PAPERS on Economics & Evolution

1022

Role Models that Make You Unhappy: Light Paternalism, Social Learning and Welfare

by

Christian Cordes
Christian Schubert

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

Role Models that Make You Unhappy: Light Paternalism, Social Learning and Welfare

Christian Cordes and Christian Schubert*

January 2011

Abstract

Behavioral (e.g. consumption) patterns of boundedly rational agents can lead these agents into learning dynamics that appear to be "wasteful" in terms of well-being or welfare. Within settings displaying preference endogeneity, it is however still unclear how to conceptualize well-being. This paper contributes to the discussion by suggesting a formal model of preference learning that can inform the construction of alternative notions of dynamic well-being. Based on the assumption that interacting agents are subject to two biases that make them systematically prefer some cultural variants over others, a procedural notion of well-being can be developed, based on the idea that policy should identify and confine conditions that generate dynamic instability in preference trajectories.

Keywords: Social Learning, Preference Change, Welfare, Human Cognition, Consumer Behavior

JEL Classification: D63, O12, D11, D83, C61

* Cordes (corresponding author): Max Planck Institute of Economics, Kahlaische Str. 10, 07745 Jena, Germany (e-mail: cordes@econ.mpg.de); Schubert: Max Planck Institute of Economics, Kahlaische Str. 10, 07745 Jena, Germany (e-mail: schubert@econ.mpg.de). The authors are grateful for helpful suggestions from participants at the 2009 joint IAREP/SABE conference, Halifax, Canada. The usual caveat applies.

1. Introduction

Behavioral economics has reached a state where its methodology and key insights are acknowledged even by the mainstream of economics. Despite this success, it is still largely unclear whether and in which sense these insights translate into viable policy implications. This is particularly puzzling in light of the fact that there is a variety of policy issues that only become "visible" through the lens of psychologically informed theorizing: consider the problems related to overborrowing (Thaler and Sunstein, 2008, ch. 8), obesity (Anand and Gray, 2009), gambling (Benjamin and Laibson, 2003), the perverse effects of the welfare state (Beaulier and Caplan, 2007), and status races (Frank, 2008), to name just a few. In all these case, individual preferences seem to depart, in a systematic way, from what is in the individual's own "best interests", thereby generating predictable ex post regret. In order to improve policy advice in these fields, behavioral economics requires not only instrumental knowledge about means-ends-relationships, but also standards of well-being or welfare.[1] Unfortunately, the standard concepts of Paretian welfare economics cannot be used, since there well-being is ultimately defined as the satisfaction of "given" and perfectly consistent ("rational") preferences.[2] In contrast, behavioral economics typically takes preferences to be endogenous and prone to inconsistency (causing "mistakes"). While the search for alternative notions of welfare is well under way, it has so far failed to come up with convincing results (Loewenstein and Haisley, 2008), thereby limiting the practical relevance of behavioral economics.

The present paper suggests a way out of this deadlock, focusing on the specific case of status-oriented consumption and interdependent, dynamic status races. Inspired by Veblen (1898), there is a growing literature on the motivation of status consumption (Johansson-Stenman and Martinsson, 2006; Rege, 2008; De Fraja, 2009), the behavioral dynamics it generates (e.g. Friedman and Ostrov, 2008), and possible implications for, e.g., utility levels (Hopkins and Kornienko, 2004), social capital (Eaton and Eswaran, 2009), tax policy (Ng, 1987a; Ireland, 2001; Frank, 2008) and the provision of public goods (Ng, 1987b).[3] We contribute to this literature by arguing that in order to be consistent, policy implications with respect to status races have to be based on a notion of welfare that can be applied in a "behavioral", dynamic world

[1] These terms will be used interchangeably in the following.
[2] This problem has apparently been neglected by the early literature on behavioral economics (see Berg, 2003).
[3] Seminal contributions are Duesenberry (1949, ch. 3), Hirsch (1976) and Frank (1999).

where preferences change over time and tend to be inconsistent. Standard Paretian concepts of welfare cannot do the job. However, those non-standard notions of welfare that have so far been proposed in the literature – such as "idealized choice", happiness, capability and "opportunity" – suffer from shortcomings as well. Two issues are particularly important: first, most of these notions ("opportunity" being the exception) are outcome-oriented rather than focused on the process of preference development. Second, all of them are insufficiently grounded in empirical evidence on human choice in a "behavioral" world of inconsistent and variable preferences. Our argument starts from the intuition that in such a world, policy should follow procedural rather than static concepts and focus on the ability of agents to engage in the ongoing instrumentally effective learning of new wants and preferences. Put differently, the market's institutional framework should give agents maximum freedom to try out new preferences. At the same time, however, it should make sure that systematic (hence foreseeable) self-defeating preference dynamics are avoided. If these two goals conflict, the first goal should take precedence, which calls for the use of freedom-preserving instruments such as "nudges" (Thaler and Sunstein, 2008).

In order to identify self-defeating preference dynamics, we introduce a formal model of cultural transmission where preferences are shaped by biased individual and social learning processes. There is mounting empirical evidence that this applies to status-oriented consumer behavior. Based on interdependent utility functions, spending on status-signaling goods affects the reference frames of peers, leading to status races that typically involve self-augmenting cycles of resource spending with only a transient positive effect on the agents' own levels of well-being. In this sense, such races can be considered socially wasteful. Recently, policies have been proposed, and policy tools devised, that may help steer individual behavior in such a way as to align it with the individuals' own self-interest, as perceived by themselves. This "steering" typically involves either the design of incentives (through the taxing of luxury goods, say) or attempts by government to intervene in the underlying processes of social norm or preference formation. The latter strategy has been referred to as "libertarian" or "asymmetric" paternalism (Thaler and Sunstein, 2008; Camerer et al., 2003; Benjamin and Laibson, 2003).[4]

[4] See also Lessig (1995) for examples involving the regulation of informal social norm formation. We leave aside the intricate instrumental issue of how policy can effectively shape informal social norms (see Parisi and von Wangenheim, 2006).

It is in particular proposals of the latter kind that tend to provoke resistance. This is directly related to their unclear normative foundations: intuitively, most economists reject the paternalistic implications of attempts by government to directly influence processes of norm or preference formation (Sugden, 2008; Glaeser, 2006). Interventions of this kind do indeed raise thorny normative issues. On which grounds can they (ever) be justified, even when involving only minimum degrees of coercion? Due to both the complexity of the normative concepts and the magnitude of the social welfare issues involved, the debate is far from being concluded. On a deeper level, it cannot be settled without clarifying what exactly is meant by notions such as "individual well-being" and "social welfare" in settings where preferences are endogenous and potentially inconsistent. This task is made even more demanding in light of the widely shared intuition that procedural considerations (involving, e.g., the agents' "autonomy") should play a role in the assessment of policies and states of affairs.

The procedural notion of welfare that we suggest is inspired and informed by our formal model of cultural learning. We assume that interacting boundedly rational agents are subject to a role model (or indirect) bias and a content-related (direct) bias that make them systematically prefer some cultural variants - for example, certain consumption behaviors - over others. These preferred behaviors or cultural variants include those displayed by prestigious individuals or those suited to signal status via consumption. There is a plethora of evidence for the existence of these kinds of biases provided by social psychology, the cognitive sciences, and other disciplines (see, e.g., Aronson et al., 2002; Norenzayan and Heine, 2005; Richerson and Boyd, 2005). Depending on the relative strengths of these two biases, their interplay can generate self-augmenting learning dynamics that lead agents away from sustainable and welfare enhancing consumption paths and that may ultimately even undermine the learning process itself. This is the basis for distinguishing between desirable and undesirable paths of preference change. We suggest that desirable preference paths represent what may be labeled "effective preference learning" and that this should be the standard of welfare guiding public policy. Hence, policy should foster the agents' chances to engage in effective preference learning by leaving their freedom to choose untouched, while at the same time employing "nudges" in order to help agents avoid undesirable preference cycles and the concomitant risk of learning failure.

The paper is organized as follows. Section 2 suggests a set of criteria for assessing whether welfare measures can be applied in our context and briefly surveys the existing literature on

alternative notions of individual well-being and social welfare, i.e., notions that are in principle applicable in a setting where individuals continue to learn new and possibly inconsistent preferences over time (instead of sticking to "given" and perfectly "rational" preferences). Section 3 presents our formal model of cultural transmission that shows how boundedly rational agents acquire consumption norms in a systematically biased way. Section 4 suggests a new welfare notion of "effective preference learning" that draws on a key insight of the model. Section 5 concludes.

2. Welfare with inconsistent and variable preferences

Until quite recently, welfare economics as the study of how to conceptualize, justify, model and operationalize normative standards and policy goals had almost vanished from the agenda of mainstream economics (Sen, 1987; Atkinson, 2009). As the insights of behavioral and evolutionary economics into the complexity of human behavior – giving rise to "non-standard models of choice" – are now increasingly acknowledged in the profession, this has changed dramatically. For many of these new insights raise important policy problems: is it legitimate to contain status races by taxing certain goods? Is it acceptable to "nudge" people toward certain (e.g., environment-friendly) behaviors? Should policy promote public information campaigns, even if they cannot be strictly neutral with respect to people's preferences? All these are contentious issues. They cannot be answered, however, without a thorough inquiry into the underlying concept of welfare. This is a non-trivial task: when individual preferences tend to be inconsistent and subject to endogenous change, traditional preference-based notions of welfare can no longer be applied (Sugden, 2004; Bernheim, 2009). An agent subject to biased social learning processes does not necessarily choose options that enhance her own well-being. Economists have to think anew about how to define and conceptualize welfare.

