PAPERS on Economics & Evolution

0917

Welfare Notions for Soft Paternalism

by

Till Grüne-Yanoff

The *Papers on Economics and Evolution* are edited by the Evolutionary Economics Group, MPI Jena. For editorial correspondence, please contact: evopapers@econ.mpg.de

ISSN 1430-4716

Max Planck Institute of Economics
Evolutionary Economics Group
Kahlaische Str. 10
07745 Jena, Germany
Fax: ++49-3641-686868

Welfare Notions for Soft Paternalism

Till Grüne-Yanoff
Helsinki Collegium of Advanced Studies, till.grune@helsinki.fi

1. Introduction

Paternalism, the interference with a person, against their will, for her own good, has long been held in disregard. Yet recently, new arguments for paternalistic government interventions have been proposed. These arguments draw on two premises. First, the proposed paternalism is characterised as 'soft', hence distinguishing it from previous 'hard' versions of paternalism. Second, the proposed paternalism justifies interventions using empirical findings of systematic biases and mistakes in reasoning, deliberation and decision-making.

These two premises stand in a certain tension to each other. 'Softness', on the one hand, ultimately rests on an internalist intuition: an adequate account of a person's good must effect a motivational connection between a person and her (non-moral) good. Under the most plausible readings, this implies a preferentialist welfare notion. As - by definition - paternalism is justified as an intervention that improves people's own good,[1] it follows that paternalism is justified with reference to what people prefer, or maybe what they *really* prefer.

The empirical studies of mistakes and biases, on the other hand, show that human decision makers often deviate from standard decision-theoretic accounts. In particular, these studies show that for many situations, decision-makers do not have well-ordered preferences. The tension - what I call the Soft Paternalist's Paradox - lies in the use of a preferentialist welfare notion when preferences are inconsistent.

Defenders of soft paternalism acknowledge this tension. Yet their *ad hoc* proposals for remedy are unsatisfactory, because they drain most of the force out of the general argument for soft paternalism. Instead, I discuss two different solutions to the Soft Paternalist's Paradox. The *weak interpretation* views the tension merely as a problem of preference measurement, and seeks to resolve the tension by finding the correct measurement procedure. Yet sophisticated measurements, I will argue, still require choosing which of the inconsistent preferences to take as welfare-relevant. This selection problem cannot be solved by improved measurements alone, but requires a conceptual analysis of the selection criteria themselves. The *strong interpretation* addresses this selection problem directly, and seeks to resolve the Soft Paternalist's Paradox by finding a preferentialist welfare notion based on *reconstructed* preferences.

[1] Gerald Dworkin, 'Paternalism', in The Stanford Encyclopedia of Philosophy (Winter 2005 Edition), edited by Edward N. Zalta (2005). URL: <http://plato.stanford.edu/archives/win2005/entries/paternalism/>.

Preference reconstruction can proceed along different paths. The *full information* account seeks to reconstruct preferences along a hypothetical ideal of fully informed agents. The full information account is likely to run foul of the internalist intuition again. In contrast, I advocate the *integrity* account, which seeks to reconstruct preferences by eliminating those from inconsistent sets that are least integrated in the agent's 'web of preferences'. This account shows how agents regain preference consistency by revising some of their preferences in a least intrusive fashion. This account, I argue, offers a reconstructed basis for a preferentialist welfare account that respects the internalist intuition and hence can serve as the basis of soft paternalism.

The paper is structured as follows. Section 2 introduces soft paternalism at the hand of three concrete policy cases, and argues that it is characterised by its commitment to internalism, not by its preservation of choice. Section 3 sketches the Soft Paternalist's Paradox, and soft paternalists' replies to it. Section 4 rejects attempts to resolve the paradox by non-preferentialist welfare notions. Section 5 argues against attempts to resolve the paradox by finding new measurement procedures for preferentialist welfare notions. Section 6 discusses the full-information reconstruction attempt. Section 7 presents the alternative integrity account. Section 8 concludes.

2. Soft Paternalism

Many observational and experimental studies give support to the claim that human decision makers often deviate from standard decision-theoretic accounts. From this descriptive claim, some authors draw the normative conclusion that human decision-makers suffer from systematic mistakes and biases disadvantageous to them. This conclusion, they believe, justifies government intervention to help individuals avoid systematically mistaken behavior.

These arguments are exemplified in *Libertarian Paternalism* and in *Asymmetric Paternalism.* Let me illustrate these approaches with three concrete cases.

In the *cafeteria* case, the cafeteria manager has to decide which foods to serve and how to arrange the choices. Studies show that individuals are prone to select items placed earlier and at eye level in a line of food items. Hence, the manager's decision likely has an effect on customers' diet. Sunstein and Thaler argue that she cannot simply decide to give customers what they want, because these wants are influenced by her very decision. Nor should she choose and arrange foods at random. Instead, she should choose what she thinks will make customers better off, all things considered. Intuitively, she should arrange food in such a way that customers preferring to eat healthily find it easy to control themselves, while those set to indulge unhealthy food are able to find it, albeit in more remote places: 'Once the cost of self-control are incorporated into the analysis, we can see that some diners would prefer this arrangement, namely those who would eat a dessert if it were put in front of them but would resist temptation if given a little help'.[2]

[2] Cass R. Sunstein and Richard H. Thaler, 'Libertarian Paternalism is Not an Oxymoron', *University of Chicago Law Review* 70(4) (2003): 1159-1202, p. 1184.

