow, 1971).

The standard von Neumann–Morgenstern (vNM) expected utility model allows for the representation of such risk attitudes but only by equating them with changing marginal utilities. By definition (18), the utility of lottery [\$100, 0.5; \$0] is the probability-weighted

sum of its outcomes; it can differ from the utility of \$50 for certain only if the utilities of outcomes are not linear with the monetary amounts. Consequently, general risk aversion in the vNM model is identical to diminishing marginal utility.

This is conceptually and empirically problematic. Conceptually, evaluations of quantities and of uncertainty seem distinct. For example, Bengt Hansson (1988) imagines a professional gambler who turns down an offer to trade a single copy of a book he likes for an even-chance gamble between receiving no copy of the book and three copies. According to the vNM model, the gambler must be risk averse. Yet the gambler might reject such an analysis: being a professional gambler, he has habituated himself to being risk neutral. The reason he turns down the gamble, he says, is simply that the second and third copies are of almost no worth to him. Thus, he exhibits diminishing marginal utility for additional books but no risk aversion.

Furthermore, the vNM account cannot represent various empirically observed choice patterns, whose interpretation in terms of risk preferences seems normatively legitimate. For example, a person rejecting a small-stakes gamble (e.g., [\$100, 0.5; –\$100]) can be represented in the vNM model only with a diminishing utility function for money, which would entail that the same person would reject very reasonable large-stakes gambles (e.g., [$\$10^{10}$, 0.5; –\$100]) (Rabin, 2000). Other examples concern choice patterns exhibited in the famous Allais and Ellsberg paradoxes, which cannot be represented by vNM utilities but are often considered expressions of rational attitudes toward risks (see chapter 8.2 by Peterson, this handbook).

The root of these problems seems to be the identification of risk attitudes with nonlinear evaluation of quantity of goods in the vNM model.[2] Various solutions to these problems thus propose to separate the two. One approach is to replace (18) with a risk-weighted expected utility model (Buchak, 2013; Quiggin, 1982), where the agent's risk function r weighs the probabilities p in a nonlinear fashion:

$$u([x_1, p_1; x_2, p_2; \ldots; x_n, p_n]) = \sum_{i=1}^{n} r(p_i) \times u(x_i). \qquad (19)$$

In particular, risk aversion would be represented by a convex risk function (e.g., $r(p) = p^2$). This representation, however, syntactically separates risk attitudes from evaluations. Critics have argued that risk attitudes are a kind of desire and therefore propose to model them in the Jeffrey framework as desirabilities over chance propositions instead (Stefánsson & Bradley, 2019).

3. What Do Preferences and Utilities Represent?

The above two sections concerned formal properties of representational tools. Although these properties of course impose constraints, it turns out that these constraints are not sufficient to uniquely determine what preferences and utilities represent. Instead, multiple coherent positions have been developed, and lively controversies in philosophy and the social sciences continue between them. I will distinguish three broad categories: (1) those that consider utilities as representing mental states—like pleasure or happiness—and consider preferences as derived from utilities, (2) those that consider preferences as representing mental states—like comparative likings or judgments—and derive utilities from preferences, and (3) those that consider preferences as representing choice patterns and derive utilities from preferences.

3.1 Utility as Representing Mental States

Although earlier authors had proposed the maximization of happiness as a moral principle, is it only with Bentham (1789/1988) that happiness is explicitly identified as the presence of pleasure and the absence of pain that people feel as the consequence of action. Bentham argued that happiness can be measured in its intensity, duration, uncertainty, and remoteness and proposed to call this aggregate measure *utility*.

Bentham proposed utility in this sense only as a normative criterion of what should be considered morally good. Early marginalist economists like Gossen (1854/1983) and Jevons (1871) instead aimed to explain social phenomena, based on the motives of (representative) human individuals, and they used the utility concept for that purpose.[3] In Jevons's (1871) words, they "attempted to treat Economy as a Calculus of Pleasure and Pain." Furthermore, they created the utility function as a cardinal representation of mental states. In this framework, the notion of preference, to the extent that it was used at all, was merely derived from hedonistic utility.