This is maybe most apparent in the ongoing debate about the normative foundations of approaches advocating "libertarian", "light", "soft", "benign" or "asymmetric" (henceforth LLSBA) versions of paternalism (see Loewenstein and Haisley, 2008, for a survey). According to Thaler and Sunstein (2003), a policy counts as paternalistic "if it is selected with the goal of influencing the choices of affected parties in a way that will make those parties better off" (ibid.,

p. 175).[5] But what does "better off" mean, i.e., to what notion of well-being or welfare does it refer to? The same problem arises in the context of applying the concept of "asymmetric paternalism" proposed by Camerer et al. (2003): they require that the net benefit of any paternalistic intervention (such as, e.g., a default rule stimulating increased saving) accruing to irrational consumers should outweigh the aggregate costs to rational consumers (ibid., pp. 1212, 1219-20). As Loewenstein and Haisley (2008) rightly note, "[t]o evaluate costs and benefits … once again requires some concept of welfare."

2.1 A catalog of criteria

An alternative notion or metric of welfare is thus clearly needed. In order to organize and facilitate the search process and restrict the range of possible candidates, four separately necessary and jointly sufficient conditions can be suggested that any such notion should satisfy:

(1) First, the notion of welfare should not be detached from what real-world individuals (i.e. those depicted in the behavioral economics literature) are able to care about and effectively do care about, according to our best available empirical or experimental evidence. In philosophical terms, it should respect the "internalist intuition" (Grüne-Yanoff, 2009). As a consequence, it should be broadly and at least indirectly based on the individuals' own subjective perceptions about what makes them better off.

(2) Second, the notion of welfare has to come to terms with the empirical fact that in many contexts relevant for economic policy-making, individual preferences tend to be inconsistent and do not perfectly reflect the agent's well-being or "utility". Systematic errors are common. Hence, there is a gap between, (i), what individuals "actually" prefer and what they choose and, (ii), between what they choose and what they either "like" (what makes them subjectively happy) or what is "good" for them from some objective standpoint. This dissociation between preference, choice and well-being is excluded by definition in standard accounts of welfare that are based on the weak axiom of revealed preference (Samuelson, 1948), but it is made explicit as a

[5] Dworkin (2010) defines it as "the interference of a state or an individual with another person, against their will, and defended or motivated by a claim that the person interfered with will be better off or protected from harm."

background assumption in the mainstream Behavioral Economics literature and in the literature on LLSBA paternalism.[6]

(3) Third, the notion of welfare should not be based on the assumption that individual preferences are fixed and exogenously "given". Rather, when confronted with a new choice situation they either "construct" or "learn" new preferences.[7] On a theoretical level, this aspect is less explicit in the contributions to LLSBA paternalism, but it clearly underlies the practical settings in which the recommended policies are to be applied (consider entering the notorious cafeteria in Thaler and Sunstein, 2008). In order to integrate this aspect into a theory of well-being, either the theory has to define well-being independent of the agent's preferences, or it has to account for (and be able to evaluate) the process through which an agent acquires her preferences. To illustrate the latter approach, the process of preference formation may be subject to manipulation or indoctrination, which may then be taken to indicate sub-optimality (and possibly the need for some kind of intervention).

(4) Fourth and finally, the notion of welfare should not by itself exclude considerations related to the (instrumental or inherent) value of individual freedom or "autonomy". In the literature on LLSBA paternalism, this condition plays an important role at the level of instrumental reasoning, where policy tools are being devised that are supposed to respect individual autonomy by leaving choice-sets (almost or perfectly) unchanged or by even extending them. Incorporating the dimension of autonomy into the welfare calculus has of course been a long-standing desideratum in welfare economics (Sen, 1987; Sugden, 2010), but so far it is still largely unclear how "autonomy" could be conceptualized.

2.2 Four candidate notions of well-being

Given this catalog of four requirements, it is possible to assess the four main candidates that have so far been suggested in the literature, viz., idealized choice, experienced utility, capability, and opportunity. Two preliminary remarks are in order, though. First, in our view, the complexity of the task to construe a new and "richer" notion and criterion of welfare calls for the broadening of its "informational basis", i.e., for basing the new concept on an empirically informed (rather

[6] See, e.g., the distinction between "decision utility" and "experienced utility" in Kahneman et al. (1997).
[7] Evidence that in many economically relevant situations, agents have difficulty in evaluating the utility derived from experiences and in constructing new preferences is provided by Ariely et al. (2006).

than purely axiomatic) account of human behavior and learning. Second, for the sake of the argument, we will abstract from any deeper discussion of operational difficulties, such as measurement problems, for these difficulties concern all accounts of welfare that have so far been suggested (including our own). Notice, however, that as scientific progress in the behavioral and neurosciences is quite impressive, methodological problems that appear insurmountable at the moment may be solvable in the near future. That is why we will stick to conceptual and theoretical problems in the following. Representing the four main notions of welfare that have been suggested so far, we will now briefly discuss the accounts proposed by by Bernheim and Rangel (2009), Kahneman et al. (1997), Sen (1985; 2006) and Sugden (2004; 2006; 2010).

Among these accounts, Bernheim and Rangel's (2009)[8] sticks most closely to orthodox ideas of welfare. While basically retaining the traditional choice-based concept of welfare (hence, *prima facie*, respecting condition 1), they modify it by arguing that inconsistent or "distorted" choices should be excluded from the welfare calculus. Hence, only a subset of an agent's choices (read: preference satisfactions) shall be taken to indicate an effective increase in that agent's well-being. This implies that a LLSBA intervention would be deemed legitimate to the extent that it corrects for "objective" inconsistencies in agents' revealed choices.

In trying to find a basis on which to distinguish between "distorted" and "non-distorted" choices, the authors concede that non-choice data (such as evidence from brain scans) will be necessary. Assuming that all the necessary data are available, the account suggested by Bernheim and Rangel presupposes that those choices that are not affected by, for example, memory failures, inability to learn or utility misprediction, are indeed reliable indicators of welfare. In that sense, the notion of welfare accommodates the existence of preference inconsistency, hence satisfying condition (2). What about condition (3)? Imagine a person who really wishes to pursue a given activity. After having thoroughly reflected on it, using perfectly rational information processing capacities and all available information on the consequences, she gets what she "truly" wants. Does this necessary imply that her welfare level has increased? According to Sen (1987), preferences may be shaped by the agent's circumstances. It may be the case that "the hopelessly deprived lack the courage to desire much" (ibid., p. 46). To reduce cognitive dissonance, agents may downgrade their aspirations and preferences when confronted with meager opportunity sets. Elster (1982) calls this phenomenon "adaptive preference change". Importantly, adaptation works

[8] see also Bernheim (2009).

independent of the quality and amount of information that has entered the decision, which implies that our condition (3) is violated: the informed choice account of welfare cannot adequately cope with the phenomenon of preference variability – at least not with the sub-phenomenon of adaptive preference change.

There is a second issue involved here. The idea put forward by Bernheim and Rangel (2009) closely relates to a family of welfare accounts suggested three decades ago by, for example, Brandt (1979, ch. 6), Griffin (1986, chs. 1-2), and Harsanyi (1982).[9] According to Harsanyi (1982, p. 55), for instance, an agent's revealed preferences should enter the social welfare calculus only to the extent that they reflect his "true" preferences, these being "the preferences he would have if he had all the relevant information, always reasoned with the greatest possible care, and were in a state of mind most conducive to rational choice." In an axiomatic sense, "perfect information" boils down to Bernheim and Rangel's "pruned" choices (see above). That is why these ideal preferences may have nothing to do with the actual ("manifest") preferences that a cognitively constrained, real-world agent happens to have learned over time, for example, via imperfect channels of cultural transmission (see our model below, Section 3). It is unclear whether such a real person would endorse her allegedly "true", but counterfactual, preferences if they were presented to her with the expert hint that she would have them had she had all relevant information and had she processed that information in a perfectly rational way. This however means that, contrary to first appearances, our "internalism" condition (1) is not satisfied. Closely related to this, it is questionable whether an account that, in a world populated by at least some irrational agents, necessarily leads to the immediate exclusion of some of the agents' manifest preferences satisfies the "autonomy" condition (4).

Among these shortcomings, from a methodological perspective at least the violations of conditions (3) and (1) are related in the following way: Harsanyi's "true" preferences are those of a perfectly informed and perfectly rational homo economicus. For such an agent, neither Sen's divergence between objective well-being and preference satisfaction (see above) nor a divergence between manifest and "true" preferences can possibly occur. This is due to the fact that Harsanyi's account is not based on an empirically informed positive theory of human behavior or learning, but rather on the axiomatic account of rational choice embodied in the homo

[9] See also Sobel (1994).

economicus concept. We stipulate that this is at the heart of this (and Bernheim and Rangel's) account's shortcomings as a theory of human well-being.

