Interpersonal utility comparisonsa perspective on selected models

Afschin Gandjour

Interpersonal utility comparisons – a perspective on selected models

Afschin Gandjour

Frankfurt School of Finance & Management, Frankfurt, Germany

Corresponding author: Prof. Afschin Gandjour, Frankfurt School of Finance & Management, Adickesallee 32-34, 60322 Frankfurt am Main, Germany; phone: +49-(0)69-154008832; fax: +49-(0)69-1540084832; e-mail: a.gandjour@fs.de

Interpersonal utility comparisons – a perspective on selected models

Abstract

Recently, new models for comparing the strength of individual preferences have been proposed. This perspective article discusses these models within the scope of different accounts of how people attribute mental states to others. The paper shows that the new models suffer from the same shortcoming as Harsanyi's Equiprobability Model of Moral Value Judgments, which entails an interpersonal comparison of strengths of preferences.

Key words: interpersonal utility comparison; mental states; empathy; sympathy

JEL Classification: B41, D63

Interpersonal comparison of preferences and preference strengths has been a controversial topic for many decades. Harsanyi published in 1953 and extended over the following two decades (1955, 1977a) a seminal model for interpersonal comparison of strengths of preferences, which suggests that Bayesian rationality postulates together with interpersonal utility comparisons entail an average utilitarian theory. This model presents an axiomatic foundation for utilitarian morality (Harsanyi 1978). Also called the Equiprobability Model of Moral Value Judgments, it not only considers egoistic preferences but also moral value judgments about the utility distribution in society. These moral judgments are made by an impartial and rational observer behind a veil of ignorance. The observer has an equal probability of being any member of society. Each member of society possesses von Neumann-Morgenstern (vNM) preferences over lotteries. The observer has preferences over the positions in society that are also represented by a vNM utility function. These preferences are called extended preferences. They are "morally valid preferences", which exclude irrational and antisocial preferences (Harsanyi 1975).

Importantly, Harsanyi's model is based on empathy as a means of assessing people's strength of preferences. According to Harsanyi, the observer shows empathy with another person, puts herself in the other person's shoes, and imagines living the other person's life. Thus, the observer completely takes over the preferences of the other person. That is, preferences of the observer and the target become the same (a condition that Harsanyi calls "similarity postulate" (1977b, p. 639)).

Harsanyi's model has been subject to a number of criticisms (see Gandjour (2021) for a recent summary); a central one being that the vNM utility function is not a cardinal representation of utility. Reacting to this critique, Harsanyi suggested the use of conversion ratios to "convert all these utility functions into the same common utility unit" (Harsanyi 1977a, p. 57). Thus, the observer is able to perform interpersonal comparisons without the need to portray preferences of others on her personal scale and using her personal preferences. The latter implication is desirable as it aligns with what Harsanyi calls the "principle of acceptance": He points out that "[t]he interests of each individual must be defined fundamentally in terms of his own personal preferences and not in terms of what somebody else thinks is "good for him"" (Harsanyi 1977a, p. 52). Hence, the observer must accept preferences of others.

A recently published account by Adler (2014) builds upon Harsanyi's Equiprobability Model. Yet, instead of using empathy as a means of conducting intrapersonal or interpersonal utility comparisons, Adler's account relies on sympathy. Thus, it aims at avoiding two key problems he attributes to Harsanyi's account: (i) satisfaction of some preferences does not contribute to well-being; and (ii) the empathetic observer may not possess information about certain attributes of other individuals such as birth dates. I will discuss Adler's critique on Harsanyi's account as well as his own model in greater detail below.