While this perspective has been largely replaced in 20th-century social science, it has recently seen something of a revival in Daniel Kahneman's theory of *experienced utility*. Kahneman assumes that people at every moment experience what he terms *instant utility*—meaning pleasure and/or pain. This utility has quantity and valence, with a neutral point on the boundary between desirable and undesirable, between pleasure and pain, and can be measured on a ratio scale. By integrating instant utility over a period, the *total utility* for that period can be computed. Given that on this approach, utility can be measured independently, utility maximization becomes

an empirical, testable hypothesis. To the extent that utility maximization is seen as a criterion of rationality, the experienced-utility approach claims to offer an empirical test for the rationality of individuals in particular domains of behavior (Kahneman, Wakker, & Sarin, 1997).

3.2 Preferences as Representing Mental States

In the social sciences, the preference concept became important for explanatory and predictive purposes with Irving Fisher's (1892/1961) and Vilfredo Pareto's (1906/2014) methodological criticisms of hedonic cardinal utility. Pareto argued that because an accurate measurement procedure for cardinal hedonic utility was not available, social scientists should constrain themselves to merely ordinal comparisons (Bruni & Guala, 2001). This argument turned preference into a fundamental notion of the social sciences, replacing (hedonic) utility.

Economists in the 1930s (Hicks & Allen, 1934) radicalized Pareto's idea and argued that cardinal utility should be excluded in order to purge economics of psychological hedonism. However, their concept of preference retained psychological content: people were assumed to act purposefully and therefore to have preferences that really constitute mental evaluations, rather than being ex post rationalizations of behavior (Lewin, 1996).

Based on (18) and stated preferences over lotteries with known probabilities, von Neumann and Morgenstern devised a measurement of interval utility. Later, Savage (1954/1972) and Jeffrey (1965/1983) devised simultaneous measurements of utilities and probabilities from stated preferences. These new concepts of utility, however, were very different from the older hedonic concept: here the preference concept is basic and the cardinal utility function merely derived.[4]

Because utility is derived from preferences using representation theorems, the assumptions necessary for these theorems are built into the utility representation. Utility representations of this kind thus already assume rationality properties like transitivity, independence, and continuity. It would therefore be pointless to try to test whether such utility representations meet these rationality criteria (cf. Davidson's arguments discussed in section 2.3).

3.3 Preferences as Representing Choice Patterns

In economics, the revealed-preference approach defined preference in terms of choice. Historically, this approach developed out of the pursuit of behavioristic foundations for economic theories—that is, the attempt to eliminate the mental-preference interpretation altogether. Specifically, the downward slope of the demand function—one of the central features of microeconomics—could be derived only from properties of marginal utilities, yet these properties were difficult to measure with the mentalistic Allan–Hicks utility concept. Consequently, Paul Samuelson in 1938 proposed to "start anew . . . dropping off the last vestiges of the utility analysis" (Samuelson, 1938, pp. 61–62).

The basic idea was to assume that *choice* satisfied certain rationality postulates and then define a *revealed preference* relation $x \succcurlyeq_R y$ between two goods bundles x and y as the choice of x in conditions where either x or y could have been chosen.[5] In other words, if an individual at current prices can afford both purchasing x and purchasing y, given her budget, and she chooses to purchase x, then she is said to revealed-prefer x to y. Based on the above definitions (1) and (2), revealed indifference and revealed strict preferences can be derived from $\succcurlyeq_R$. Furthermore, by determining revealed preferences over relevant lotteries and applying one of the expected utility frameworks discussed in section 2.2, these preferences can be represented on an interval utility scale. Note, however, that in this case, the utility function does not represent any mental content at all but only recorded choice patterns.