The hedonic approach, most prominently endorsed by Kahneman et al. (1997),[10] replaces the axiomatic basis of orthodox welfare economics by an empirically informed foundation. Here, welfare is redefined as "experienced utility" which in turn is inferred from self-reported levels of hedonic well-being. In this view, LLSBA interventions would be judged legitimate to the extent that they correct preferences that systematically lead to suboptimal hedonic outcomes. More generally, government is supposed to maximize an empirical hedonic social welfare function or at least use happiness data to assess the desirability of alternative social states or trade-offs involved in, for example, weighing inflation against unemployment (Di Tella et al., 2001), or more mundane matters such as airport noise against money (van Praag and Barsma, 2005). As these accounts rely on data on self-reported well-being, they easily satisfy our "internalist" condition (1). In the context of the present paper, though, the main advantage of a hedonic account of welfare lies in its independence of the choices real-world individuals make. Hence, experienced utility can be used to critically evaluate the "rationality" – and, hence, normative weight (Ng, 1999) – of alternative preferences, their paths, and choices. For this reason, it can be argued that this account satisfies condition (2). If suitably amended, it may also be able to satisfy condition (4), for the value of autonomy may well be defined as instrumentally valuable within a general hedonic framework.

The main problem associated with this approach is that reported experienced utility, too, is subject to adaptation over time: people tend to adapt their subjective level of hedonic utility to both fortunate and unfortunate circumstances (Frederick and Loewenstein, 1999)[11]. Hence, it violates our condition (3): in the presence of adaptation effects, basing public policy solely on a happiness metric bears the risk of generating counter-intuitive implications. The issue is closely related to the adaptation problem in the context of preference formation, discussed above. As Sen (1987, pp. 45-46) puts it, "[t]he hopeless beggar, the precarious landless labourer, the dominated housewife … may all take pleasures in small mercies … but it would be ethically deeply mistaken to attach a correspondingly small value to the loss of their well-being because of this survival strategy". In a different context, such a metric could imply that a health-impaired

[10] See also Ng (2003), Layard (2005), Frey and Stutzer (2002).
[11] The locus classicus is Brickman and Campbell (1971).

person's level of "welfare" would positively depend upon her personal ability to adapt – leading to the practical consequence that, for example, a welfare-sensitive legislator or judge would make damages owed to this person depend negatively on her ability to adapt.

Hence, experienced utility does not appear to be an attractive candidate for an alternative concept of welfare that could be applied in our context. Let us finally discuss the two objectivist approaches to human well-being. They are objectivist in the sense that they define well-being (at least partly, as in Sen's case) independent of the agents' subjective attitudes or mental states. Consider Sen's *capability* approach: here, an agent's well-being depends on what the agent is able to achieve. Put differently, it is constituted by the vector of functionings that are effectively available to her (Sen, 1980; 1985).[12] A functioning is defined, very broadly, as "an achievement of a person: what she or he manages to do or be" (Sen, 1980, p. 10). Examples include "being adequately nourished", "having a basic education" or "being able to appear in public without shame". Functionings refer to the use a person makes of the commodities she commands and her ability to transform commodities into personal quality of life (Clark, 2006). An agent's capabilities are defined as the set of alternative combinations of functionings that she is able to achieve (e.g., the ability to achieve the state of "being adequately nourished"). Sen also refers to capabilities as the agent's "substantive freedoms he or she enjoys to lead the kind of life he or she has reason to value" (Sen, 1999, p. 87). In contrast to utility-based notions of welfare, in the capability approach an agent's quality of life does not only depend on the achievements (outcomes) which are realized in the end, but also on the extent to which she has the freedom to choose among alternative options. Given this account of well-being, social arrangements are then supposed to expand people's capabilities (rather than promote economic growth). Hence, LLSBA policies would be legitimate to the extent that they promote the provision of functionings by, e.g., restricting behavior that is individually self-defeating in terms of these functionings.

Turning now to our catalog of four criteria, it is not entirely clear whether the capability approach satisfies criterion (1). On the one hand, Sen and other contributors to this strand of literature (e.g., Nussbaum, 2000) take great pains to argue that their lists of functionings reflect what real-world agents really do care about. Consider the procedural dimension of choice which is neglected in standard utility-based accounts of well-being, but which obviously is valued by most people. On the other hand, given that this approach is objectivist, the agents' subjectivist

[12] See Clark (2006) for a recent survey of the field.

perceptions of their own well-being are discounted (Sugden, 2006). Agents may have a say in establishing, by means of public deliberation, the lists of functionings (there is not *one* "definite" list, at least not in Sen's version of the capability theory), but beyond that "constitutional" stage, their well-being is assessed according to an external set of criteria.

It is also easy to verify that this approach satisfies our criteria (2) and (4): as to (2), preference inconsistency does not pose a challenge, since the capability approach does not conceptualize well-being in terms of consistent preference-satisfaction. As to (4), autonomy considerations are explicitly integrated into this approach (see above). It is less obvious, though, whether the capability approach can accommodate preference endogeneity (condition 3). Again, the phenomenon of adaptation enters the scene. According to Sen, the relevant list of valuable functionings should not exclusively be set up by some scientific armchair theorist; rather, it should also depend on the subjective valuations of the agents whose quality of life is to be assessed. This gives the theory a partly subjectivist outlook. At the same time, it creates the risk that the subjective valuations themselves will be shaped by environmental influences, leading to adaptation effects analogous to the ones discussed above for the case of preferences and happiness: "personal values are also notoriously subject to influence by accustomed social conditions" (Sumner, 1996, p. 66). Hence, in this case it is again the adaptation problem that precludes the satisfaction of condition (3).[13]

Finally, consider the *opportunity* criterion of well-being advanced by Robert Sugden in a series of recent papers (2004; 2006; 2008). Taking the tendency of empirical preferences to be inconsistent as his starting point, Sugden suggests to view and embrace it as a fact of life rather than a source of suboptimality (*qua* inconsistency). He proposes to replace the standard preference-satisfaction account of welfare by an alternative notion, defining well-being in terms of the opportunity to act on whatever preferences one may turn out to hold in future periods. It is not the actual satisfaction of these preferences, but rather the unrestricted opportunity to acquire them and to act on them (whether they happen to be consistent or not) that generates well-being. All this holds in the domain of preferences for private goods and services. As an underlying normative principle of consumer sovereignty, Sugden suggests that the individual has to assume responsibility – and bear the consequences – for whatever preferences she may then hold.

[13] Notice that this argument presupposes that subjective attitudes are in general susceptible to adaptation effects of the kind described. That implies in particular that "feeling" and "critical reasoning" (the process where personal valuation is brought to bear) do not differ in this regard, a presumption that Sen (2006, p. 93) objects to.

It is as yet unclear to what extent this approach supports any LLSBA intervention. Sugden himself argues that in this framework any intervention would directly affect the agent's opportunity set – and, hence, well-being – in a negative way, hence precluding any such policy (Sugden 2008).[14] On the other hand, LLSBA paternalists argue that most policy measures that they propose do leave the agents' opportunity sets unaffected. It seems that this argument cannot be resolved without further clarification on what it means to enjoy and act upon one's own "opportunity set". For our purposes, suffice it to say that it is at least conceivable that even with Sugden's notion of welfare in place, LLSBA paternalism would not be totally excluded. On the conceptual level, Sugden's approach obviously satisfies our criterion (2): preference inconsistency is not an issue here. It also satisfies criterion (3), since it is based on the prediction that individual will continue to learn and acquire new preferences, which includes adaptation to changing environmental conditions. Given the fact that all competing accounts of welfare failed at this point (see above), this is a remarkable achievement. Sugden's account also reserves a special place for autonomy or freedom considerations (witness his responsibility norm), hence satisfying criterion (4).

There is only one problem: this approach violates the internalism criterion (1), and this violation is largely due to its neglect of the psychological mechanisms driving preference change. For in order to back his normative requirement of responsibility and to provide his approach with formal consistency, Sugden has to assume at the outset that agents are willing to endorse any preference whatsoever they will acquire in future periods. This does not rule out the possibility that preferences will be inconsistent, but it does rule out that in light of this possibility agents will wish to engage in prudential *self-commitment*. This quasi-axiomatic statement is, however, at odds with empirical evidence: behavioral economics tells us that individuals who are aware of their own cognitive constraints and limitations of willpower, often wish to engage in self-commitment. This already applies in cases where there are only a limited number of "given" preference orders (say, a "cold" and a "hot" one) the agent holds in turn.[15] It is even more pertinent in the case of expected preference change, when in light of current preferences, the agent may wish to restrict the option to satisfy a subset of possible future preferences. Hence, by

[14] See also Sugden (2009).
[15] Note, though, that the "hot-cold empathy gap" may make such self-commitment difficult in practice (see Ariely and Loewenstein, 2006).

overriding this deep-seated wish, Sugden's approach partly overrides what individuals themselves care about, which violates our criterion (1).