In the *retirement default* case, the employer must decide (or the government must make a regulative decision) whether to automatically enroll employees in a pension savings plan, allowing them to opt out, or to keep employees uncovered, unless they explicitly opt in. Studies show that enrollments significantly depend on the default: under automatic enrollment, enrollment figures were about 80% higher than under non-automatic enrollment. Sunstein and Thaler argue that employers should opt for automatic enrollment (or regulators should mandate them to do so), because 'most employees would prefer to join the 401(k) plan if they took the time to think about it and did not lose the enrollment form'.[3]

In the *home solicitation* case, the government has to decide whether it honours contractual obligations from home solicitation sales, or whether it grants buyers an (non-waivable) right to rescind any purchase within a certain number of days. Studies show that people in transient emotionally or biologically 'hot' states overestimate how long these states will last, and tend to overweigh the short-term benefits of indulging their current state of mind. Camerer et al. argue that the regulator should introduce 'cooling-off periods that force people to delay taking action for some duration – and in particular, allow them to reevaluate their decisions free from heat-of-the-moment impulses'.[4]

In each of these cases, an intervention is called for with the aim of improving the agent's situation. In order to argue that the intervention results indeed constitute an improvement for the agent, the agent's preferences, or hypothetical preferences, or some subset of her preferences are appealed to. For example, the cafeteria case appeals to the preferences of easily tempted diners. The retirement default case appeals to preferences of those employees who properly reflected on the option. The home solicitation case appeals to customers' cooled-off re-evaluations. More generally, libertarian paternalists seek to 'improve the choosers own welfare'[5] and help people 'make choices that are in their best interest or at the very least are better, by their own lights'.[6]

This is in accord with some philosophical notion of 'soft' or 'weak' paternalism.[7] A weak paternalist believes that it is legitimate to interfere with the means that agents choose to achieve their ends, if those means are likely to defeat those ends.[8] The crucial point of soft paternalism thus understood is that such interventions heed a form of *internalist intuition*:

[3] Ibid., p. 1172-3. Cf. also Colin Camerer, Samuel Issacharoff, George Loewenstein, Ted O'Donoghue and Mathew Rabin, 'Regulation for Conservatives: Behavioral Economics and the Case for Asymmetric Paternalism', *University of Pennsylvania Law Review*, 1151(3) (2003): 1211-1254, p. 1127-1230.
[4] Ibid., p. 1239.
[5] Sunstein and Thaler, 'Libertarian Paternalism is Not an Oxymoron', p. 1162.
[6] Ibid., p. 1163.
[7] Not, however, with all such notions. John Stuart Mill, for example, held the view – ofte called Soft Paternalism - that the only conditions under which state paternalism is justified is when it is necessary to determine whether the person being interfered with is acting voluntarily and knowledgably (cf. J. S. Mill 1859, *On Liberty*, Indianapolis: Bobbs-Merrill, 1956).
[8] Dworkin, 'Paternalism'.

> **(II)** It is a necessary condition for something to be good for a person that she is capable of caring about it.[9]

Such capability requires some pro-attitude towards this thing, which under appropriate ideal conditions would manifest as an all-things-considered preference. By (implicitly) respecting this intuition, soft paternalism is defined considerably narrower than hard paternalism, and is intuitively more palatable to many.

Note that the above interpretation of soft paternalism appeals to a property of the person or action interfered with. In contrast to this, there is another interpretation that both libertarian and asymmetric paternalism also use. It appeals to properties of the interference, not the person or action interfered with.

> Libertarian paternalism is a relatively weak and nonintrusive type of paternalism, because choices are not blocked or fenced off. In its most cautious forms, libertarian paternalism imposes trivial costs on those who seek to depart from the planner's preferred option.[10]

According to this view, interventions are 'soft' in the sense that they do not prohibit choice, or restrict people's options. I think this is a red herring. It is rarely possible for the government to intervene in a way that actually eliminates a choice option. Instead, almost all regulative measures focus on making 'bad' options more costly, and 'good' options less costly. Whether these costs are incurred via financial incentives or prison sentences is – at best – a matter of degree, not of qualitative difference.

Further, the claim that these costs are 'trivial' is hard to defend. Every effective paternalist intervention, while it may differ in the absolute cost it imposes, will be identical in the relative cost it imposes. If the government intends to prevent an individual from choosing in a certain way, it has to shift the individual's cost-benefit balance in such a way that the choice becomes not beneficial in the eyes of that individual. For example, a policy designed to effectively prevent someone from overeating may impose a cost that seems 'trivial' in absolute terms; a policy to prevent someone from committing a crime difficult to detect but very gainful to the criminal may have to impose much higher costs. Yet in either case, an effective policy must impose a cost that *effectively* interferes with the agent. Thus, relative to the behavioural impulse it seeks to counteract, the imposed costs must be the same for all sorts of effective paternalist intervention. From this relative perspective the distinction between 'soft' and 'hard' intervention collapses.

Libertarian and Asymmetric Paternalism are somewhat ambiguous about what their distinguishing characteristics are. In the light of the above argument, I will concentrate on the first meaning of 'soft' in soft paternalism, and discuss to what extent one can hope to

[9] David Brink, Moral Realism and the Foundations of Ethics (Cambridge: Cambridge University Press, 1989), p. 40. Stephen Darwall, Impartial Reason (Ithaca: Cornell University Press, 1983), p. 54-55.
[10] Sunstein and Thaler, 'Libertarian Paternalism is Not an Oxymoron', p. 1162.

find a welfare measure that satisfies the internalist intuition. That this is not an easy feat I will show in the next section.

3. The Soft Paternalism Paradox

There is a tension between endorsing the internalist intuition and relying on the empirical claim that human reasoning and deciding is often biased and mistaken. Respecting internalism not only requires a commitment to people's relevant pro-attitudes, it also requires that the soft paternalist finds procedures to measure welfare in terms of these pro-attitudes. The possibility of such a measurement necessitates that all the relevant pro-attitudes are consistent. For example, if an agent has a circular preference ordering over options {*A*,*B*,*C*} where she prefers *A* to *B*, *B* to *C*, yet *C* to *A*, it is impossible to say which option is best for her, and even difficult to say which option is better for her in an unambiguous sense.[11] If one can argue that only *A* and *B* are welfare-relevant, one could focus on the subset {*A*,*B*}, thus yielding *local consistency*.

Local consistency of all welfare-relevant pro-attitudes and their associated choices is a minimal condition for the measurability of welfare. Yet such local consistency is often not satisfied, as a look at the three cases from section 2 shows. In the cafeteria case, diners' choices depend on the arrangement of food items. They eat a dessert when they find it right at the cash register, but they forgo it when desserts are on offer behind the salad bar. In the retirement default case, employees' choices depend on the default option. Many more will choose enrolment if this is the default option than if it is not. In the home solicitation case, consumers choose differently upon reflection than on their first instincts. Many will return the gadget they bought on a whim, if they have a right to do so. If we can believe these studies, each of these choices violates WARP, the weak axiom of revealed preferences.[12] If one tried to rationalize these choices, the resulting (strong) preference order would not only be cyclical, but symmetric: a preference for dessert/enrolment/gadget inferred from these choices would always be directly contradicted by a preference against dessert/enrolment/gadget. These rationalizations do not yield even local consistency, and hence cannot be used to measure the individual's welfare.