Another model with the intention to enable interpersonal comparisons of preferences was presented earlier by Davidson (1986, 2004). According to Davidson we naturally compare our mental states - beliefs, desires, pretending, knowledge, etc. - with those of other people when interpreting their behavior. To this end, we project certain aspects of our mental states on others. Thus, the basis for interpersonal comparisons is inherent in the very activity of interpretation. Weintraub (1998) contends that from interpreting another person's behavior does not follow that we are able to compare preferences in terms of their strength or intensity. Instead, we are "force[d] (...) to attribute the same utility scale to all agents; to assume, that is, that agents' utilities straddle the same interval." (p. 309). Hence, according to Weintraub, the model by Davidson does not imply an interpersonal comparison of the intensity of preference satisfaction. Yet, the very same criticism applies to Harsanyi's model using conversion ratios as discussed above. This is because the use of conversion ratios aims at producing the same utility scale for everyone and hence does not account for the intensity of preferences (Weymark 1991).

A fairly recent paper by Rossi (2011) modifies Davidson's model to enable a comparison of the strength of individual preferences and avoid the criticism put forward by Weintraub (1998). Specifically, Rossi makes the following case: When we interpret other people's behavior and ascribe mental states to others, we believe that other people would form the same preferences and mental states (and, as I infer from his writing, the same strength of preferences as well) if they were subject to the same circumstances. Thus, assuming the same circumstances, we are able to assess the strength of preferences of other people and interpret their behavior. Still, the author contends that this interpretation of behavior is subject to the *belief* that under the same circumstances people would form the same preferences (what he calls the "principle of similarity"). If the belief is justified, it is possible to make interpersonal comparisons of

preference strengths. Note that the "principle of similarity" closely resembles Harsanyi's "similarity postulate" (Harsanyi 1977b).

The purpose of this perspective article is to discuss the Davidson/Rossi solution as well as Adler's sympathy-based model within the scope of different accounts of how people attribute mental states to others. The paper shows that both models are incomplete as they only allow for interpersonal utility comparisons on an ordinal scale, which is also a well-known limitation of Harsanyi's original account. To provide a classification of the various models and define recommendations for further development, we first discuss different theories of mental state attribution.

Accounts of mental state attribution

Two accounts of attributing mental states to self and others stand out in the current literature: theory-theory accounts (e.g., Carruthers (1996)) and simulation-theory accounts (e.g., Goldman (2006)). Both can also be considered accounts of mindreading, which "is the activity of representing specific mental states of others, for example, their perceptions, goals, beliefs, expectations, and the like" (Gallese 1998, p. 50). These mental states are "invoked to explain and predict behavior" (Gallese 1998, p. 50). According to Gallese and Goldman (1998, p. 52) the core difference between the two accounts is that while theory-theory accounts depict "mindreading as a thoroughly 'detached' theoretical activity", simulation-theory accounts depict "mindreading as incorporating an attempt to replicate, mimic, or impersonate the mental life of the target agent". According to theory-theory accounts, we attribute mental states to others essentially by theoretical reasoning in accordance with causal laws of behavior. That is, we start with initial information about the target's beliefs and desires and use general principles to generate a prediction about the target's mental states and behavior. We pull the mental concepts used to predict behavior out of a body of implicit knowledge we all possess (Savaki 2010). Importantly, theory-theory accounts do not use empathy as a process of attributing mental states to others.

In contrast, simulation theorists argue that we understand others by mentally simulating them. We take the position of the other person, adopt pretend beliefs and pretend desires that we think the other person has, and use these pretend mental states to understand the other person's behavior (Gallagher 2001, Spaulding 2012). Hence, simulation theorists make appeal to empathy.

As a word of caution, both accounts of mental state attribution exist in several versions. In addition, there are hybrid versions that combine the two.

The Davidson/Rossi solution

Rossi's account does not address the fundamental processes used by individuals to ascribe mental states to others (and to themselves). Rather, it is a higher-level account of the conditions that render any such attribution possible. Hence, Rossi's solution can be compatible with different accounts of attributing mental states to others. If Rossi's solution invoked empathy as the relevant mechanism, it would result in an account of interpersonal comparisons of preference strengths similar to the one by Harsanyi (but without the use of conversion ratios). The alternative way of

attributing mental states to others would be through prediction based on some implicit knowledge, which is the principle underlying theory-theory accounts.