We conclude this brief survey of existing non-standard accounts of well-being by observing that all accounts suffer from the lack of empirically informed "positive" foundations, notably with respect to the processes or mechanisms involved in preference change. Assuming that adaptation effects and the desire to self-commit are essential components of real-world processes of preference formation, it is easy to see that the existing non-standard accounts of welfare either require the exclusion of the former (Bernheim and Rangel, Kahneman, Sen) or the dismissal of the latter (Sugden).

It seems that in order to get an account of well-being that satisfies our criteria (1)-(4), we need to be able to integrate both phenomena. In order to do this, we suggest to follow an intuition expressed many years ago by Elster (1982): in a world of preference endogeneity, it does not make sense to use outcome or "end-state" notions of welfare. Rather, well-being should be made a function of the historical genesis of an individual's wants (ibid.: 237-38). This implies, first, to eschew all criteria discussed above, except Sugden's. Given the need to take seriously the concerns of real-world individuals facing preference endogeneity, it implies, second, to inquire into the distinction between "desirable" and "undesirable" ways preferences can develop over time. In order to do this, it seems necessary to dwell on an empirically informed model of preference change and learning. In this context, preference learning paths are "undesirable" to the extent that – due to biased social learning processes – they lead individuals away from those states that they themselves would consider to be welfare-enhancing. Such a model will be presented in the next Section. On its basis it may then be possible to develop an account of well-being that distinguishes between "beneficial" and "harmful" (or self-defeating) ways to change given preferences and acquire new preferences.

3. A model of biased cultural learning of consumption norms

Formal models of cultural evolution analyze how cognitive processes of human agents combine with patterns of social interaction on the population level to generate the distributions and dynamics of cultural variants, for example, the preferences underlying different consumption

behaviors (Cavalli-Sforza and Feldman, 1981; Boyd and Richerson, 1985; Henrich and Boyd, 2002; Richerson and Boyd, 2005). We draw on this approach to account for preference learning dynamics in a population of heterogeneous consumers. As shown, we are then able to distinguish different preference learning trajectories and their corresponding welfare implications.

Cultural transmission or learning is biased; people tend to acquire some cultural variants rather than others. We assume two biases here: a role model bias and a content-related, direct bias. First, the choice of a cultural trait can be based on the observable attributes of the individuals who exhibit the trait (Richerson and Boyd, 2005, 69; Harrington Jr., 1999). Such a model-based bias includes a predisposition to imitate successful or prestigious individuals. In general, such an indirect bias results if social learners use the value of a secondary character that characterizes a model (e.g., her observable consumption level) to determine the attractiveness of that individual as a model for the primary character (e.g., the preference concerning the level of consumption one should aspire). Hence, one trait of a role model is taken as an indicator whether it is worthwhile to copy another aspect of this individual's behavior or attitude. Second, individuals are more likely to adopt some cultural variants rather than others based on their content (Boyd and Richerson, 1985, 135; Richerson and Boyd, 2005, 69). Such a content-based or direct bias can result from cognitive structures that cause people to preferentially adopt some cultural behavior rather than others (e.g., Cordes, 2005). In general, a cultural transmission rule is characterized by direct bias if one behavioral variant, for example, status-signaling consumption behavior, is more attractive than others. While the general strive to signal status is partly innate and partly learned, the kind of commodity suitable to do so depends to a great degree on the cultural environment. A directly biased transmission creates a force that increases the frequency of the culturally transmitted variant that is favored by the bias.

To formally depict some normatively relevant facets of cultural evolution, we draw on a model based on a set of recursion equations that first was applied to quantitative genetics (see Fisher, 1930; Lande, 1981; Kirkpatrick, 1982) and adapt it to our purposes and the specificities of cultural dynamics (for accessible introductions to models of social evolution of this type see Cavalli-Sforza and Feldman, 1981; Boyd and Richerson, 1985, ch. 8; McElreath and Boyd, 2007, ch. 8). We begin with the effects of biased cultural learning on an individual's preference and consumption levels. In a next step, we will then show how these learning dynamics translate into consequences on the population level. We consider an individual's preference level to be

cardinally measurable. In principle, a shift to an ordinal scale should not alter the model's general mathematical properties and thus the argument developed here.

Let x_0 be the measure of a consumer's initial consumption level and p_0 the measure of this individual's preferred level of consumption. Moreover, we assume that each individual is exposed to n role models' consumption behaviors $(x_1,...,x_i,...,x_n)$. To allow the relevance of different models (e.g., individuals in different social positions) to differ, we assign different basic weights, denoted by α_i, to them, whereby $\sum_i \alpha_i = 1$. Moreover, cultural transmission of consumption behaviors is subject to a direct bias, i.e., individuals choose role models by comparing their notions of an appropriate level of consumption with the observed consumption behaviors of models in the population. We assume that, in the beginning, an individual's preference concerning consumption, p_0, exceeds her initial consumption level given by x_0, i.e., the direct bias favors higher consumption levels. This reflects a preference for, for example, status-signaling goods and a general human concern about status or relative position in a group. This implies that another part of the influence of the ith role model, i.e., beside her basic weight α_i, depends on her revealed consumption level, x_i, and the preference level of the observing individual, p_0, captured by the function $\beta(x_i, p_0)$. In this context, the preference level affects which consumption levels or behaviors an individual finds attractive. The observed consumption level, x_i, that maximizes $\beta(\bullet)$ is the most attractive one. Role models that exhibit this value will, on average, have the greatest influence in the transmission of consumption behaviors. The farther x_i is away from p_0, the lower the weight of this role model. Below, we will specify a Gaussian form for $\beta(\bullet)$ and clarify this idea.

In total, the weight of the ith model in the transmission of consumption behaviors is given by her basic weight, α_i, and the direct bias function, $\beta(\bullet)$. The recursion equation that determines the level of consumption after cultural transmission, x_0', is

$$x_0' = \frac{\sum_i^n x_i \alpha_i \beta(x_i, p_0)}{\sum_i^n \alpha_i \beta(x_i, p_0)}. \quad (1)$$

The role models' weights are normalized by the denominator so that Equation (1) gives the influence of the ith model relative to the other models encountered by the individual in question.

To represent the evolution of an individual's preference level, p_0, which is subject to our second, indirect bias, we again assume that the importance of the ith model in cultural transmission depends on a basic weight, α_i, modified by an indirect bias function, $\theta(x_i, p_0)$. As in the case of the direct bias, the latter part of the weight of the ith model is a function of her consumption behavior, captured by x_i, and the observing individual's preference level, p_0. p_i denotes the n models' preference levels concerning appropriate levels of consumption (p_1,..., p_i,..., p_n). Then, the individual's preference level after cultural transmission is given by

$$p_0' = \frac{\sum_i^n p_i \alpha_i \theta(x_i, p_0)}{\sum_i^n \alpha_i \theta(x_i, p_0)}. \tag{2}$$

Here, social learners are indirectly biased: they use the value of a secondary character that characterizes a model – the observed consumption behavior – to determine the attractiveness of that individual as a model for the primary character – the preference concerning a certain consumption level or behavior. Therefore, in determining their own preference level, consumers acquire beliefs from their influential role models about who should be imitated, i.e., about which consumption level one should strive for. In addition, the most influential agents – those characterized by a level of consumption close to the preference level of the individual in question – will prefer higher consumption levels than the population as a whole. We assume an individual's level of consumption to be positively correlated with her preference level. Via cultural transmission, agents adopt from their set of role models a higher preference level than they had before entailing a higher aspired consumption level. Furthermore, two individuals exposed to the same set of cultural models will, on average, adopt different cultural variants of traits affected by the role model bias.

After having defined the recursions that describe how individuals choose their consumption level and how they attain their preferences for a certain consumption level via cultural

transmission, we now turn to the evolution of the mean values of x_i, $\bar{x}$, and p_i, $\bar{p}$, in a population consisting of individuals behaving in the described way. To do so, we look at the changes in these mean values given by the following expressions:

$$\Delta\bar{x} = \bar{x}' - \bar{x} = \frac{\sum_i^n x_i\alpha_i\beta(\bullet) - \bar{x}\sum_i^n \alpha_i\beta(\bullet)}{\sum_i^n \alpha_i\beta(\bullet)} \tag{3}$$

and

$$\Delta\bar{p} = \bar{p}' - \bar{p} = \frac{\sum_i^n p_i\alpha_i\theta(\bullet) - \bar{p}\sum_i^n \alpha_i\theta(\bullet)}{\sum_i^n \alpha_i\theta(\bullet)}. \tag{4}$$