The tension between the two premises thus leads to the *Soft Paternalism Paradox*: soft paternalism distinguishes itself from other forms of paternalism by stressing its respect for the internalist intuition, but its empirical antecedents often imply a very fundamental form of preference inconsistency that makes the measurement of an internalism-respecting notion of welfare impossible. By using a preferentialist welfare notion when

[11] I am using preferences as a special kind of pro-attitudes here and elsewhere in the paper for illustrative purposes. Other kinds of pro-attitudes are subject to different consistency requirements.

[12] WARP requires that if *x* is chosen when *y* is available, then there can be no budget set containing both alternatives for which *y* is chosen and *x* is not. Thus, asymmetry of the revealed preference relation $>_C$ is secured.

preferences are inconsistent, soft paternalism at best seems ineffective, and at worst incoherent.

Defenders of soft paternalism are aware of this paradox. They acknowledge that in many situations in which they advocate paternalistic intervention,

> people lack clear, stable, or well-ordered preferences. What they choose is strongly influenced by details of the context in which they make their choice… These contextual influences render the very meaning of the term "preferences" unclear.[13]
>
> If the arrangement of the alternatives has a significant effect on the selections the customers make, then their true "preferences" do not formally exist.[14]
>
> …in the contexts in which such studies are used, people do not have clear or well-formed preferences, and hence it is unclear that people have straightforward "values" that can actually be found.[15]

Yet the implications of these observations are not clearly investigated. Some authors seem to conclude that because individuals do not have consistent preferences over the welfare-relevant options, a paternalistic regulator can impose its own welfare criteria. Yet at the same time, these authors feel still bound to the internalist intuition. This yields a rather eclectic mix of context-sensitive welfare criteria that – so the authors speculate - people would possibly accept if they were more consistent, more reflected, or more knowledgeable. Amongst the more prominent suggestions, one finds the following:

- Select the approach that the majority would choose if explicit choices were required and revealed.[16]
- Select the approach that minimizes the number of opt-outs.[17]
- Material payoff of an activity, or expected payoffs.[18]
- Privilege certain temporal perspectives. For example, they judge higher 401(k) participation beneficial, because of 'people's self-reports that they save less than they would like'.[19]

While some of these options may be reasonable guesses at welfare criteria that satisfy the internalist intuition, such an approach drains most of the force out of the argument for soft paternalism.

General arguments for paternalism depend on a systematic notion of welfare. Once that systematic basis is created, it is then a second and more context sensitive task to measure

[13] Sunstein and Thaler, 'Libertarian Paternalism is Not an Oxymoron', p. 1161.
[14] Ibid., p, 1164.
[15] Ibid., p. 1177.
[16] Ibid., p. 1195.
[17] Ibid., p. 1196.
[18] Ibid., p. 1127.
[19] Ibid., p. 1127.

the potential welfare gains or losses for specific situations. Yet the above authors propose to replace the systematic notion of welfare with proposals that can only be applied in particular circumstances, hence mixing the question of welfare concept and welfare measurement. These *ad hoc* proposals leave little room to argue for or against soft paternalism in general, and hence are unsatisfactory for the present debate.[20] Instead, a solution to the Soft Paternalism Paradox can be found only by offering a welfare notion that does not conflict with the preference inconsistencies found in the relevant cases. I will discuss three different directions for such a search in the next section.

4. Resolving the Soft Paternalism Paradox

Maybe the most obvious response to the Soft Paternalism Paradox is to avoid a preferentialist welfare notion altogether. If preferences are inconsistent, defenders of this approach suggest, there is no point in trying to establish a welfare measure on their basis. Indeed, for reasons quite independent of the Soft Paternalism Paradox, such positions have been explored and advocated. Contemporary ethics standardly makes a tripartite distinction between preferentialist accounts of welfare, hedonist accounts according to which well-being consists in the greatest balance of pleasure over pain, and objective list theories, which list items constituting welfare consisting neither merely in pleasurable experience nor in preference-satisfaction.[21]

Yet as a means to rescue soft paternalism, these alternative welfare notions are of little use. This is obvious for objective lists accounts: they explicitly include items that the individuals concerned do not value. Thus, objective list accounts, for all the virtues they may have, do not satisfy the internalist intuition and hence let the soft paternalism project collapse into standard, hard paternalism.[22]

One might think that hedonist accounts avoid this problem, as pleasure and pain are psychological states and hence 'internal' to the individual whose welfare we are interested in. However, hedonist accounts give rise to welfare notions that occasionally conflict with preferentialist notions. This is presumably intended by the defenders og hedonistic accounts, as they see preference satisfaction as a mere means – and a fallible at that – for the ultimate end of happiness. Yet when one inspects some of these cases of conflict, it seems that people often do *not* regard preference satisfaction as a 'fallible means'.

For example, measuring the level of felt satisfaction of married couples across child-

[20] Soft paternalists, after all, do not merely argue for paternalistic interventions in specific situations. Rather, they push an 'anti-anti-paternalism' agenda, which seeks to show that certain *kinds* of paternalistic regulations are permissible. The stress here is on kinds, which requires a generalising argument, not one remaining in the particular.

[21] For an overview, see James Griffin, Well-Being: Its Meaning, Measurement, and Moral Importance (Oxford: Clarendon Press, 1986).