Remember that Rossi's "principle of similarity" starts from the "same circumstances"; yet, a complete account of interpersonal utility comparison also needs to address how utility amounts compare across different circumstances. That is, we need to compare the utility amount of person i in situation x (by mindreading person i) with the utility amount of person j in situation y (by mindreading person j). Yet, when comparing preferences across different circumstances, Rossi's account enables such comparison only on a ratio scale but not on an absolute scale.[1] A ratio scale fixes the zero point of the utility scale (Bradley 2008) and thus allows measuring the relative strength of preference, for example, the strength of preference for x relative to y (Bradley 2008; see Barrett (2019) for a similar approach based on desire strength). Thus, we can compare utility ratios of different individuals. Bradley (2008) argues that oftentimes it is sufficient to have a ratio scale, for example, if we want to reach aggregate judgments about the relative desirability of two courses of action (i.e., the desirability for x relative to y).[2] According to Bradley (2008), the utility value '0' should be fixed in correspondence with the ethically neutral proposition, that is, the proposition whose realization is a matter of indifference to the individual (Bradley 2008, pp. 95-96). In contrast, the common zero-one rule fixes the utility value '0' in correspondence with the worst (Hausman 1995, p. 480) or least preferred option. Bradley's proposal is not the same as the common zero-one rule because the ethically neutral proposition does not necessarily coincide with an individual's least preferred option (Bradley 2008, p. 96). Still, Bradley explicitly allows for prospects that are less desirable than the ethically neutral proposition, by assigning them negative utility values.

But regardless of how the zero point is operationalized, taking the position of a utilitarian, whose aim it is to assess the aggregated utility of a certain course of action, it is still necessary to portray strength of preferences on an absolute scale, that is, a scale which is absolutely unique to each individual (cf. Davidson 1955). The reason is that a utilitarian wants to know to what degree the

[1] A cardinal or metric scale can determine differences and/or proportions of the outcomes of the characteristic of interest. There are three different types of cardinal scales: interval, proportional (ratio), and absolute scale (Mittag 1993). Both the proportional (ratio) and the absolute scale start at the natural origin zero (Mittag 1993). In contrast to a proportional (ratio) scale, however, an absolute scale has natural units (Mittag 1993).

[2] From my interpretation of Bradley's account, relative desirability is not an ordinal measure because relative desirability is quantifiable.

utility of those whose preference is satisfied is able to compensate the utility of those whose preference is dissatisfied. If this analysis is conducted in comparison to the status quo, it is equivalent to a comparison of the utility gains and losses (utility increments) compared to the status quo (Narens 2020).

Rossi's solution is different from Harsanyi's in another respect. According to Harsanyi, the "similarity postulate" is justified by reference to pragmatic considerations, that is, the "similarity postulate" is the simplest, most parsimonious, and least arbitrary amongst alternative hypotheses (Harsanyi 1982, p. 51). By contrast, Rossi's "principle of similarity" is the principle that makes interpretation of other people's behavior possible. If so, the arguments offered by Rossi and Harsanyi to defend the possibility of justified interpersonal utility comparisons are different: Rossi offers a 'modest' transcendental argument, whereas Harsanyi offers a pragmatic argument.