As is shown in the Appendix A (see also Price, 1970), (3) and (4) can be reformulated to

$$\Delta\bar{x} = \frac{\mathrm{cov}(x, \beta(x, p))}{\bar{\beta}} \tag{5}$$

and

$$\Delta\bar{p} = \frac{\mathrm{cov}(p, \theta(x, p))}{\bar{\theta}}. \tag{6}$$

For values of x and p close to $\bar{x}$ and $\bar{p}$, the bias functions $\beta(\bullet)$ and $\theta(\bullet)$ can be approximated by using a Taylor series approximation (see Appendix B). We then obtain for the changes in the mean values of x and p

$$\Delta\bar{x} \approx \mathrm{var}(x)\frac{\partial \ln\beta(\bullet)}{\partial x}\bigg|_{\bar{x},\bar{p}} + \mathrm{cov}(x, p)\frac{\partial \ln\beta(\bullet)}{\partial p}\bigg|_{\bar{x},\bar{p}} \tag{7}$$

and

$$\Delta\bar{p} \approx \mathrm{cov}(x, p)\frac{\partial \ln\theta(\bullet)}{\partial x}\bigg|_{\bar{x},\bar{p}} + \mathrm{var}(p)\frac{\partial \ln\theta(\bullet)}{\partial p}\bigg|_{\bar{x},\bar{p}}. \tag{8}$$

In this context, $\text{cov}(x, p)$ indicates the covariance between a certain level of consumption, measured by x, and the preference level, p, that determines the consumption level perceived as appropriate by a consumer. This is the assumed positive correlation between an individual's consumption level x and her notions about an "adequate" level of consumption, p. The other term, beginning with $\text{var}(x)$ or $\text{var}(p)$, gives the change due to the direct effect of variation in that variable, for example, how a change in an individual's consumption level affects her influence as a role model. In order to facilitate an evaluation of this two-dimensional system, we need to define particular functional forms for the direct and indirect bias expressions. Plausible forms of $\beta(\bullet)$ and $\theta(\bullet)$ are Gaussian bias functions

(9) $$\beta(x, p) = \exp\left\{-a(x-p)^2\right\}$$

and

(10) $$\theta(x, p) = \exp\left\{-b(x-p)^2\right\}$$

where the parameter a measures the strength of the direct bias, and the parameter b measures the strength of the indirect bias. Both forces are subject to genetic dispositions and an individual's idiosyncratic learning history (see, e.g., Richerson and Boyd, 2005, p. 66). These two functions measure the influence of role models in cultural transmission as a function of their exhibited consumption levels. This influence is maximized for $x = p$ and decreases the farther x is away from p. Given these assumptions and letting G_x and G_p indicate the variance in x and p respectively, while B_{xp} indicates the constant covariance between x and p, the recursions for the two means are[16]

(11) $$\Delta\bar{x} = G_x\left\{-2a\bar{x} + 2a\bar{p}\right\} - B_{xp}\left\{-2a\bar{p} + 2a\bar{x}\right\}$$

and

(12) $$\Delta\bar{p} = B_{xp}\left\{-2b\bar{x} + 2b\bar{p}\right\} - G_p\left\{-2b\bar{p} + 2b\bar{x}\right\}.$$

[16] Like most modelers using this kind of formal approach, we assume that the variance evolves to its equilibrium value independent of changes in the mean and is constant thereafter (see, e.g., McElreath and Boyd, 2007, p. 302).

Next, we solve for $\hat{\bar{x}}$ and $\hat{\bar{p}}$ denoting the equilibrium values of the consumption and the preference levels in a population of interacting consumers. It is easy to see that if $\bar{x} = \bar{p}$, both recursion equations equal zero. Therefore, $\hat{\bar{x}} = \hat{\bar{p}}$ is an equilibrium of this two-dimensional dynamic system. Given a graph with $\bar{x}$ on the horizontal axis and $\bar{p}$ on the vertical axis, $\bar{x} = \bar{p}$ is a line through the origin with slope one and any point on this line is an equilibrium. To further analyze the model's dynamic properties, it is helpful to compute its trajectory in such a diagram:

$$(13) \qquad \frac{\Delta\bar{p}}{\Delta\bar{x}} = \frac{B_{xp}(-2b\bar{x} + 2b\bar{p}) - G_p(-2b\bar{p} + 2b\bar{x})}{G_x(-2a\bar{x} + 2a\bar{p}) - B_{xp}(-2a\bar{p} + 2a\bar{x})}$$

$$= \frac{b(B_{xp} + G_p)}{a(B_{xp} + G_x)}.$$

If $\bar{x}$ changes one unit as a result of biased cultural transmission, the preference level, $\bar{p}$, changes $\frac{b(B_{xp} + G_p)}{a(B_{xp} + G_x)}$ units as a consequence of the correlated effects.

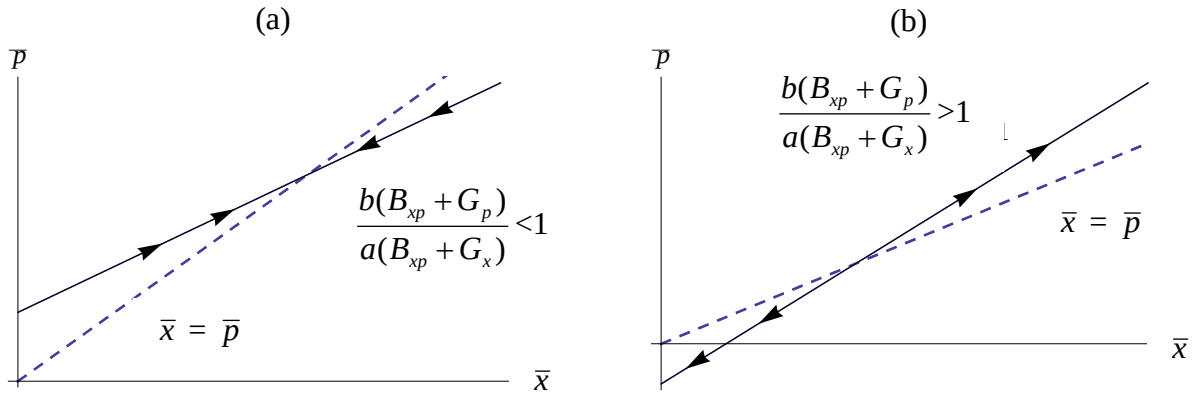

Figure 1 The phase plots of the model's evolutionary learning trajectories for (a) $a>b$ and (b) $a<b$ ($B_{xp}=3$, $G_x=1$, $G_p=1$).

Figure 1 shows two phase plots of the model's evolutionary trajectories that visualize its dynamics. In both cases, the dashed 45-degree lines through the origins with slope one are the lines of equilibrium, where $\bar{x} = \bar{p}$. Both, $\bar{x}$ and $\bar{p}$ increase above this line of equilibrium and

decrease below it. The other lines in Figure 1 (a) and (b) give the trajectories of populations of social learners that begin at different initial points and have slope $\frac{b(B_{xp}+G_p)}{a(B_{xp}+G_x)}$. Given the case depicted in (a), the slope of the trajectory is less than the slope of the line of equilibrium and the system is stable approaching an equilibrium on the 45-degree line, as is indicated by the arrows. On the other hand, in case (b), if the trajectory's slope is greater, the population evolves towards higher and higher preference levels and respective consumption levels when starting from a point above the line of equilibrium.

Therefore, if $a(B_{xp}+G_x) > b(B_{xp}+G_p)$, i.e., if the direct bias, whose strength is measured by the parameter a, is larger than the indirect bias effect, measured by b, then the learning dynamic eventually comes to rest at some point along the line of equilibrium. This situation is depicted in Figure 1 (a). On the other hand, if the indirect bias effect of role models is greater, i.e., if $b(B_{xp}+G_p) > a(B_{xp}+G_x)$, then the mean of the preference level in the population, $\bar{p}$, is increasing faster than the mean of the consumption level, measured by $\bar{x}$, and both traits "run away"; the system is unstable (case (b) in Figure 1 for initial values lying above the line of equilibrium). As a consequence, the distance of both mean values from the line of equilibrium is increasing from one learning step to the next ($\bar{p}' - \bar{x}' > \bar{p} - \bar{x}$). Certainly, this process cannot continue forever: some economic or ecological constraints will eventually restrain this dynamic. Some factors not accounted for in the model will eventually limit the evolution of the cultural traits in the population. To conclude, a situation with $\bar{p} > \bar{x}$ (inducing a rise in $\bar{x}$) combined with the role model bias function, $\beta(\bullet)$, that continuously updates the preference levels upward, makes social learners adopt ever more accentuated preferences relative to an "appropriate" level of consumption. This results in a self-augmenting "treadmill" of consumption choices.

Figure 2 offers another way to visualize the system's dynamic properties. It displays the development of a representative individual's preference level, p_0, and consumption level, x_0, by iterating the two-dimensional system consisting of Equations (11) and (12) for many social learning steps. As can be seen in Figure 2 (a), the role model bias in social learning gives rise to a widening gap between an individual's preference and consumption levels. This might also be interpreted as a growing discontent of an agent with her material condition. Figure 2 (b), in contrast, assumes that no role model bias continuously updates a consumer's preference level. In

this case, an individual's level of consumption finally reaches her preference level. The closing of this gap might entail a lasting satisfaction of this consumer's preferences.