Although this approach was highly influential at the time, not all economists followed Samuelson in this radical proposal—so that today, the mentalist and the revealed-preference interpretation of preferences are often both presented in textbooks (e.g., Mas-Colell et al., 1995). Indeed, it might be the case that Samuelson himself later changed his mind, shifting from the *definition* of preference in terms of choice to a *measurement* of preference based on choice data (Hands, 2014). Nevertheless, there are many economists today who defend the behaviorist interpretation, for example, by denying that preferences represent the causes of choice (Binmore, 2009) and by insisting on "mindless economics" (Gul & Pesendorfer, 2008). Concurrently, there is an ongoing discussion among philosophers whether the current concept of preference used by economists is a separate theoretical concept—defined against the common use as capturing choice patterns—or whether it indeed is based on the mental, "folk-theoretic" notion (Mäki, 2000; Ross, 2014).

Critics of the behaviorist interpretation have pointed out that the common "folk" explanations of action require both a motivational component (represented by, e.g., preferences or utility) and an epistemic component (i.e., beliefs represented by, e.g., probabilities). The revealed-preference account neglects this epistemic component and therefore cannot provide folk explanations (Hausman, 2012; List & Dietrich, 2016; Sen, 1993). While the defenders of revealed preference might concede as much (see above), Hausman further argues that

the folk-psychological account already provides a lexical definition of preferences and that it is counterproductive to craft a stipulative definition of preferences that is supposed to replace this folk notion in economics.

However, defenders of revealed preference further argue that it is necessary to distinguish between those agents who indeed have preferences as states of minds (e.g., humans and maybe higher animals) and those agents who do not (e.g., machines, plants, or institutions). The former category might choose on the basis of their mentalistically interpreted preferences, and hence the above-discussed effort can aim at eliciting the preferences on which their choices are based. The latter category, despite their lack of states of mind, might nevertheless exhibit behavior that can be interpreted as rational choice. In those cases, one can only speak of preferences reconstructed from choice, without claiming that these preferences describe mental states at all (Gul & Pesendorfer, 2008; Ross, 2014).

4. Preferences Combination

A preference relation $\succcurlyeq$ might be constructed from a vector of preferences $\langle \succcurlyeq_1, \ldots, \succcurlyeq_n \rangle$. Such a relationship has a number of different interpretations. For example, on an *intrapersonal* interpretation, $\succcurlyeq$ might be the *total preference* of an individual—that is, an overall comparison of two relata, taking all relevant considerations into account. The items of $\langle \succcurlyeq_1, \ldots, \succcurlyeq_n \rangle$ might be the partial preferences of this individual—for example, desirable characteristics of economic goods (Lancaster, 1966) or different reasons that one may have to prefer one of the options to another (Dietrich & List, 2013; Pettit, 1991). For example, $\succcurlyeq_1$ might represent evaluations of the respective health benefits of different foodstuffs, while $\succcurlyeq_2$ might represent the comparative sustainability of their production.

An *interpersonal* interpretation, in contrast, would interpret $\langle \succcurlyeq_1, \ldots, \succcurlyeq_n \rangle$ as the collection of total preferences of individuals $1, \ldots, n$ in a group and $\succcurlyeq$ as the group preference. For example, the managing board of a company might consist of n members, each of whom has a preference for a certain course of action. The final decision of which action to take might be interpreted as the expression of the group preference $\succcurlyeq$.

Some authors have argued that for intrapersonal cases, total preferences are always the result of an aggregation of partial preferences (e.g., Hausman, 2012). These authors assume that a total preference relation is uniquely determined by the partial preference relations through a process of aggregation.

Other authors reject the idea that total preferences are uniquely derivable from partial preferences. Instead, they claim that total preferences are constructed at the moment of elicitation and thus influenced by contexts and framings of the elicitation procedure that are not encoded in preexisting partial preferences (Payne, Bettman, & Johnson, 1993). Total preferences seem to be influenced by direct affective responses that are independent of cognitive processes (Zajonc, 1980). For instance, food preferences seem to be partly determined by habituation and are therefore difficult to explain as the outcome of a process exclusively based on well-behaved partial preferences. According to this view, partial preferences are in many cases ex post rationalizations of total preferences rather than the basis from which total preferences are derived.