Sympathy-based account

In line with other authors (Parfit 1984, p. 494; Scanlon 1996; Arneson 1999, p. 124; Darwall 2002, p. 53), Adler (2014) points out that satisfying preferences sometimes does not contribute to well-being. That is, observers can have "non-self-interested preferences". Adler calls this the "wrong kind of preference" problem. He provides the following example, which is cited in full because it is an important (but mistaken) reason for his alternative proposal discussed below:

> "[I]magine that there are five people in the population: *i*, *j*, *k*, *l*, *m*. Outcome *x* is one in which individuals' incomes range in \$20 000 increments from \$20 000 to \$100 000. Individual *i* has income \$20 000, and individual *j* has income \$100 000; while the other three have, respectively, incomes of \$40 000, \$60 000 and \$80 000. Individual *i* has tastes R_i, etc.
>
> Then $(A_i(x), R_i)$ is the bundle (having an income of \$20 000; having tastes R_i; being part of a population of five individuals where the other incomes are \$40 000, \$60 000, \$80 000, \$100 000 and where the other individuals have tastes R_k, R_l, R_m and R_j). And $(A_j(x), R_j)$ is the bundle (having an income of \$100 000; having tastes R_j; being part of a population of five individuals where the other incomes are \$20 000, \$40 000, \$60 000, \$80 000 and where the other tastes are R_i, R_k, R_l and R_m).

> Imagine, now, that *k* is an *impartial* spectator. In the exercise of ranking hybrid bundles, she assumes an attitude, not of self-interest, but rather of impartiality between her interests and everyone else's. If so *k* will be *indifferent* between the bundles $(A_i(x), R_i)$ and $(A_j(x), R_j)$. She doesn't care, from this impartial perspective, whether she is the one with $20 000 and particular tastes in a given population distribution of income and tastes, or she is the one with $100 000 and particular tastes in the very same distribution of income and tastes. But, of course, (*x*; *i*) and (*x*; *j*) are not equally good for well-being. It is worse for well-being, ceteris paribus, to be the person with the lowest income in a given distribution of income, rather than the person with the highest (at least if R_i and R_j both include a taste for more income rather than less)."

The other key problem Adler (2014) discerns in Harsanyi's account is that individuals might possess attributes that the observer necessarily lacks and cannot acquire without changing who she is. Therefore, the observer cannot really put herself in the individual's shoes. Adler provides the example of an observer who is supposed to formulate extended preferences over two lives, one living in the first century BC and the other in the 16th century AC. He goes on to argue that an observer who was born in 1980 necessarily lacks the essential attribute of birth timing of the other two lives.

For these reasons Adler (2014) proposes a "sympathy-based conception of extended preferences". That is, in ranking an individual's outcomes[3] x and y the observer "does *not* engage in the thought experiment of acquiring" an individual's causal factors ("attributes") of x and y. Instead, her extended preference is reduced to an outcome preference under a condition of unreserved sympathy[4] for the individual (while her preference is still represented by a vNM

[3] According to Adler, outcomes are defined as "arbitrarily detailed specifications of possible worlds, [which] do not specify individuals' preferences".

[4] According to the Stanford Encyclopedia of Philosophy sympathy is defined as follows (Stueber 2018): "In contrast to affective empathy, sympathy is not an emotion that is congruent with the other's emotion or situation such as feeling the sadness of the other person's grieving for the death of his father. Rather, sympathy is seen as an emotion sui generis that has the other's negative emotion or situation as its object from the perspective of somebody who cares for the other person's well being (Darwall 1998). In this sense, sympathy consists of "feeling sorrow or concern for the distressed or needy other," a feeling for the other out of a "heightened awareness of the suffering of another person as something that needs to be alleviated"."

utility function). As the preference exercise is reduced to an assessment of outcomes, the observer does not need to consider non-mental attributes such as birth dates. Still, the observer "can take account of all of the subjects' attributes (…) in arriving at [his] well-being judgements". The observer can take account of the individual's preferences, "without requiring the observer to take those preferences as decisive". From my interpretation, the observer may therefore ignore attributes, causal factors, and background information of an individual even if they are not missing. In that case, she would suppress some of the background information of an individual and perhaps bring in her own background information. In any case, the observer is never asked to imagine acquiring someone else's identity (p. 158).