[22] The exception are lists, which include items that are in accord with the internalist intuition, e.g. autonomy. I will return to this later in this section.

rearing periods, various separate studies show that marital satisfaction decreases dramatically after the birth of the first child and increases only when the last child leaves home.[23] Further, taking care of their children brings women significantly less enjoyment than does sex, socializing, eating, exercising, shopping, napping and watching television.[24] Child-raising repeats itself from generation to generation; presumably, its impact on happiness is part of folklore. Yet why people then not respond to this information and correct their 'mistaken' preferences in such a way as to maximise happiness? The plausible answer is that parents seek to realise a plan, a value or a desire for bringing up children, and that this seeking is not seriously undermined by feelings of frustration or exhaustion, which in turn may lead to reports of diminished satisfaction. It is these plans, values or desires, and not felt satisfaction, that parents are seen to care about and which therefore explain why they continue having and raising children, even if it may not be best for them according to the hedonic welfare notion.

I do not know whether the above constitutes a case for paternalistic intervention on the hedonist account. But any such intervention, it seems, would have to go against what people care about. Hence, despite being a psychological state, hedonic measures do not satisfy the internalist intuition and again let the soft paternalism project collapse into standard, hard paternalism.

Thus, both hedonic and objective list accounts, despite their many advantages, are not useful for the rescue of soft paternalism, as they turn this project into something that its defenders have explicitly wished to differentiate it from.

A conclusion to similar effect must be drawn with respect to an alternative welfare criterion suggested by economist Robert Sugden.[25] He seeks a welfare criterion that does not require preference consistency. His basic intuition is that it is good for an individual to have a wide range of alternative options from which to choose, whether or not her choices reveal any internally consistent set of judgments about welfare.[26] By inscribing this concept of consumer sovereignty into his account of welfare, Sugden *a fortiori* satisfies the internalist intuition. Yet by insisting on consumer sovereignty, he also denies the soft paternalist any ground for action. By definition of the opportunity criterion, there cannot be any intervention in people's decisions that would improve their welfare. Again, instead of rescuing soft paternalism, the proposed alternative welfare notion deflates the soft paternalism project.

In the absence of any other welfare concept that may do the job, I conclude from the above that this avenue is not likely to rescue soft paternalism. Rather, any attempt to

[23] For a recent overview over this research, see J. M. Twenge, W. Campbell W and C. A. Foster, 'Parenthood and marital satisfaction: A meta- analytic review'. *Journal of Marriage & the Family* 65(3) 2003: 574–583.
[24] D. Kahneman, Krueger, A. B., Schkade, D. A. Schwarz, N., & Stone, A. A. 'A survey method for characterizing daily life experience: The day reconstruction method', Science 306 (2004): 1776- 1780.
[25] Robert Sugden, 'The Opportunity Criterion: Consumer Sovereignty Without the Assumption of Coherent Preferences', American Economic Review, 94(4) (2004): 1014-1033.
[26] Sugden's account therefore I akin to an objective list account that stresses autonomy. Cf. footnote 21.

replace the preferentialist notion of welfare for the advance of soft paternalism is likely to throw the baby out with the bathwater. Instead, I will now turn to approaches that stick with the preferentialist welfare notion but seek to make it operational under the empirical antecedents of soft paternalism. Two kinds of approaches must be distinguished.

The *weak interpretation* interprets the Soft Paternalism Paradox as merely a problem of measurement. The welfare-relevant preferences are out there, they just are not properly measured with the current techniques. If the measurement procedure is corrected, then the Soft Paternalism Paradox is solved. No actual or hypothetical changes in the individual's preferences are required for this resolution of the paradox.

In contrast, the s*trong interpretation* interprets the paradox as a problem of the preference state itself. (Locally) inconsistent preferences cannot be welfare-relevant, hence the paradox can only be resolved by *reconstructing* preferences in such a way that preferentialist welfare measure is possible. I will discuss these strategies in the next three sections.

5. The Weak Interpretation

Not all of an individual's preferences are necessarily welfare-relevant preferences in the sense that a social planner concerned about the individual's good has to take them into account. Hence, there may be cases in which an individual exhibits inconsistent, cyclical choices – yet a meaningful preferentialist welfare index can still be derived from them. This is the intuition behind a number of recent papers that propose alternative methods to rationalise seemingly irrational choice – i.e. that derive consistent preference orderings from choices violating basic properties like WARP. Consistency here is usually understood as the minimal condition of acyclicity of strong preference.[27] These approaches all treat the Soft Paternalism Paradox as a measurement problem, and hope to resolve it by suggesting alternative rationalisations.

Two strategies can be distinguished in the recent economic literature. The first is to differentiate choice situations in terms of behaviour-affecting but welfare-irrelevant conditions, and identify *unambiguous choices* across such choice situations, from which consistent preferences can be derived. The second is to posit that individuals choose with reference to *multiple rationales*, and rationalise choices with their help. I will discuss some exemplars of each strategy.

As an aexample of the first strategy, Bernheim and Rangel replace the standard revealed preference relation with an unambiguous choice relation: roughly, x is (strictly) unambiguously chosen over y (written xP^*y) if y is never chosen when x is available,

[27] Kotaro Suzumura, 'Rational Choice and Revealed Preference' Review of Economic Studies 43(1) (1976): 149-58.

under all ancillary conditions d.[28] Ancillary conditions are features of the choice environment that may affect behavior, but are not taken as relevant to a social planner's evaluation. Under weak assumptions, P^* is acyclic and therefore suitable for welfare analysis; it is also the most discerning welfare criterion that never overrules choice. The authors suggest that the domain of unambiguous choice relation can be extended by reducing the number of ancilliary conditions considered. Choices under certain conditions may be less relevant for welfare than others; by 'pruning' these conditions from the welfare relevant-domain, 'the remaining choices coherently reveal "true" objectives'.[29]

As an example of the second strategy, Salant and Rubinstein propose an *extended choice function* C_c, which assigns a chosen element to every pair (A, f) where A is a set of alternatives, and f is a frame.[30] A frame includes observable information that is irrelevant in the rational assessment of the alternatives, but affects choice as a result of procedural or psychological factors. They show that C_c can be rationalized by some asymmetric and transitive relation $>$ iff (i) for every frame f, there exists an ordering $>_f$ such that $c(A,f)$ is the $>_f$ –maximal element in A, and (ii) if $c(A,f)=x$ and $c(B,g)=x$, then there exists a frame h such that $c(A\cup B,h)=x$.[31]