If total preferences are the result of an aggregation of partial preferences, the question arises what this aggregation process might be. Two kinds of approaches can be distinguished. First, one could seek to develop *intra*personal aggregation by following models of social choice that describe *inter*personal aggregation of preferences (Arrow, 1963; Sen, 1970). However, such models do not allow for trade-offs between partial preferences. That is, if a particular x is much better than a y in one dimension k, the preference $y \succ x$ might nevertheless be possible because the strength of y in other dimensions over x outweighs dimension k. This approach requires (i) a cardinal measure of preference intensity and (ii) the comparability of these preference intensities across dimensions (Keeney & Raiffa, 1993).

5. Preference Change

Preferences sometimes change. This poses difficult predictive and explanatory challenges for social scientists, many of which remain currently unsolved. In this section, I will instead focus on the question of what might constitute *rational preference change*, discussing three approaches: temporal-discounting models, consistency-restoring preference revision models, and doxastic preference change (for a broader overview, see Grüne-Yanoff & Hansson, 2009).

5.1 Temporal-Discounting Models

Sometimes preferences change because the temporal distance between the time of evaluation and the realization of the evaluated event changes (see chapter 10.4 by Raub on rational choice, this handbook). For example, many people consider a tedious task to be performed in three months' time less bad than the same task to be performed

now, even if all background conditions are equal. As this is an intensity change, preferences must be represented on an interval scale. The standard model then represents the utility $u(x,t)$ of an alternative x at time t as equal to the time-independent utility of x (which one might think of as the utility of x if it were realized now), discounted by the delay:

$$u(x,t) = u(x)/(1+r)^t, \tag{20}$$

where r is the discount rate. This is the *exponential discounting model*, proposed by Samuelson (1937), which still dominates economic analysis. It is a fact that people at least sometimes do not discount in agreement with the exponential model. Various nonexponential models have been proposed, of which the *hyperbolic discounting model* is the most prominent.

While it is obvious that the behavior of many people can only be described by assuming some kind of time-discounting, the question remains whether such discounting is rational and justifiable. At least two questions need to be distinguished—first, whether discounting utility of a prospect merely for its distance in time is ever justified, particularly whether steep discount rates can be justified—irrespective of what the particular form of the discounting function is. Critics argue that one should want one's life, as a whole, to go as well as possible and that counting some parts of life more than others interferes with this goal (Pigou, 1920; Ramsey, 1928; Rawls, 1971). According to this view, it is irrational to prefer a smaller immediate good to a greater future good, because now and later are equal parts of one life. Choosing the smaller good or the greater bad makes one's life, as a whole, turn out worse: "Rationality requires an impartial concern for all parts of our life. The mere difference of location in time, of something's being earlier or later, is not a rational ground for having more or less regard for it" (Rawls, 1971, p. 293). Critics of temporal discounting often attribute apparent departures from temporal neutrality to a cognitive illusion, which causes people to see future pleasures or pains in some diminished form.

Against the temporal neutrality of preferences, some have argued that there is no enduring, irreducible entity over time to whom all future utility can be ascribed; they deny that all parts of one's future are equally parts of oneself (Parfit, 1984). They argue, instead, that a person is a succession of overlapping selves related to varying degrees by memories, physical continuities, similarities of character and interests, and so on. On this view, it may be just as rational to discount one's "own" future preferences as to discount the preferences of another, distinct individual, because the divisions between the stages of one's life may be as "deep" as the distinctions between individuals.

The second question concerns the particular form of discounting. One of the central properties of Samuelson's exponential model is that it is *order-preserving*. That is, an individual who prefers $(y,t+i)$ to (x,t) at some time previous to t will also prefer $(y,t+i)$ to (x,t) at time t. The only difference between these two evaluations is the increase in utility for both $(y,t+i)$ and (x,t). The exponential model ensures that these intensities increase proportionally, so that the preference ranking of options never changes.