Adler's view that sympathy does not require mindreading through empathy is supported by the literature. For example, Sober and Wilson (1998, p. 236) have argued that empathy requires one to be a psychologist, but that sympathy does not:

> "Empathy entails a belief about the emotions experienced by another person. Empathic individuals are "psychologists" (..); they have beliefs about the mental states of others. Sympathy does not require this. You can sympathize with someone just by being moved by their objective situation; you need not consider their subjective state. Sympathetic individuals have minds, of course; but it is not part of our definition that sympathetic individuals must be psychologists."

Similarly, Stueber (2018) argues that "sympathy does not necessarily require feeling any kind of congruent emotions on part of the observer, a detached recognition or representation that the other is in need or suffers might be sufficient". According to this view, Adler's approach could be classified as a theory-theory account.

Furthermore, Adler distinguishes between paternalistic[5] and non-paternalistic sympathetic preferences. In fact, his account can accommodate both types of preferences. While paternalistic altruism is utility derived from another's consumption, non-paternalistic altruism is utility derived from another's own utility. As vividly described by Hoffmann (2006):

[5] Making decisions for other people rather than letting them take responsibility for their own lives (Cambridge Dictionary)

> "Parents have paternalistic concern for their children when they care about their children's health or consumption in and of itself, not because of what the child likes. A classic example of paternalistic caring is the parent's admonishment, "Eat your spinach. I don't care if you don't like it. It's good for you." Parents have non-paternalistic concern for their children when they care about the child's consumption or health because it makes the child happy."

Adler argues that limiting his account to non-paternalistic preferences would result in the same ranking of outcomes by the observer and the subject.[6] In support of his reasoning, economists now widely accept that non-paternalistic altruism leads to double counting of individual utility and hence should be excluded from preference elicitation exercises (Bergstrom 1982). That is, individual utility is counted once in the utility function of the individual in question and once in the utility function of the individual demonstrating non-paternalistic altruism (Bergstrom 1982). Hence, compared to self-interested preferences, non-paternalistic altruism does not change the allocation of resources and ranking of outcomes.

In the following I provide a critique of Adler's "sympathy-based conception". First, his motivating example cited above is mistaken. That is, it is not an example of the "wrong kind of preference" problem. The literature cited by Adler in reference to the "wrong kind of preference" problem (Parfit 1984 (p. 494), Scanlon 1996, Arneson 1999 (p. 124), Darwall 2002 (p. 53)) exclusively concerns *egoistic* preferences whose satisfaction does not lead to well-being. Yet, in his example Adler criticizes that satisfying *moral* preferences does not lead to well-being. Contrary to Adler, Harsanyi (1982) thinks that this is actually desirable. He explains (Harsanyi 1982): "Otherwise [the observer's] assessment will not be a genuine moral value judgement but rather will be merely a judgement of personal preference". That is, satisfying moral preferences should, in fact, not contribute to the well-being of the observer. Furthermore, while Adler presents an account of interpersonal comparisons of well-being for their own sake, Harsanyi's account addresses "moral value judgements." In Harsanyi's account, moral preferences matter by

[6] "Strong non-paternalism says that the well-being ranking of a given subject's histories is identical to the subject's extended preferences over those histories." (p. 155)

definition. Hence, it is not plausible to criticize Harsanyi's account for the inclusion of moral preferences that do not contribute to well-being.