The extended choice function rests on the notion of multiple rationales.[32] A rationale is an ordering that rationalizes one or several choice sets. A tuple of orderings $[>_1,\ldots>_k]$ on X is a rationalization by multiple orderings if for every $A\in P(X)$, the choice function $c(A)$ gives the $>_k$-maximal element of A for some k. Such a procedure of course rationalizes every behavior, as in the extreme there is a different rational for every choice set. However, the authors propose to focus on those rationalizations that employ the minimal number of rationales: 'the larger [the number of rationales], the less meaningful is the rationalization by multiple rationales that can be given to c'.[33]

Manzini and Mariotti also posit that decision makers employ multiple rationales, but suggest a sequentially rationalizing procedure.[34] According to their method, the first rationale identifies a shortlist of candidates (which tie or are incomparable according to this rationale), from which the second rationale selects. Necessary and (combinedly) sufficient conditions for sequential rationalization are (i) a weakened version of WARP[35],

[28] B. Douglas Bernheim and Antonio Rangel 'Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics', Quarterly Journal of Economics 124(1) (2009): 51–104.
[29] Ibid., 55.
[30] Yuval Salant and Ariel Rubinstein, '(A, f): Choice with Frames', Review of Economic Studies 75(4) (2008): 1287-1296.
[31] Salant and Rubinstein are sceptical whether this rationalization is welfare-relevant.
[32] G. Kalai, Ariel Rubinstein and R. Spiegler, 'Rationalizing Choice Functions by Multiple Rationales', Econometrica 70 (2002), 2481-2488.
[33] Ibid., 2482.
[34] Paola Manzini and Marco Mariotti, 'Sequentially Rationalizable Choice' American Economic Review 97(5) (2007): 1824–1839.
[35] Standard WARP requires that if x is chosen when y is present, then y is never chosen when x is present. This prohibits any influence of the availability of other options on the choice, although such availability sometimes has informational contents (A. Sen, 'Internal consistency of choice'. *Econometrica* 61 (1993)

and (ii) that alternatives chosen from each of two sets are also chosen from their union. The authors expand their model to finitely many sequences, and show that even then certain simple choice functions may not be rationalizable. Nevertheless the number and order of sequences has effects on whether and how choice is rationalized.

Green and Hojman also develop an account of multiple rationales.[36] They conceptualize these rationales (which they call population of motivations) as a probability distribution λ over all strict orders on the set of alternatives. An explanation (viz. rationalization) of a choice function c is a pair (λ, v) consisting of a population of motivations λ and a voting rule v such that $c(A) \in v(\lambda, A)$ for all A of the domain of choice situations. Voting rules v assign scores to items of A by forming the weighted sum of the rank of this item in different orders. The result of a 'vote' under v by a population λ when the set of alternatives is A is the set of alternatives that receives the highest sore. Ordinal welfare conclusions are derived by identifying unambiguous votes when the set of alternatives is changing. Cardinal welfare conclusions are computed as the sum of utilities for an item across different orders, weighted by the probability distribution λ. This cardinal result is interpreted as a compromise among a set of simultaneously-held, conflicting preference relations. Because the form of neither λ nor v are limited, any choice function, no matter how irrational, can be rationalized by this method.

Surveying these various approaches, the results are rather disappointing, especially when confronted with the paradigmatic cases for soft paternalistic interventions. First, some of the above approaches are not able to rationalize every choice function. For some purposes, this may be helpful, as it secures the testability of the proposed choice correspondences. But for the purpose at hand, it turns out that choice functions that gave rise to the Soft Paternalism Paradox cannot be rationalized. According to Bernheim and Rangel, the choice functions associated with the cafeteria, retirement default and home solicitation cases do not yield an unambiguous choice. Hence no "true value" can be recovered. Similarly for Manzini and Mariotti, as for each of the paradigmatic cases, one rationale excludes the option that the other rationale would favour. Finally, in Kalai's et al. approach, as well Salant and Rubinstein's, one is left with a different rationale for each choice situation, where each of the rationales contradict each other. With dessert in front of her, the diner prefers to have a dessert, while with dessert further away, she prefers not to have it. Similar for the retirement default and the home solicitation cases. Identifying these rationales represents the conflict now as a conflict of incompatible rationales, but how could one derive a welfare criterion from that?

In each of these cases, additional criteria or principles must be appealed to in order to resolve conflicts or make rationales compatible. For example, Bernheim and Rangel suggest to prune the less welfare-relevant choice situation from the general choice set, hence yielding an unambiguous choice. But which choice situation should be pruned?

(3): 495–521). Manzini and Mariotti's weakened WARP allows such 'menu effects', but requires that if large menus have no such effect, then subsets of these menus do not either.

[36] Jerry R. Green and Daniel A. Hojman, 'Choice, Rationality and Welfare Measurement', Harvard Institute of Economic Research Discussion Working Paper Series, No. 2144 (2007).

Intuitively, one may say that the dessert-at-register, no-enrolment-default and initial-purchase situations should be pruned, but this of course begs the question: all the relevant normative content lies in the intuition, which is external to the approach.

In Mariotti's and Manzini's case, one may switch the sequences or add furher ones, but again, according to what principle?

Finally, Green and Hojman offer a utilitarian-style 'compromise' among a set of simultaneously-held, conflicting preference relations. But why would one subscribe to such a principle? Would we really say that someone conflicted over having dessert or not is best off by eating half the pudding? Should those susceptible to enrollment defaults be furnished with a smaller pension scheme? Or should those who later want to return a gadget bought on the whim be left with a smaller gadget? I take it that these compromises violate intuitions about what paternalistic improvement can achieve, and hence do not offer a solution to the problem either.

I therefore conclude that the Soft paternalism Paradox cannot be resolved by finding the right measurement procedure for welfare-relevant preferences 'out there'. The surveyed approaches either do not provide a useful welfare criterion at all, or they require assumptions external to the measurement procedure itself. Yet these conceptual assumptions remain hidden in the background, largely ignored. I therefore now turn to the strong interpretation, which addresses these conceptual issues directly, suggesting to *reconstruct* preferences in such a way that preferentialist welfare measure is possible.