Many people experience temporal preference changes that contradict this order-preserving property: they start out preferring $(y,t+i)$ to (x,t) at some earlier time—perhaps because y, when compared to x at the same time, is so much better. But as they approach t, and hence the realization of x, they experience a *preference reversal*: near to and at t, they suddenly prefer (x,t) to $(y,t+i)$. For a simple example, consider a person who prefers one apple today to two apples tomorrow but (today) prefers two apples in 51 days to one apple in 50 days. Although this is a plausible preference pattern, it is incompatible with the exponential model. It can, however, be accounted for in a model with a declining discount rate. Pioneered by Ainslie (1992), psychologists and behavioral economists have therefore proposed to replace Samuelson's exponential-discounting model with a model of *hyperbolic discounting*. The hyperbolic model discounts the future consumption with a parameter inversely proportional to the delay of consumption and hence can represent preference reversals like the above (for an overview, see Grüne-Yanoff, 2015).

Choosing according to hyperbolic-discounted utility leads to periodic abandonment of previous consumption plans in favor of new ones, which would be in turn abandoned in the next period. Such patterns of time-inconsistent choice are often related to failures of *self-control*, and hyperbolic discounting is therefore often seen as an expression of irrational preferences (Strotz, 1956).

However, some authors have argued that hyperbolically discounted preferences are adaptive for environments in which the uncertainty of obtaining a distant reward increases nonlinearly with the temporal distance (Sozou, 1998). In the light of these results, hyperbolic discounting might be rational at least in some contexts. Against this, others have argued that it might be possible to separate these considerations of uncertainty from pure time preferences (Andreoni & Sprenger, 2012) and that nonexponential discounting of pure time preferences remains a sign of irrationality.

5.2 Consistency-Restoring Preference Revision Models

If an agent forms a specific preference as the result of some experience, further changes in her overall preference state are often necessary to regain consistency. Using tools from belief revision (see chapter 5.2 by Rott on belief revision, this handbook), preference change has been modeled as an adjustment to such inputs (Grüne-Yanoff, 2013; S. O. Hansson, 1995; Liu, 2011). Changes in preference are triggered by inputs that are represented by sentences expressing new preference patterns.

For example, if a subject grows tired of her previous favorite brand of mustard, x, and starts to like brand z better, then this will be represented by a change with the sentence "z is better than x," in formal language: $z \succ x$, as an input. However, a change in which the previous preference $x \succ z$ is substituted by the new preference $z \succ x$ can happen in different ways. For instance, there may be a third brand, y, that was previously placed between x and z in the preference ordering. The instruction to make the new preference relation satisfy $z \succ x$ does not tell us where y should be placed in the new ordering. The new ordering may, for instance, be either $z \succ x \succ y$ or $z \succ y \succ x$. One way to deal with this is to include additional information in the input, for instance, specifying which element(s) of the alternative-set should be moved while the others keep their previous positions. In my example, if only z is going to be moved, then the outcome should satisfy $z \succ x \succ y$. These and other considerations make it necessary to modify the standard model of belief change in order to accommodate the subject matter of preferences.

5.3 Doxastic Preference Change

Preferences also change as a consequence of belief changes. Two kinds of beliefs are especially important here. The first is the belief that the presence of state x would make a desired state y more likely—a rise in probability of y given x produces a rise in the desirability of x, and vice versa. The second kind of belief relevant for doxastic preference change concerns prospects that influence the preference for other prospects without being probabilistically related. For example, one's preference for winning a trip to Florida in the lottery will crucially depend on one's belief about the weather there during the specified travel time, even though these two prospects are probabilistically unrelated. More generally, if $x \wedge y$ is preferred to $x \wedge \neg y$, with x and y probabilistically not correlated, then a rise in the probability of x will result in a rise in the desirability of y (even if it does not affect the probability of y), and vice versa.

Jeffrey (1977) provided a simple model of preference change as the consequence of an agent's coming to believe a proposition A to be true. His model incorporates both kinds of belief relevant for doxastic preference change. The basic idea is that $\langle u, P\rangle$ represents the unconditional preference $\succcurlyeq$ if P is the probability distribution based on the agent's actual information. The *conditional preference ordering* $\succcurlyeq_A$, in contrast, is represented by the tuple $\langle u, P_A\rangle$, where P_A is the probability distribution based on the counterfactual scenario that the agent accepts proposition A as true. That is, the agent imagines that if he changed his whole belief system from P to P_A, then he would have the preference relation $\succcurlyeq_A$ as represented by $\langle u, P_A\rangle$ (for more discussion of conditional preferences, see Bradley, 2005; Joyce, 1999; Luce & Krantz, 1971).