My second point of criticism concerns the use of paternalistic preferences in Adler's account. In Harsanyi's account preferences of individuals are "excluded" if they are irrational (Harsanyi 1982). Chang (2000) considered this to be a "paternalistic intervention to promote a person's own good". Nevertheless, the exclusion of irrational preferences in Harsanyi's account needs to be justified from a consequentialist viewpoint (Birnbacher, unpublished lecture notes). Inherent low value is an insufficient reason and a criticism of Harsanyi's account (Birnbacher, unpublished lecture notes). In contrast to Chang's (2000) interpretation of Harsanyi's account, Adler does not discuss paternalistic preferences in relation to rational preferences. Neither does he discuss paternalistic preferences in relation to moral preferences. Therefore, based on Adler's own presentation, the second point of criticism is unrelated to the first (i.e., the non-use of moral preferences in Adler's account). Specifically, I argue that Adler's account needs to present an underlying coherent theory in the first place that justifies overriding the preferences of individuals. It would have been possible for Adler to make reference to so-called moral paternalism, which intends to promote the moral well-being of a person (Dworkin 2020). However, given Adler's criticism on the inclusion of moral preferences, it would have led to a contradiction. Alternatively, Adler could have reverted to welfare paternalism, which regards for the welfare of another; however, this does not align with the rational preferences by the observer, which do not need additional correction. Therefore, Adler's account faces a dilemma: Either it is based on non-paternalistic preferences only and yields the same result as an empathy-based assessment of outcomes that *matter to the subject*; or it includes paternalistic preferences without independent theoretical foundation. The fact that, from behind the veil of ignorance, non-paternalistic preferences yield the same result as an empathy-based assessment of outcomes cannot, by itself, be a justification for reverting to paternalistic preferences. Reducing the preference exercise to a pure outcome assessment appears to be an ad hoc fix aimed at precluding the problems associated with Harsanyi's account. Instead, there needs to be a coherent theory providing a rational justification for using sympathy and its underlying psychological mechanism as a means of conducting interpersonal comparisons. The theoretical justification for use of sympathy and outcomes would need to be embedded in Harsanyi's axiomatic justification of utilitarianism.

In addition, a sympathy-based account that simply ignores all or only missing attributes runs the risk of causing a bias. The reason is that the assessment of an individual's outcomes can depend on the underlying attributes. That is, our unreserved sympathy for another person's outcomes depends on the available background information on that person. For example, an assessment of the health status of an individual likely depends on the time the individual was living. A functional limitation would perhaps be perceived as less concerning when assessing the profile of a person who lived in earlier times than today. The reason is that today's availability of health technologies and health care resources makes it possible to achieve a different level of health outcome, which could influence the level of sympathy. Therefore, lacking information on birth timing as an attribute could lead to a biased assessment. Hédoin (2021) does not consider this to be a relevant problem because Adler does not "look(..) for uniformity of extended preferences". However, this argument implicitly assumes that biases cancel out across different observers.

A final point of critique on Adler's sympathy-based approach is that it suffers from the same sort of criticism on the vNM utility function as Harsanyi's account. This holds regardless of whether Adler's approach is used to conduct interpersonal comparisons of welfare for their own sake or for the purpose of utility maximization. The only difference is that Adler's comparison is sympathy-based whereas Harsanyi's comparison is empathy-based. According to Weymark (1991), the vNM utility function used in Harsanyi's original model allows only an ordinal ranking of preferences but not an interpersonal comparison of utility where utility satisfies "cardinal unit plus comparability" (a point also stated by Sen (1986)). Yet, only the latter scale is able to capture differences in preference intensity. In fact, this is considered by Broome (2008) to be the "standard objection" to Harsanyi's model.[7] It implies that the social welfare function is linear only in terms of individual von Neumann-Morgenstern utilities but not in terms of welfare (Weymark 1991, p. 313). Therefore, Harsanyi's theorem merely calls for maximizing the sum of von Neumann-Morgenstern utilities of individuals. But it should not be interpreted as a utilitarian theorem or as a support thereof. Harsanyi's suggestion to use conversion ratios in order to establish an interpersonal comparison of preference strengths through a "common utility unit"

[7] In contrast, Hausman (1995) argues that it is unnecessary or even wrong for preference utilitarianism to define an absolute unit of satisfaction. He reasons that an absolute scale would need to capture the impact of changes in the degree of preference satisfaction on mental states. But if preference utilitarianism is a theory about preferences, then the impact on mental states is morally irrelevant.