6. The Strong Interpretation

The idea that preferences have to be (re-)constructed to obtain meaningful welfare measurements is not new. There exists a substantial psychological literature suggesting that preferences are invented rather than found,[37] more like the product of architecture than of archaeology.[38] This literature focuses on contingent evaluation methods of preference elicitation, where the questioner has various methods at hand to influence the preference expression of the questioned. For example, she may encourage him to consider alternative options and scenarios to counteract certain frames; she may provide him with relevant information and present this information in vivid ways in order to counteract inappropriate selectivity or incomprehension; or she may use explicit scale anchors or more robust scales to avoid biases in scale usage.[39] The goal of these constructive efforts is similar to those of libertarian or asymmetric paternalists:

[37] E. J. Johnson, M. Steffel and D. G. Goldstein, 'Making better decisions: From measuring to constructing preferences', Health Psychology 24 (2005): S17–S22.
[38] R. Gregory, Sarah Lichtenstein and Paul Slovic, 'Valuing environmental resources: A constructive approach' Journal of Risk and Uncertainty 7 (1993): 177–197. J. W. Payne, J. R. Bettman and D.A Schkade, 'Measuring constructed preferences: Towards a building code', Journal of Risk and Uncertainty 19 (1999): 243–270.
[39] Payne et al. 'Measuring constructed preferences', pp. 250ff.

determining policy objectives by people's own values, but seeing to it that people express their 'real' or 'true' or 'welfare relevant' preferences.

> The more that measured preferences are to play a role in an important decision, e.g. a public policy decision, the greater the weight that should be given to the better constructed preferences.[40]

Yet there is an important ambiguity in how this re-construction should take place. On the one hand, a person's preferences may be developed with the help of additional information. This follows a longstanding intuition that better-informed preferences more accurately reflect an agent's welfare. However, I will argue that this 'informed preference' approach flies in the face of soft paternalism's commitment to internalism. On the other hand, her preferences may be made more consistent, particularly identifying some of her preferences as core values and central commitments. This 'integrity account' I will argue, is the best available compromise between respecting internalism and constructing a preference ordering useful for soft paternalists.

John Harsanyi formulated the basics of the informed preference account thus:[41]

> Any sensible ethical theory must make a distinction between rational wants and irrational wants, or between rational preferences and irrational preferences. It would be absurd to assert that we have the same moral obligation to help other people in satisfying their utterly unreasonable wants as when we have to help them in satisfying their very reasonable desires…a person's true preferences are the preferences he *would* have if he had all the relevant factual information, always reasoned with the greatest possible care, and were in a state of mind most conducive to rational choice.[42]

This normative judgment seems intuitively plausible, and compatible with the internalist intuition. It also is close to the soft paternalist position, as it opens space for improving people's welfare while still respecting their preferences, in some sense. Indeed, soft paternalists sometimes allude to this approach:

> In some cases individuals make inferior decisions in terms of their own welfare—decisions that they would change if they had complete information, unlimited cognitive abilities, and no lack of self-control.[43]

Note that his approach is hypothetical. It does not seek to provide actual information on which people then may refine their preferences. Instead, it judges preferences to be irrational and hence welfare-irrelevant if people changed them, would they have more

[40] Ibid., p. 265.
[41] Related views have been proposed by moral philosophers Richard Brandt and Peter Railton. Because social scientists will be most likely familiar with Harsanyi's account, I will refer to it here.
[42] John Harsanyi, 'Morality and the Theory of Rational Behaviour', in Utilitarianism and Beyond, edited by Amartya Sen and Bernhard Williams (Cambridge: Cambridge University Press, 1982): 39 – 62, p. 55.
[43] Sunstein and Thaler, 'Libertarian Paternalism is Not an Oxymoron', p. 1162.

information or better reasoning faculties. Further, it replaces these irrational preferences with the counterfactual preferences under ideal informational and cognitive conditions. Soft paternalism, if adopting this strategy, would determine people's welfare by reference to the preferences they *would* have, *were* they perfectly informed and endowed with perfect cognitive capacities.

At the very least, to determine what a person would prefer under ideal circumstances requires understanding what it means that a person is fully informed.[44] However, as I will show, this notion stands in conflict with the internalist intuition.

To be informed is not just to be exposed to information, but to properly appreciate it. Richard Brandt describes this appreciation as that the agent

> gets the information at the focus of attention, with maximal vividness and detail, and with no hesitation or doubt about its truth.[45]

For example, the smoker cannot merely be provided with information about the dangers of smoking. Rather, the information must be vividly and repeatedly represented to her at times when reflection of the bad effects of smoking is more likely to influence her current desire – maybe just after inhaling, so that the reflection destroys any pleasure she would normally get from the cigarette. [46]

Full information accounts require that one considers the person to appreciate all relevant information simultaneously, and forms her preferences accordingly. But is such a consideration conceptually coherent? We ordinarily think that any particular person will be unable to appreciate certain facts, given what she is currently like. Because of her intellectual and psychological features, she occupies a point of view, a perspective from which she views the world and which determines what can be informing for her. Thus, certain information cannot be appreciated by certain persons, due to their personality. To overcome such barriers, these persons have to undergo education or certain experiences. Crucially, such barriers may be overcome only if a person changes prior to or due to this

[44] Although I cannot fully argue for it here, it is plausible that one may make the preferences of an agent less welfare relevant by providing her with more, but not with full information. Imagine that a person prefers the 10.30 bus to the 10.40 train, because she believes that the bus will get her to her destination earlier. However, the train actually leaves at 10.35, and the bus takes 10 minutes longer than the train. By informing her of the latter, but not the former, we (i) provide her with more information, and (ii) make her preferences less welfare relevant. She will now prefer taking the 10.40 train, which means – if she acts on the preference – that she'll miss both train and bus. Thus, the full information account has a chance to work only if we consider provision with *full* information – incremental increases of information below full information may have the opposite effect.

[45] Richard Brandt, A Theory of the Good and the Right (Oxford: Oxford University Press, 1979).

[46] Soft paternalists are quite aware of the difference between information and appreciation. Indeed, getting drastic messages on cigarette packages, appealing to visceral reactions through 'shocking' images of open heart surgery, or conveying information in vivid ways (cf. 'to win on this ticket is a likely as being hit by lightening over the course of the next week', Camerer et al. 'Regulation for Conservatives', p. 1231) are part of their policy repertoire.

education or these experiences; she overcomes them by changing her personality quite dramatically.