Notes

1. The study of these properties can be traced back to Book III of Aristotle's Topics. Since the early 20th century, several philosophers have studied the structure of preferences with logical tools. In 1957 and 1963, respectively, Sören Halldén and Georg Henrik von Wright proposed the first complete systems of preference logic (Halldén, 1957; von Wright, 1963).

2. This identification has a long tradition, going back to Daniel Bernoulli's solution to the St. Petersburg paradox (Bernoulli, 1738/1954).

3. Jevons discussed the effect of character, gender, and race on the utility function and avoided dealing with the implied heterogeneity of utility functions by invoking such a representative individual (White, 1994).

4. Defenders of the Benthamite conception call this new vNM representation *decision utility*, to distinguish it from their *experienced utility*.

5. The two main assumptions are the *weak axiom of revealed preferences* (WARP) and the *strong axiom of revealed preferences* (SARP). WARP says that if x is chosen when y is available (in some alternative-set A), then there must not be another alternative-set A' containing both alternatives for which y is chosen and x is not. SARP says that if from a set A_1 of alternatives, x is chosen when y and z are available, and if in some other set of alternatives, A_2, y is chosen while z is available, then there can be no set of alternatives containing alternatives x and z for which z is chosen and x is not. (SARP says this for chains of any length.)

References

Ainslie, G. (1992). *Picoeconomics: The strategic interaction of successive motivational states within the person*. Cambridge, England: Cambridge University Press.

Anand, P. (1993). The philosophy of intransitive preference. *Economic Journal*, *103*, 337–346.

Andreoni, J., & Sprenger, C. (2012). Risk preferences are not time preferences. *American Economic Review*, *102*(7), 3357–3376.

Arrow, K. J. (1963). *Social choice and individual values* (2nd ed.). New Haven, CT: Yale University Press. (Original work published 1951)

Arrow, K. J. (1971). The theory of risk aversion. In *Essays in the theory of risk-bearing* (pp. 90–120). Chicago, IL: Markham.

Aumann, R. J. (1962). Utility theory without the completeness axiom. *Econometrica*, *30*(3), 445–462.

Bentham, J. (1988). *An introduction to the principles of morals and legislation*. Buffalo, NY: Prometheus. (Original work published 1789)

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk (L. Sommer, Trans.). *Econometrica 22*(1), 22–36. (Original work published 1738)

Binmore, K. (2009). *Rational decisions*. Princeton, NJ: Princeton University Press.

Bradley, R. (2005). Radical probabilism and mental kinematics. *Philosophy of Science*, *72*, 342–364.

Bruni, L., & Guala, F. (2001). Vilfredo Pareto and the epistemological foundations of choice theory. *History of Political Economy*, *33*, 21–49.

Buchak, L. (2013). *Risk and rationality*. Oxford, England: Oxford University Press.

Chang, R. (Ed.). (1997). *Incommensurability, incomparability, and practical reason*. Cambridge, England: Cambridge University Press.

Cubitt, R. P., & Sugden, R. (2001). On money pumps. *Games and Economic Behavior*, *37*(1), 121–160.

Davidson, D. (1980). Hempel on explaining action. In *Essays on actions and events* (pp. 261–275). Oxford, England: Oxford University Press. (Original work published 1976)

Davidson, D., McKinsey, J. C. C., & Suppes, P. (1955). Outlines of a formal theory of value, I. *Philosophy of Science*, *22*, 140–160.

Debreu, G. (1954). Representation of a preference ordering by a numerical function. In R. M. Thrall, C. H. Coombs, & R. L. Davis (Eds.), *Decision processes* (pp. 159–166). New York, NY: Wiley.

Debreu, G. (1959). *Theory of value: An axiomatic analysis of economic equilibrium*. New Haven, CT: Yale University Press.

Dietrich, F., & List, C. (2013). A reason-based theory of rational choice. *Noûs*, *47*(1), 104–134.