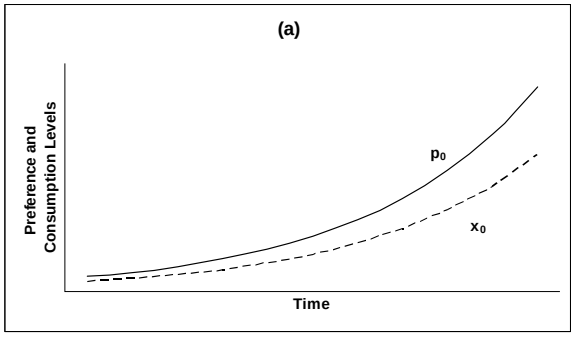

Figure 2 (a) The development of an individual's preference and consumption levels when indirect bias is greater than direct bias.

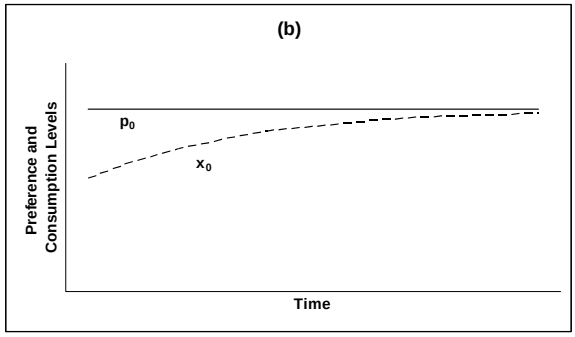

Figure 2 (b) The development of an individual's preference and consumption levels when there is no indirect bias.

To summarize our findings derived from the model, we offer the following propositions.

Proposition 1: If the preference level of an individual consumer were unaffected by indirectly biased social learning processes, following from Equation (11), $\bar{x}$ would be at equilibrium at $\bar{x} = \bar{p}_{unbiased}$. Thus, if $\bar{x} < \bar{p}_{unbiased}$ the mean value of a population's consumption level would increase; if it is greater than $\bar{p}_{unbiased}$, it would decrease.

Proposition 2: If consumers' mean consumption level, measured by $\bar{x}$, reaches their mean preference level, $\bar{p}$, they temporarily meet their corresponding needs. This state is, however, of a transitory nature only if $\bar{p}$ is continuously modified by processes of biased social learning via cultural role models. The individuals then feel deprived again and strive to adapt their consumption level to this new preference level determining the new level of consumption perceived as appropriate. However, the additional resources spend on reaching the new consumption level still only lead to satisfying one's (modified) consumption preference as it did on lower levels of consumption.

4. Towards a procedural concept of well-being

This Section presents an account of well-being that builds upon the insights on processes of cultural learning and preference formation discussed above. The model of cultural transmission informs us about the conditions under which the dynamic interplay of two learning biases can generate a treadmill of self-augmenting consumption cycles. This allows us to draw a line between those learning paths that are in dynamic equilibrium and those paths that depart from this equilibrium in an accelerating way, leading the agent into an extreme state of permanent dissatisfaction where, ultimately, exogenous (economic or ecological) constraints become binding and the latent learning dynamics cannot unfold anymore.

We suggest to use this model for normative purposes in the following way: processes of preference learning can either "work" or "fail to work" effectively. In the former case, the representative agent's gradual updating of her preference level (recall that this is a cardinal variable) settles down in a dynamic equilibrium that allows her to satisfy her preferences by matching her preference path with a corresponding consumption path (again, note that her consumption level, at any point in time, is represented by some cardinal variable). This is the process depicted in Figure 1(a). In the latter case, the agent continuously modifies her preference levels in such a way that she systematically fails to successfully satisfy her preference level (and, hence, fails to increase her well-being). If we interpret preferences as instrumental tools that should ideally enhance the agent's well-being, we may say that the agent, driven by the interplay of our two biases, ends up systematically acquiring dysfunctional tools (as shown in Figure 2(a)).

Our model of cultural transmission allows us to operationalize Elster's intuitive idea that in a world of preference endogeneity we should inquire into the historical genesis of the agents' preferences (see Section 2, above) in order to be able to distinguish between desirable and undesirable ways of preference learning. In Elster's original paper, he mentions "adaptive preference formation" and the "deliberative manipulation of wants by other people" as cases of undesirable preference change (1982, p. 223, pp. 226-230). In his account, the label "undesirable" refers to the presumed lack of "autonomy" in the acquisition of preferences. In contrast, we focus on an empirically particularly relevant case of social conditioning that involves our specific two biases. The model can be used to track the genesis of an agent's preferences by identifying the way these biases interact in driving the agent's process of preference formation. Based on the

empirically observed differential impact of these two biases, it is then possible to suggest a normative distinction between two kinds of preference dynamics.

This distinction can be made normatively relevant by suggesting the following notion of individual well-being: rather than being defined as the satisfaction of "given" and perfectly consistent preferences, in our setting well-being should be understood as the ability to engage in the effective learning of new preferences over time, with the "effective" proviso reflecting the absence of systematical frustration on the part of the agent. "Frustration" is to be understood in terms of our model of cultural learning, as the systematic non-satisfaction of preferences, due to the ongoing process of preference updating, fuelled by the influence of social role models, in combination with a consumption level that does not grow at the same pace. Consequently, although the individual's level of consumption rises, her preferences are not satisfied, since the agent is "never satisfied" with what she is able to attain. The effective learning of new preferences constitutes well-being in a partly procedural sense. It is partly inherently valuable, i.e., independent of the outcomes in terms of preference satisfaction, and its value is independent of the logical consistency of an agent's vector of preferences at any given point in time. Endogenous preferences turn out to be naturally inconsistent, so in our view it does not make sense to treat inconsistency as an argument for discounting the value of an agent's preferences.[17]

In light of this notion of well-being, status-oriented consumption qualifies as a policy problem to the extent that it generates learning dynamics that ultimately (and, perhaps, paradoxically) jeopardize the individuals' ability to "try out" new preferences in an ongoing, sustainable way. At the level of instrumental policy-making, it follows that policy should at least not restrict, but preferably foster the ability of individuals to engage in the effective learning of new preferences. In order to achieve this, it seems that two general conditions have to be met: first, agents should obviously be left free to acquire new preferences as they see fit. They should, then, be endowed with the maximum set of opportunities that is compatible with everyone else's opportunity set. Note that this first condition is identical to Sugden's criterion of well-being as "opportunity" (see above). In our framework, it is however not sufficient: our model of cultural transmission predicts that preference formation may be dynamically unstable under certain circumstances. These circumstances are targeted by our second condition: policy should design the institutional framework of markets in such a way that the tendency of preference dynamics to

[17] See Sugden (2004) for a related argument.

"collapse" by inducing self-augmenting cycles will at least not be promoted, but preferably reduced. Prima facie, this may be achieved by a variety of tools that either involve "hard" paternalism (involving taxation of status goods, say) or milder LLSBA interventions that aim directly at influencing consumers' preferences.

How can our notion of well-being be *justified*? We propose to follow Sugden (2008) in adopting a contractarian perspective that frames the issue of evaluating alternative policy rules in terms of a mutual "constitutional" agreement between the individuals that will be affected by the rules. In order to operationalize this approach, assume that all those individuals are put behind a Rawlsian "veil of ignorance" where they do not know anything about the preferences they themselves will adopt in the future "market game" with their periods of cultural transmission.[18] Assume furthermore that they *do* know about the characteristics of the model of cultural learning introduced in Section 3, above, i.e., they know about the conditions that may undermine the ongoing learning of new preferences and "trap" them in self-defeating cycles of status consumption. Given these assumptions, we claim that those individuals may agree upon assessing alternative policy rules in light of our notion of well-being. This claim would be refuted if individuals would, through a process of public deliberation, reject the idea to guide and structure policy with the help of such a notion.[19]

This claim is backed by the following proposition: Our notion of well-being satisfies all the criteria that have been posited in Section 2, above: First, it is not detached from what real-world individuals are able to care about and effectively do care about. Being an essential contributing factor to an individual's identity and sense of self, preference learning (and the ability to pursue it) can be argued to be even one of the most important things in life people care about. Hence, we argue that our concept of well-being satisfies the "internalism" condition (1).

Second, our notion of well-being allows for individual preferences to be inconsistent and departing from what is in the agent's "best interests". In our context, this means that under certain circumstances the agent may willingly acquire preferences that are self-defeating from an ex post perspective. The possible dissociation between preference, choice and well-being is build into our approach to reasoning about well-being. That is why condition (2) is satisfied.

[18] See Witt and Schubert (2008) for a related approach.
[19] This is the usual way to frame the role of normative arguments in contractarian theory, see, e.g., Vanberg (2006).

Third, our notion of well-being is explicitly tailored to a world where preferences are not "given" once and for all, but change endogenously. Specifically, they are subject to "natural" processes of social conditioning, which in our model come in the shape of two kinds of biases and their dynamic interplay. In this context, the main advantage of our concept of well-being lies in its procedural nature which reflects the procedural account of preferences supported by our positive background model. Hence, condition (3) is satisfied.

Fourth, by virtue of requiring to give agents maximum opportunities to engage in preference learning (subject only to freedom-preserving LLSBA interventions to prevent "traps" in the dynamics of preference learning, see above), our notion of well-being includes considerations related to the instrumental and/or inherent value of individual freedom and autonomy. This allows us to conclude that condition (4) is satisfied as well.