(Harsanyi 1977a, p. 56) equally fails because the information is not deducible from the vNM utility functions of individuals. Similarly, a ratio scale, which would also be able to establish cardinality but is neither invoked by Harsanyi nor Adler, is not implied by the vNM utility function either and therefore would require to have a different conceptual foundation.

Further development

There have been some recent developments around interpersonal comparisons of absolute utility. Moreno-Ternero and Roemer (2005) have suggested an extension of Harsanyi's model that allows the observer to make welfare interpersonally comparable. In detail, the observer first steps in the shoes of any person i and takes on i's risk preferences and vNM utility function. Then, the observer imagines how i would feel in terms of welfare if she (i) were to be realized as any person j with a given wealth level. Next, the observer converts j's wealth to the welfare-equivalent wealth for i. By taking on every person i's viewpoint the observer has an n number of wealth distributions where n is the total number of individuals. The observer may then take the average of these wealth distributions to assess the utility of an action. The principle idea is thus to convert wealth into utility and utility again into wealth as wealth can be compared across individuals. This conversion is also able to deal with the satisfaction of egoistic preferences that do not contribute to well-being. That is, if satisfying egoistic preferences does not lead to well-being, this will not be captured by the welfare-equivalent measure. If we apply a theory-theory account to this mechanism, we would need to predict other people's welfare-equivalent wealth based on implicit knowledge (as opposed to determining the welfare-equivalent wealth through empathy).

As a final note, given that preferences not only depend on wealth but also on health, longevity, and other factors, the proposal by Moreno-Ternero and Roemer (2005) has been further refined by suggesting the use of life years in perfect utility as a measure of welfare ([Redacted]).

References

1. [Redacted]
2. Adler MD. Extended preferences and interpersonal comparisons: a new account. Economics and Philosophy 2014;30(2):123-162.
3. Arneson RJ. Human flourishing versus desire satisfaction. In: Paul EF, Miller FD, Paul J, ed. Human flourishing. Cambridge: Cambridge University Press; 1999:113-142.
4. Barrett J. Interpersonal comparisons with preferences and desires. Politics, Philosophy & Economics 2019;18(3):219–241.
5. Bergstrom TC. When is a Man's life worth more than his human capital? In: Jones-Lee MW, ed. The value of life and safety. Amsterdam: North-Holland; 1982:3–26.
6. Binmore KG. Rational decisions. Princeton University Press; 2008.
7. Bradley R. Comparing evaluations. Proceedings of the Aristotelian Society 2008;108:85–100.
8. Broome J. Can there be a preference-based utilitarianism? In: Fleurbaey M, Salles M, Weymark JA, eds. Justice, political liberalism, and utilitarianism: themes from Harsanyi and Rawls. Cambridge: Cambridge University Press; 2008.
9. Cambridge Dictionary. Paternalistic. https://dictionary.cambridge.org/de/worterbuch/englisch/paternalistic
10. Carruthers P. Simulation and self-knowledge: a defense of the theory-theory. In: Carruthers P, Smith PK, eds. Theories of theories of mind. Cambridge: Cambridge University Press; 1996.
11. Chang HF. A Liberal Theory of Social Welfare: Fairness, Utility, and the Pareto Principle. The Yale Law Journal 2000;110(2):173-235.
12. Darwall SL. Welfare and rational care. Princeton, NJ: Princeton University Press; 2002.
13. Davidson D, McKinsey JC, Suppes P. Outlines of a formal theory of value, I. Philosophy of Science 1955;22(2):140-160.
14. Davidson D. Judging interpersonal interests. In: Elster J, Hylland A, eds. Foundations of social choice theory. Cambridge: Cambridge University Press; 1986:195-211.
15. Davidson D. Problem of rationality. Oxford: Clarendon Press; 2004.
16. Dworkin, Gerald, "Paternalism", The Stanford Encyclopedia of Philosophy (Fall 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2020/entries/paternalism/>.
17. Gallagher S. The practice of mind. Theory, simulation or primary interaction?. Journal of Consciousness Studies 2001;8(5-6):83-108.