Fishburn, P. C. (1970). *Utility theory for decision making*. New York, NY: Wiley.

Fisher, I. (1961). *Mathematical investigations in the theory of value and prices*. New Haven, CT: Yale University Press. (Original work published 1892)

Gossen, H. H. (1983). *Die Entwicklung der Gesetze des menschlichen Verkehrs und der daraus fließenden Regeln für menschliches Handeln* [The laws of human relations and the rules of human action derived therefrom]. Boston, MA: MIT Press. (Original work published 1854)

Grüne-Yanoff, T. (2013). Preference change and conservatism. *Synthese*, *190*(14), 2623–2641.

Grüne-Yanoff, T. (2015). Models of temporal discounting 1937–2000: An interdisciplinary exchange between economics and psychology. *Science in Context*, *28*(4), 675–713.

Grüne-Yanoff, T., & Hansson, S. O. (Eds.). (2009). *Preference change: Approaches from philosophy, economics and psychology* (Theory and Decision Library). Dordrecht, Netherlands: Springer.

Gul, F., & Pesendorfer, W. (2008). The case for mindless economics. In A. Caplin & A. Schotter (Eds.), *The foundations of positive and normative economics: A handbook* (pp. 3–42). Oxford, England: Oxford University Press.

Halldén, S. (1957). *On the logic of "better."* Lund, Sweden: Library of Theoria.

Hands, D. W. (2014). Paul Samuelson and revealed preference theory. *History of Political Economy*, *46*, 85–116.

Hansson, B. (1988). Risk aversion as a problem of conjoint measurement. In P. Gärdenfors & N.-E. Sahlin (Eds.), *Decision, probability, and utility: Selected readings* (pp. 136–158). Cambridge, England: Cambridge University Press.

Hansson, S. O. (1995). Changes in preference. *Theory and Decision*, *38*, 1–28.

Hansson, S. O. (1996). Decision making under great uncertainty. *Philosophy of the Social Sciences*, *26*(3), 369–386.

Hansson, S. O. (2001). *The structure of values and norms*. Cambridge, England: Cambridge University Press.

Hausman, D. M. (2012). *Preference, value, choice, and welfare*. New York, NY: Cambridge University Press.

Hicks, J. R., & Allen, R. G. D. (1934). A reconsideration of the theory of value. Part I. *Economica*, *1*(1), 52–76.

Jeffrey, R. C. (1977). A note on the kinematics of preference. *Erkenntnis*, *11*, 135–141.

Jeffrey, R. C. (1983). *The logic of decision* (2nd ed.). Chicago, IL: University of Chicago Press. (Original work published 1965)

Jevons, W. S. (1871). *The theory of political economy*. London, England: Macmillan.

Joyce, J. M. (1999). *Foundations of causal decision theory*. Cambridge, England: Cambridge University Press.

Kahneman, D., Wakker, P. P., & Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *Quarterly Journal of Economics*, *112*(2), 375–406.

Keeney, R. L., & Raiffa, H. (1993). *Decisions with multiple objectives: Preferences and value tradeoffs*. Cambridge, England: Cambridge University Press.

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy*, *74*(2), 132–157.

Leontief, W. W. (1941). *The structure of American economy, 1919–1929: An empirical application of equilibrium analysis*. Cambridge, MA: Harvard University Press.

Lewin, S. B. (1996). Economics and psychology: Lessons for our own day from the early twentieth century. *Journal of Economic Literature*, *34*(3), 1293–1323.

List, C., & Dietrich, F. (2016). Mentalism versus behaviourism in economics: A philosophy-of-science perspective. *Economics & Philosophy*, *32*(2), 249–281.

Liu, F. (2011). *Reasoning about preference dynamics*. Dordrecht, Netherlands: Springer.

Luce, R. D. (1956). Semiorders and a theory of utility discrimination. *Econometrica*, *24*, 178–191.

Luce, R. D., & Krantz, D. H. (1971). Conditional expected utility. *Econometrica*, *39*, 253–271.