These considerations have led some to argue that there is an inherent tension between what it is like to be a particular person and what is required in order for a person to be fully informed.[47] This led to the view that it is conceptually impossible that any person is able to appreciate the full range of all relevant information for her possible future lives.

I think this is an important argument against the full information account. However, I will make use of a weaker version of this argument, stressing only the need to change one's personality quite dramatically in order to appreciate some information fully. In Stigler and Becker's example, a person changes her preferences when systematically exposing herself to music in a way similar to the substance abuser who exposes himself to heroin.[48] Imagine that such changes would be counterfactually performed for all your relevant preferences. Each one would be refined with all the relevant information, so as to see what your refined self would prefer. Would it then not be surprising if the well-being of the two of you, your informed self and your ordinary self, consisted in different things? Although purportedly you, the refined "you" may not be someone whose judgments you would recognize as authoritative. You may well just not care about the preferences and values this refined "you" holds.

Put differently, if a policy maker bases her policies on preferences thus reconstructed, she could well encounter an outcry of public protest. If a soft paternalist sought to nudge people to do things that such a refined person would prefer, he may reasonably be accused of acting as a hard paternalist, imposing values on people that they do not share. Thus, the full information account of preferences violates the internalist intuition, and hence is incompatible with soft paternalism.

7. The Integrity Account

Vicarious policy makers or soft paternalists could defend themselves against such charges by pointing out that the reconstructed preferences on which the policy is based are sufficiently close to the person's actual preferences that this person must reasonably care about them. As I argued above, such closeness is not guaranteed by the full information approach. Citizens may reasonably claim that they do not care about modern art, and hence do not want to support it with their tax money, even if they would desire seeing this art being made, had they appreciated all relevant available information. Their hypothetically refined selves might just not close enough to their actual preferences that they would have any rational obligation to care about those refined preferences. The full-

[47] Connie S. Rosati, 'Persons, Perspectives, and Full Information Accounts of the Good', Ethics 105(2) (1995): 296-325, p. 317.
[48] George J. Stigler and Gary S. Becker, 'De Gustibus Non Est Disputandum', The American Economic Review, 67(2) (1977): 76-90.

information account permits too many hypothetical changes of individuals to sufficiently respect the internalist intuition.

Yet in other cases, where an agent holds conflicting preferences *P* and ¬*P*, the semantic content of her *other* preferences may justify the policy maker to reconstruct her preference ordering in such a way that she would hold *P* but not ¬*P*. Take for example a person who is generally very cautious to prevent bodily harm to herself, but who has conflicting preferences about wearing a helmet when bicycling (for example, she may hold a preference for wearing when she envisions consequences of an accident, but hold a preference against wearing when envisioning herself with a helmet). Her general risk preference is manifest in various concrete preferences, expressed in her automotive behaviour, work choice, travelling preferences, and her attitudes towards adventure sports. From these behaviours and expressed attitudes, her preferences for a certain cost-to-risk trade-off could be quantified. Now she may have other reasons not to wear a bicycle helmet, for example out of fashion concerns. One may be able to determine how central these concerns are, and hence also quantify her actual preferences for the cost-to-risk trade-off in the bicycle helmet case. Let's assume that when comparing this particular trade-off with her general preference, the two contradict each other. Thus her preferences would not be very well integrated, in the sense that some of her preferences over token states contradict her preferences over type states.[49] The policy maker then could say that (i) by her own reasons, as expressed in other preferences, she should prefer wearing a helmet, and (ii) her reason not to wear it, e.g. out of fashion concerns, is not central enough to stand against her reasons for safe behaviour, and hence should not stand up again wearing the helmet.

This form of reconstruction still allows the confrontation of individuals with correct information. Yet only such information is allowed which facilitates the derivation of preferences from other preferences, or which increases preference consistency. Thus, the integrity account is a sub-case of the full-information account.

To make sense of the idea that some preferences can be more integrated than others, I refer to the idea of the integrity of a person. I understand integrity in terms of the preferences that people identify with most deeply, as constituting what they consider their life is fundamentally about. Preferences of this kind are called 'identity-conferring commitments' or sometimes 'ground projects'. The idea is that for people to abandon an identity-conferring commitment is for them to lose grip on what gives their life its identity, or individual character. An identity-conferring commitment is 'the condition of my existence, in the sense that unless I am propelled forward by the conatus of desire, project and interest, it is unclear why I should go on at all'.[50]

[49] Cf. Luc Bovens, 'Sour Grapes and Character Planning', Journal of Philosophy 89(2) (1992): 57-78, for a discussion of this semantic approach in the context of adaptive preferences.
[50] Bernard Williams, 'Persons, Character and Morality', in Moral Luck: Philosophical Papers 1973–1980 (Cambridge: Cambridge University Press 1981): 1–19, p. 12.

Importantly, I propose that the notion of identity-conferring commitment is determined *endogenously*. A certain preference becomes identity-conferring if it relates to many other preferences, some of which may be identity-conferring as well. In analogy to Quine's idea of a web of belief, I propose a *web of preference*.[51] In such a web, some preferences are more central than others, in the sense that more central ones are more connected than others. For example, in the helmet-wearing case, the pro-wearing preference was connected to a type preference for cautious behaviour, which in turn was connected to many token preferences about travel, work, sports, etc. The contra-preference, however, was connected to a type preference for fashionable appearance, but this in turn was connected to only a few token preferences. Thus the cautious-behaviour preference would be more central than the fashionable-appearance preference, conferring a higher degree of integration on the pro-wearing preference than on the contra-wearing one.

Centrality thus defined is a relational concept that can be quantified in degrees. A core preference has a high centrality, and a peripheral preference low centrality. Centrality influences change dynamics in the following way: preferences near the edge of the web are more susceptible to change in the light of new judgments, emotional experiences or experiences from pervious choices. Preferences near the centre are harder to dislodge. Change in such a network proceeds (i) by gradually reducing the centrality of core preferences, through the elimination of some of the preferences they connect to, and (ii) by the gradual increase of the centrality of some other preferences, though addition of connections to other preferences. Thus, what counts as an core may be in continuous flow, determined through changes in the peripheral preferences and their connections to other core preferences. No external criterion for what counts as central is needed.