At a more applied level, our procedural notion of well-being (with its positive background model of cultural learning) may be used to assess the desirability of the usually proposed sets of policy tools to shape people's consumption of status goods. Consider "hard" paternalism first. Contrary to a widely shared view in the literature on conspicuous consumption and "positional externalities" (Frank, 2008), our model shows that the taxation of status goods is ineffective in reducing the consumption of status-signaling goods and the accompanying participation in "status races". Recall the key relationship between $\bar{x}$ and $\bar{p}$ (Figure 1, above). If $\bar{x} < \bar{p}$, a tax imposed on the consumption of the good in question would merely increase the difference between $\bar{x}$ and $\bar{p}$ without removing the consumers' general motivation to meet their preference level by increasing their personal consumption levels and, thus, finally $\bar{x}$. The role model-based acquisition of modified preference levels via social learning remains unaffected. In the case of taxation, "status races" would still take place by means now affordable within the boundaries of the new economic restrictions. Hence, this kind of consumption dynamic may occur on the basis of rather different classes of commodities depending on the available income. The deeper motivational factors driving status-seeking would remain unaffected.

Hence, interventions that target processes of preference and social norm formation appear to be the best (and probably only) tools available to counter self-defeating preference dynamics. We will briefly consider two kinds of policies that shape preference formation (the proper realm of LLSBA paternalism), two kinds of policies that shape social norm formation by targeting the "social meaning" of activities (Lessig, 1995), and finally the design of institutional frameworks

that satisfy the individuals' need for self-esteem by providing procedural utility (Frey et al., 2004).

As to the first set of policies, we suggest two kinds of interventions. First, policy may influence agents' preferences by providing information about the characteristics and "side-effects" of consumer goods that will not be provided in the marketplace. This information provision may proceed on more or less neutral terms, i.e. involve the "coloring" of information by using framing effects or other tricks from the behavioral economics toolbox. Our criterion of well-being would endorse the non-neutral provision of information as long as agents are not made subject to manipulation "behind their back", i.e., as long as they can *in principle* (i.e., if they explicitly wish to do so) realize the degree of non-neutrality and act against the policy-maker's goals.[20]

Second, given that, (i), an important precondition of status consumption and the ensuing status races is the individuals' ability to incur debt and that, (ii), widespread overborrowing is a key effect of people's participation in status races, we argue that LLSBA instruments should be used to discourage people from behavior that leads to excessive indebtedness. There is ample empirical evidence that the manipulation of default options can effectively influence people's attitudes towards saving and dissaving (Thaler and Sunstein, 2008).

As to the second set of policies, targeting processes of social norm formation, we again suggest two specific interventions. They aim at influencing the "social meaning" of activities, which Lessig (1995, p. 951) defines as the locally uncontested "semiotic content attached to various actions, or inactions, or statuses, within a particular context". They are derived from socially shared understandings or expectations concerning the appropriate course of action in specific situations. To illustrate, consider "what it means" to be black in the American South in the 1960s, to have a whisky in the office in the early 1960s, to have the same whisky in the office in today's business world, or to smoke as a woman in the early 20th century. Within limits, policy may use the conservation or manipulation of social meanings to achieve certain aims, such as the aim to discourage status-oriented consumption. First, it may regulate advertisement by banning the direct appeal to people's role-model bias in the media. Second, it may exploit the very same role model bias by promoting public campaigns that associate prestigious (prominent and

[20] See Bovens (2008) for a similar suggestion.

glamorous) role models with an attitude that is immune against the lure of status good consumption (by appealing to people's "autonomy" or "self-determination", say).

Finally, a third set of policies could be implemented to confront directly one of the key motivational drivers of status good consumption, namely people's needs for self-esteem, respect, and a positive self-image (e.g. Johansson-Stenman and Martinsson, 2006). Happiness research tells us that the very same set of needs can also be satisfied by organizing human relationships within the political arena or at the workplace in such a way that people gain "procedural utility" (Frey et al., 2004). This may be a supplementary step towards reducing individuals' desire to satisfy these needs through expensive status good consumption, hence to decrease the demand for status good consumption, the engagement in "wasteful" status races, and the risk of decreasing individuals' well-being in terms of their ability to self-improve by learning new preferences.

5. Concluding Remarks

While behavioral economics provides ample evidence for the fact that in many economically relevant situations real-world individuals may systematically fail to choose what is in their best interests (as perceived by themselves), little progress has been made to derive convincing policy implications from this. We have argued that this is due to the fact that we cannot formulate policy advice without a concept of well-being or welfare, and that the policy-minded literature in the field still lacks a concept of well-being which could be consistently applied in a "behavioral" setting, i.e., in a setting where individual preferences change and tend to be inconsistent.

Our contribution to the literature has been twofold. First, we have tried to facilitate the quest for a "behaviorally adequate" non-standard concept of well-being by introducing a catalog of four conditions that any such concept ought to satisfy. We have used this catalog to gauge the usefulness of those (four) non-standard notions of well-being that have so far been discussed in the literature. In our view, this exercise already sheds light on the contours of a practically applicable concept (for instance, it tells us that such a concept will certainly be of a procedural nature). Second, following the intuition that additional insights into the process of individual preference formation may shed further light on this issue, we have introduced a formal model of preference learning. The model predicts that the interplay of two specific biases can generate self-

augmenting learning dynamics that lead agents away from sustainable and welfare-enhancing consumption paths. This prediction allowed us to make a distinction between "desirable" and "undesirable" paths of preference learning. This distinction is, thus, not axiomatic, but grounded in empirical evidence, and it can form the basis of a procedural notion of well-being. According to this view, well-being is seen as residing in the individual's ability to engage in the ongoing, not systematically frustrated ability to acquire new preferences. Hence, status races call for policy intervention to the extent that they risk decreasing well-being, thus defined. We hasten to add an important caveat: this policy intervention should not be organized in a top-down way, by some omniscient social planner. Rather, we have suggested to frame the discussion in terms of the contractarian paradigm: our concept of well-being should be used to inform the individuals themselves (as citizens) about possible ways to think about the problem of status races and their impact on well-being. On the basis of this information, the individuals may agree upon a social contract, stipulating a set of rules that help them avoid the welfare losses associated with status races, while essentially maintaining their personal freedom of choice. Think of it as the analog of a set of traffic rules: they do not prescribe any particular activity (they don't tell you where to go), but they channel interdependent – and potentially harmful – individual behavior in such a way that harmful accidents are avoided as far as possible.

(january 26, 2011)
(approx. 10.080 words)

References

Anand, P., Gray, A., 2009. Obesity as Market Failure: Could a 'Deliberative Economy' Overcome the Problems of Paternalism? Kyklos 62, 182-190.

Ariely, D., Loewenstein, G., 2006. The Heat of the Moment: The Effect of Sexual Arousal on Sexual Decision Making. Journal of Behavioral Decision Making 19, 87-98.

Ariely, D., Loewenstein, G., Prelec, D., 2006. Tom Sawyer and the Construction of Value. Journal of Economic Behavior & Organization 60, 1-10.

Aronson, E., Wilson, T. D., Akert, R. M., 2002. Social Psychology.Prentice-Hall, Upper Saddle River.

Atkinson, A.B., 2009. Economics as a Moral Science. Economica 76, 791-804.

Beaulier, S., Caplan, B., 2007. Behavioral Economics and Perverse Effects of the Welfare State. Kyklos 60, 485-507.

Benjamin, D.J., Laibson, D.I., 2003. Good Policies for Bad Governments: Behavioral Political Economy. In: Federal Reserve Bank of Boston (ed.), How Humans Behave. FRBB Conference Series No. 48.

Berg, N., 2003. Normative Behavioral Economics. Journal of Socio-Economics 32, 411-427.

Bernheim, B.D., 2009. Behavioral Welfare Economics. Journal of the European Economic Association 7, 267-319.

Bernheim, B.D., Rangel, A., 2009. Beyond Revealed Preference: Theoretic Foundations for Behavioral Economics. Quarterly Journal of Economics 124, 51-104.

Bovens, L., 2008. The Ethics of Nudge. In: Grüne-Yanoff, T., Hansson, S.O. (eds), Preference Change: Approaches from Philosopgy, Economics and Psychology. Springer, Berlin, pp. 207-220.

Boyd, R., Richerson, P. J., 1985. Culture and the Evolutionary Process. University of Chicago Press, Chicago.

Brandt, R., 1979. A Theory of the Good and the Right, Clarendon Press, Oxford.

Brickman, P., Campbell, D.T., 1971. Hedonic Relativism and Planning the Good Society. In: Apley, M.H. (ed.), Adaptation-Level Theory: A Symposium, Academic Press, New York, pp. 287-302.

Camerer, C., Issacharoff, S., Loewenstein, G., O'Donoghue, T., Rabin, M., 2003. Regulation for Conservatives: Behavioral Economics and the Case for 'Asymmetric Paternalism. University of Pennsylvania Law Review 151, 1211-1254.