Mäki, U. (2000). Reclaiming relevant realism. *Journal of Economic Methodology*, *7*(1), 109–125.

Mas-Colell, A., Whinston, M. D., & Green, J. R. (1995). *Microeconomic theory*. New York, NY: Oxford University Press.

McClennen, E. F. (1990). *Rationality and dynamic choice*. Cambridge, England: Cambridge University Press.

Pareto, V. (2014). *Manual of political economy: A critical and variorum edition*. Oxford, England: Oxford University Press. (Original work published 1906)

Parfit, D. (1984). *Reasons and persons*. Oxford, England: Oxford University Press.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge, England: Cambridge University Press.

Pettit, P. (1991). Decision theory and folk psychology. In M. Bacharach & S. Hurley (Eds.), *Foundations of decision theory* (pp. 147–175). Cambridge, MA: Blackwell.

Pigou, A. C. (1920). *The economics of welfare*. London, England: Macmillan.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior and Organization*, *3*(5), 323–343.

Quinn, W. S. (1990). The puzzle of the self-torturer. *Philosophical Studies*, *59*, 79–90.

Rabin, M. (2000). Risk aversion and expected utility theory: A calibration theorem. *Econometrica*, *68*(5), 1281–1292.

Ramsey, F. P. (1928). A mathematical theory of saving. *Economic Journal*, *38*, 543–559.

Ramsey, F. P. (1950). Truth and probability (original manuscript, 1928). In R. B. Braithwaite (Ed.), *The foundations of mathematics and other logical essays* (pp. 156–198). London, England: Routledge & Kegan Paul.

Rawls, J. (1971). *A theory of justice*. Oxford, England: Oxford University Press.

Ross, D. (2014). *Philosophy of economics*. New York, NY: Palgrave Macmillan.

Samuelson, P. A. (1937). A note on measurement of utility. *Review of Economic Studies*, *4*, 155–161.

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, *5*(17), 61–71.

Savage, L. J. (1972). *The foundations of statistics* (2nd ed.). New York, NY: Dover. (Original work published 1954)

Schumm, G. F. (1987). Transitivity, preference and indifference. *Philosophical Studies*, *52*, 435–437.

Sen, A. (1970). *Collective choice and social welfare*. San Francisco, CA: Holden Day.

Sen, A. (1993). Internal consistency of choice. *Econometrica*, *61*(3), 495–521.

Slovic, P. (1995). The construction of preference. *American Psychologist*, *50*(5), 364–371.

Smith, A. (1904). *An inquiry into the nature and causes of the wealth of nations*. London, England: Methuen. (Original work published 1776)

Sozou, P. D. (1998). On hyperbolic discounting and uncertain hazard rates. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, *265*(1409), 2015–2020.

Stefánsson, H. O., & Bradley, R. (2019). What is risk aversion? *British Journal for the Philosophy of Science*, *70*(1), 77–102.

Strotz, R. H. (1956). Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies*, *23*(3), 165–180.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, *76*, 31–48.

von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior* (2nd ed.). Princeton, NJ: Princeton University Press.

von Wright, G. H. (1963). *The logic of preference*. Edinburgh, Scotland: Edinburgh University Press.

Voorhoeve, A., & Binmore, K. (2006). Transitivity, the sorites paradox, and similarity-based decision-making. *Erkenntnis*, *64*(1), 101–114.

White, M. V. (1994). Bridging the natural and the social: Science and character in Jevons' political economy. *Economic Inquiry*, *32*, 429–444.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, *35*(2), 151–175.

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Knauff, Markus, editor. | Spohn, Wolfgang, editor.
Title: The handbook of rationality / edited by Markus Knauff and Wolfgang Spohn.
Description: Cambridge : The MIT Press, 2021. | Includes bibliographical references and index.
Identifiers: LCCN 2020048455 | ISBN 9780262045070 (hardcover)
Subjects: LCSH: Reasoning (Psychology) | Reason. | Cognitive psychology. | Logic. | Philosophy of mind.
Classification: LCC BF442 .H36 2021 | DDC 153.4/3—dc23
LC record available at https://lccn.loc.gov/2020048455