Formal models of preference dynamics can accommodate the idea of a web of preferences, and that allow for a formal definition of centrality.[52] The basic idea is that the connections between preferences consist in the logical relations of the preferences' intentional content. When agents adopt or develop new preferences (in processes initiated from external sources, e.g. fashions, aging, biases, etc.) these preferences are accommodated in the existing web. This may mean that certain other preferences must be removed, in order to retain consistency. Adoption of new preferences need not be unconditional, as it may turn out that new preferences may conflict with present core preferences. Depending on the respective degrees of centrality, either new preferences may be rejected, old peripheral preferences may be removed, or even old core preferences may be demoted to peripheral preferences and eventually removed.

[51] W. V. Quine and J. S. Ullian, The Web of Belief. (New York: McGraw-Hill, 2nd edition, 1978). Cf. also the notion of 'web of self' in Robert Noggle, 'Integrity, the Self, and Desire-Based Accounts of the Good', Philosophical Studies 96(3) (1999): 303-331, pp. 318-21.

[52] Sven-Ove Hansson, 'Changes in Preferences', Theory and Decision 38 (1995): 1-28. Richard Bradley, 'The Kinematics of Belief and Desire' Synthese 156 (2007): 513-535. Till Grüne-Yanoff and Sven-Ove Hansson, 'From Belief Revision to Preference Change' in Modelling Preference Change. Approaches from Philosophy Economics and Psychology edited by Till Grüne-Yanoff and Sven-Ove Hansson (New York: Springer, Theory and Decision Library, 2009): (pp. 1-27).

Preference reconstruction may proceed with the help of such models in the following way. First, the policy maker identifies an agent's (inconsistent) preference orderings over all relevant options. Second, she determines the centrality of each preference. Third, she sketches the various ways the preference ordering could be revised in order to make it consistent again. Fourth, she ranks the various revision possibilities by the number of preference relations changed, weighed by their respective centrality. Fifth, if this procedure leaves preferences undetermined that are relevant for the evaluation of a policy, the policy maker derives these preferences from a more central preference on the basis of all available information.[53] Note that such a derivation may not possible, due to the absence of relevant core preferences. In that case, paternalist intervention should be avoided altogether.

This procedure satisfies the internalist intuition soft paternalism was trying to respect. It clearly does not respect every preference of a person, if her preference ordering is inconsistent.[54] Yet it respects the preferences that are most central to a person, and constituent of her individual character. These preferences likely constitute judgments about states that a person most cares about. Leaving these intact in the preference reconstruction, and using them to return the ordering to consistency, most likely avoids the case where the person disavows the reconstruction result as normatively not authoritative for herself.[55]

The result of such a reconstruction is normatively relevant for three reasons. First, because it not only respects a person's core desires, but sees to it that all her preferences are in accord with these, the reconstruction increases the person's integrity, which might be a value in itself. Second, it actively promotes what the person most cares about, to the detriment of objectives that are more peripheral to the person. Third, it is likely that preferences derived or linked to core desires are more stable than those that lack such connections. Thus, seeking to satisfy preferences connected to a core is more likely to satisfy existent preferences than seeking to satisfy preferences not so connected.[56]

[53] The information requirement employed here applies only to instrumental derivations in the specific case where reconstruction yields a preference indeterminate between two relevant options, and hence is much weaker than the requirement discussed in section 6.

[54] It may also not respect all her preferences if her ordering is consistent but so impoverished that additional welfare-relevant preferences have to be derived in the way described above.

[55] If such disavowal nevertheless occurs, it is qualitatively different from that discussed in the pervious section. The citizen who never had a chance to refine his musical tastes may reasonably object to another concert hall, even if his fully informed self demands its construction. Yet the music lover who has no conflicting preferences but professes to lack a token preference for this concert hall will likely appear as insincere.

[56] This is a concern about standard soft paternalist strategies. People may be nudged into adopting (and satisfying) preferences that are not connected to their core values and that they may loose after a while. For example, an employee may stay on a diet as long as he goes to work and eats in the softly paternalist cafeteria, but then reverts to gluttony on weekend and during vacations. While such cases raise serious doubts about soft paternalism (cf. Luc Bovens 'The Ethics of Nudge' in Modelling Preference Change: Perspectives from Economics, Psychology and Philosophy edited by Till Grüne-Yanoff and Sven-Ove Hansson (Heidelberg and New York: Springer, 2009): 207-220), I think that a focus on reconstructed preferences can avoid this problem.

Thus, if the diner has core commitments for a healthy lifestyle but fails to choose a healthy meal in the cafeteria, then a nudge is justified. Similarly with the employee in the pension default case and the customer in the home solicitation case. But if these people do not have the relevant core commitments, then a soft paternalist intervention to nudge them into such choices would neither satisfy the internalist intuition, nor would they be beneficial to these people in the sense of a preferentialist welfare notion.

8. Conclusion

Soft paternalism distinguishes itself by respecting the internalist intuition, but it is based on cases where people's preferences are often not consistent. The ensuing Soft Paternalist Paradox can only be resolved by rethinking the welfare notion that justifies paternalistic intervention. I argued that non-preferentialist alternatives fail because they disregard the internalist intuition. Yet a mere correction of the welfare measurement, as recently proposed in economics, is ineffective or begs the question. Instead, a reconstruction of people's preference orderings is required before welfare conclusions can be drawn from them. Such an approach can take two different avenues. The full information approach, I argue, disrespects the internalist intuition. Instead, I propose the integrity approach, which respects people's core preferences and reconstructs a consistent preference ordering around them. I show that the integrity approach respects the internalist intuition and produces welfare-relevant preference orderings. It therefore can resolve the Soft Paternalist Paradox.

I leave it to others to decide whether the proposed cure is worse than the disease. I showed that if Soft Paternalism is committed to internalism, then something like the integrity account is required to make sense of the welfare claims the Soft Paternalist is making. Should one think such an account infeasible for policy making, one may also consider this an argument against Soft Paternalism.