18. Gallese V, Goldman A. Mirror neurons and the simulation theory of mind-reading. Trends in Cognitive Sciences 1998;2(12):493-501.
19. Gandjour A. Interpersonal Comparison of Welfare Based on Harsanyi's Equiprobability Model for Moral Value Judgments. Ethical Perspectives 2021;28(4):385-416.
20. Goldman AI. Simulating minds: the philosophy, psychology, and neuroscience of mindreading. New York: Oxford University Press; 2006.
21. Harsanyi JC. Cardinal utility in welfare economics and in the theory of risk-bearing. Journal of Political Economy 1953;61:434-35.
22. Harsanyi JC. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. Journal of Political Economy 1955;63:309-21.
23. Harsanyi JC. Preferences and utilitarian theory: Some comments. Erkenntnis 1975;13(3):397-399.
24. Harsanyi JC. Rational behavior and bargaining equilibrium in games and social situations. Cambridge: Cambridge University Press; 1977a.
25. Harsanyi JC. Morality and the theory of rational behavior. Social Research 1977b;44:623-656.
26. Harsanyi JC. Morality and the theory of rational behaviour. In: Sen A, Williams B, eds. Utilitarianism and beyond. Cambridge: Cambridge University Press; 1982:39-62.
27. Hausman DM. The impossibility of interpersonal utility comparisons. Mind 1995;104(415):473-90.
28. Hédoin C. Social contract, extended goodness, and moral disagreement. Erasmus Journal for Philosophy and Economics 2021;14(2):25-52.
29. Hoffmann S. Since Children Are Not Little Adults-Socially-What's an Environmental Economists to Do. Duke Environmental Law & Policy Forum 2006;17:209-32.
30. Moreno-Ternero JD, Roemer JE. Impartiality and priority: part 1: the veil of ignorance (2005). Cowles Foundation Discussion Paper No. 1477A. Available at SSRN: http://ssrn.com/abstract=585705
31. Narens L, Skyrms B. The Pursuit of Happiness: Philosophical and Psychological Foundations of Utility. Oxford: Oxford University Press; 2020. |
32. Parfit D. Reasons and persons. Oxford: Clarendon Press; 1984.
33. Mittag HJ, Rinne H. Statistical methods of quality assurance. London: Chapman & Hall; 1993.

34. Rossi M. Transcendental arguments and interpersonal utility comparisons. Economics and Philosophy 2011;27:273-295.
35. Savaki HE. How do we understand the actions of others? By mental simulation, not mirroring. Cognitive Critique 2010;2:99-140.
36. Scanlon TM. The status of well-being (1996). http://tannerlectures.utah.edu/_documents/a-to-z/s/Scanlon98.pdf (accessed March 19, 2020).
37. Sen AK. Social choice theory. In: Arrow K, Intriligator M, eds. Handbook of mathematical economics. Volume 3. North Holland; 1986.
38. Sober E, Wilson DS. Unto others: The evolution and psychology of unselfish behavior. Harvard University Press; 1998.
39. Spaulding S. Mirror neurons are not evidence for the simulation theory. Synthese 2012;189:515-534.
40. Stueber K. Empathy. The Stanford Encyclopedia of Philosophy (Spring 2018 Edition), Zalta EN, ed. https://plato.stanford.edu/archives/spr2018/entries/empathy/ (accessed March 19, 2020).
41. Weintraub R. Do utility comparisons pose a problem? Philosophical Studies 1998;92:307-319.
42. Weymark JA. A reconsideration of the Harsanyi-Sen debate on utilitarianism. In: Interpersonal comparisons of well-being. Elster J, Roemer JE, eds. Cambridge: Cambridge University Press; 1991.