Cavalli-Sforza, L.L., Feldman, M.W., 1981. Cultural Transmission and Evolution: A Quantitative Approach, Princeton University Press, New Haven.

Clark, D.A., 2006. The Capability Approach: Its Development, Critiques and Recent Advances. In: Clark, D.A. (ed.), The Elgar Companion to Development Studies, Edward Elgar, Cheltenham, pp. 32-45.

Cordes, C., 2005. Veblen's 'Instinct of Workmanship,' its Cognitive Foundations, and Some Implications for Economic Theory. Journal of Economic Issues 39, 1-20.

De Fraja, M., 2009. The origin of utility: Sexual selection and conspicuous consumption. Journal of Economic Behavior and Organization 72, 51-69.

Di Tella, R., MacCulloch, R.J., Oswald, A.J., 2001. Preferences over Inflation and Unemployment: Evidence from Surveys of Happiness. American Economic Review 91, 335-341.

Duesenberry, J.S., 1949. Income, Savings, and the Theory of Consumer Behaviour. Harvard University Press, Cambridge.

Dworkin, G., 2010. Paternalism. Stanford Encyclopedia of Philosophy, available at: http://www.seop.leeds.ac.uk/entries/paternalism.

Eaton, B.C., Eswaran, M., 2009. Well-Being and Affluence in the Presence of a Veblen Good. Economic Journal 119, 1088-1104.

Elster, J., 1982. Sour Grapes - Utilitarianism and the Genesis of Wants. In: Sen, A.K., Williams, B.A. (eds.), Utilitarianism and Beyond, Cambridge University Press, Cambridge, pp. 219-238.

Fisher, R.A., 1930. The Genetical Theory of Natural Selection. Clarendon Press, Oxford.

Frank, R.H., 1999. Luxury Fever, New York: The Free Press.

Frank, R.H., 2008. Should Public Policy Respond to Positional Externalities? Journal of Public Economics 92, 1777-1786.

Frederick, S., Loewenstein, G., 1999. Hedonic Adaptation. In: Diener, E., Kahneman, D., Schwarz, N. (eds.), Well-Being: The Foundations of Hedonic Psychology. Russell Sage Foundation, New York, pp. 302-329.

Frey, B.S., Stutzer, A., 2002. What Can Economists Learn from Happiness Research? Journal of Economic Literature 60, 402-435.

Frey, B.S., Benz, M., Stutzer, A., 2004. Introducing procedural utility: not only what, but also how matters. Journal of Institutional and Theoretical Economics 160, 377-401.

Friedman, D., Ostrov, D.N., 2008. Conspicuous consumption dynamics. Games and Economic Behavior 64, 121-145.

Glaeser, E., 2006. Paternalism and Psychology. University of Chicago Law Review 73, 133-156.
Griffin, J., 1986. Well-Being. Clarendon Press, Oxford.
Grüne-Yanoff, T., 2009. Welfare Notions of Soft Paternalism. Papers on Economics & Evolution, #0917, Max Planck-Institute of Economics, Jena, Germany.
Harrington Jr., J.E., 1999. Rigidity of Social Systems. Journal of Political Economy 107, 40-64.
Harsanyi, J.C., 1982. Morality and the Theory of Rational Behavior. In: Sen, A.K., Williams, B.A. (eds.), Utilitarianism and Beyond. Cambridge University Press, Cambridge, pp. 39-62.
Henrich, J., Boyd, R., 2002. On Modeling Cognition and Culture. Journal of Cognition and Culture 2, 87-112.
Hirsch, F., 1976. Social Limits to Growth. Harvard University Press, Cambridge.
Hopkins, E., Kornienko, T., 2004. Running to Keep in the Same Place: Consumer Choice as a Game of Status. American Economic Review 94, 1085-1107.
Ireland, N.J., 2001. Optimal income tax in the presence of status effects. Journal of Public Economics 81, 193-212.
Johansson-Stenman, O., Martinsson, P., 2006. Honestly, why are you driving a BMW? Journal of Economic Behavior and Organization 60, 129-146.
Kahneman, D., Wakker, P.P., Sarin, R., 1997. Back to Bentham? Explorations of Experienced Utility. Quarterly Journal of Economics 112, 375-405.
Kirkpatrick, M., 1982. Sexual Selection and the Evolution of Female Choice. Evolution, 36, 1-12.
Lande, R., 1981. Models of Speciation by Sexual Selection on Polygenic Traits. Proceedings of the National Academy of Science 78, 3721-3725.
Layard, R., 2005. Happiness: Lessons from a New Science. Penguin, London.
Lessig, L., 1995. The Regulation of Social Meaning. University of Chicago Law Review 62, 943-1045.
Loewenstein, G., Haisley, E.C., 2008. The Economist as Therapist: Methodological Ramifications of 'Light' Paternalism. In: Caplin, A., Schotter, A. (eds.), The Foundations of Positive and Normative Economics. Oxford University Press, Oxford, pp. 210-245.
McElreath, R., Boyd, R., 2007. Modeling the Evolution of Social Behavior: A Guide for the Perplexed. University of Chicago Press, Chicago.
Ng, Y.-K., 1987a. Diamonds are a Government's Best Friend: Burden-Free Taxes on Goods Valued for Their Values. American Economic Review 77, 186-191.
Ng, Y.-K., 1987b. Relative income effects and the appropriate level of public expenditure. Oxford Economic Papers 39, 293-300.
Ng, Y.-K., 1999. Utility, Informed Preference, or Happiness: Following Harsanyi's Argument to its Logical Conclusion. Social Choice & Welfare 16, 197-216.
Ng, Y.-K., 2003. From Preference to Happiness: Towards a More Complete Welfare Economics. Social Choice and Welfare 20, 307-350.
Norenzayan, A., Heine, S.J., 2005. Psychological Universals: What Are They and How Can We Know? Psychological Bulletin 131, 763-784.
Nussbaum, M., 2000. Women and Human Development: The Capabilities Approach. Cambridge University Press, Cambridge.
Parisi, F., von Wangenheim, G., 2006. Legislation and Countervailing Effects from Social Norms. In: Schubert, C., von Wangenheim, G. (eds.), Evolution and Design of Institutions, London: Routledge, pp. 25-55.
Price, G.R., 1970. Selection and Covariance. Nature 227, 520-521.
Rege, M., 2008. Why do people care about social status? Journal of Economic Behavior and Organization 66, 233-242.
Richerson, P. J., Boyd, R., 2005. Not by Genes Alone: How Culture Transformed Human Evolution. University of Chicago Press, Chicago.
Samuelson, P.A., 1948. Consumption Theory in Terms of Revealed Preference. Economica 60, 243-253.
Sen, A.K., 1980. Equality of What. In: McMurrin, S.M. (ed.), The Tanner Lectures on Human Values. University of Utah Press, Salt Lake City, pp. 197-220.
Sen, A.K., 1985. Commodities and Capabilities. Basic Books, New York.
Sen, A.K., 1987. On Ethics and Economics. Basil Blackwell, London.
Sen, A.K., 1999. Development as Freedom. Knopf, New York.
Sen, A.K., 2006. Reason, Freedom and Well-Being. Utilitas 18, 80-96.
Sobel, D., 1994. Full Information Accounts of Well-Being. Ethics 104, 784-810.
Sugden, R., 2004. The Opportunity Criterion: Consumer Sovereignty Without the Assumption of Coherent Preferences. American Economic Review 94 1014-1033.

Sugden, R., 2006. What We Desire, What We Have Reason to Desire, Whatever We Might Desire: Mill and Sen on the Value of Opportunity. Utilitas 18, 33-51.
Sugden, R., 2008. Why Incoherent Preferences Do Not Justify Paternalism. Constitutional Political Economy 19, 226-233.
Sugden, R., 2009. On Nudging: A Review of Nudge: Improving Decisions about Health, Wealth, and Happiness, International Journal of the Economics of Business 16, 365-373.
Sugden, R., 2010. Opportunity as Mutual Advantage. Economics & Philosophy 26, 47-68.
Sumner, L.W., 1996. Welfare, Happiness and Ethics. Clarendon Press, Oxford.
Sunstein, C.R., Thaler, R.H., 2003. Libertarian Paternalism Is Not an Oxymoron. The University of Chicago Law Review 70, 1159-1202.
Thaler, R.H., Sunstein, C.R., 2008. Nudge: Improving Decisions about Health, Wealth, and Happiness. Princeton University Press, New Haven.
van Praag, B.M.S., Barsma, B.E., 2005. Using Happiness Surveys to Value Intangibles: The Case of Airport Noise. Economic Journal 115, 224-246.
Vanberg, V.J., 2006. Human intentionality and design in cultural evolution. In: Schubert, C., von Wangenheim, G. (eds), Evolution and Design of Institutions. Routledge, London, pp. 197-212.
Veblen, T., 1898. The Theory of the Leisure Class. MacMillan, New York.
Witt, U., Schubert, C., 2008. Constitutional Interests in the Face of Innovations: How Much Do We Need to Know About Risk Preferences? Constitutional Political Economy 19, 203-225.