The Case for Ends Paternalism: Extending Le Grand and New’s Framework for Justification of Government Paternalism

Dan Acland

ABSTRACT

Le Grand and New, in their recent book, “Government Paternalism: Nanny State or Helpful Friend,” present a novel definition of paternalism and a framework for thinking about whether any given paternalistic policy can be considered justifiable. I show that their framework is flawed in that it restricts justifiable paternalism to that which is intended to alter individuals’ judgment about the means they use to pursue their self-determined ends. I show that the principles they use to justify certain kinds of means paternalism also justify certain kinds of ends paternalism. In particular, when there is a body of rigorous social-science evidence that individuals select ends that they themselves, if they had adequate information or experience would prefer not to pursue, and when other conditions are met, ends paternalism may be considered to improve the wellbeing of the individual as determined by the individual themselves. I present examples of policies that could be justified under this framework, and offer cautionary notes.
1 Introduction

In *Government Paternalism: Nanny State or Helpful Friend*,\(^1\) Julian Le Grand and Bill New (hereafter LN) provide what they consider to be an unambiguous and comprehensive definition of government paternalism, upon which they build a framework for analyzing specific paternalistic policies to determine whether they can be considered justifiable. Their analytical framework is a timely and important contribution to the conversation about government paternalism that has been revived by behavioral economists. It can be thought of as providing a clear and practicable tool for determining when paternalism can be justified, and, equally if not more importantly, when it cannot. But their analytical framework is incomplete.

A key element of LN’s framework is the distinction between “ends paternalism” — intended to shift people away from the ends they choose to pursue in their quest for well-being and life satisfaction and towards those that the paternalist believes they should pursue — and “means paternalism,” which is intended merely to improve upon the means individuals use to pursue their own chosen ends. LN argue that if means paternalism is based on solid social-scientific evidence that individuals, as a result of systematic “reasoning failures,” fall short of optimally achieving their own chosen ends in specific situations, then policies intended to help them overcome those reasoning failures should be considered justifiable, provided that the cost of the inevitable infringement on personal autonomy is outweighed by the benefit of getting people closer to their goals. Ends paternalism, they claim, cannot be justified in this way, as there is no evidence of systematic reasoning failure with respect to ends, and indeed, in their view, no way to imagine what form such evidence would take.

But the distinction between ends and means paternalism does not provide the justificatory bright line LN believe it does, because, as I will show, there are well-documented cases of reasoning failure that cause people to fail to optimally identify the set of ends that will bring them the greatest happiness and life satisfaction, as defined by themselves. I will argue that such “ends-related” reasoning failures, mostly having to do with our systematic failure to accurately forecast our affective response to future conditions and outcomes, should be considered *a priori* grounds for consideration of paternalistic government policy, in just the same way as the means-related reasoning failures LN discuss. The cases of ends-related reasoning failure I have in mind fall into the category LN refer to as “limited imagination,” broadly defined as “limited ability to imagine or predict [one’s] utility in alternative situations or at different times.”\(^2\)

LN understand limited imagination to be a source of exclusively means-related reasoning failure, but in this they err. They do correctly show how

---

\(^1\)Le Grand and New (2015).
\(^2\)Le Grand and New (2015, p. 90).
certain kinds of limited imagination can cause individuals to systematically fail to achieve their own self-identified ends. They cite, for example, the accumulation of evidence that many people systematically fail to predict, or fully appreciate, the fact that they will in all likelihood have the same taste for immediate gratification in the future that they have today, and thus will be no more likely to complete an aversive task with long-term benefits tomorrow than today, a reasoning failure that can lead to very costly procrastination (O’Donoghue and Rabin, 1999). But LN do not recognize that other forms of limited imagination can just as easily cause individuals to pursue the wrong ends, by which I mean a set of ends, or a balance among competing ends, which, even if successfully achieved, fails to maximize the individuals’ own well-being, as defined by the individuals themselves. Or rather, the individuals’ own wellbeing as they themselves would define it, if they were not subject to limited imagination. And as with the forms of limited imagination that can cause people to have the wrong means, the kinds of limited imagination that can cause people to have the wrong ends have also been well documented by psychologists and behavioral economists.

As an example, an addict who has no experience with a life of successful abstinence and long-term sober living, and who is unable to imagine how rewarding such a life could be, may have a set of ends that does not include achieving sustained abstinence, or that places a low value on that goal. In this case LN would say that a paternalist is unjustified in calling the addict’s self-identified ends into question, but only whatever means they use to pursue them. I will argue, to the contrary, that the addict, due to limited imagination, may simply have failed, as yet, to identify their most preferred ends. If they were to experience the rewards of sober living, and thus overcome their limited imagination, they might very well discover that they actually have a clear preference for the end of maintaining abstinence and a sober way of life over the end of experiencing an addictive high and suffering the discontents and depredations of a life of active addiction. If so, we would be able to observe, by asking them, that they themselves felt they had previously been pursuing the wrong ends, as defined by them. An ends-paternalistic policy intended to move them from addiction to sobriety could then be considered potentially justifiable, on precisely the same grounds, and subject to precisely the same conditions, that LN apply to means paternalism.

An objection can be raised that the so-called “end” of sober living, while perhaps the proximal goal of cessation behaviors, is actually just a means to the higher end of overall life satisfaction, and that in fact for any given individual

---

3I am not referring to failure to imagine what the withdrawal process or early period of abstinence would be like, but failure to imagine what a long-term sober lifestyle might be like — typically replete with fundamental shifts in perspective and life orientation, and concomitant dramatic improvements in life outcomes and life satisfaction, all of which typically develop over a period of months or years.
there exists what we might call a “hierarchy of ends” in which anything we define as an end that an individual pursues can always be redefined as a means to some higher end, in a potentially infinite regress. If we take this position in the case of the addict above, then the LN framework for analyzing means paternalism automatically extends beyond the choice of which means to use — in pursuit of either the end of sober living, or the end of active addiction — to the choice between the pursuit of sober living and the pursuit of active addiction, as means to the higher end of life satisfaction. Indeed, if we take this approach, in which any goal we choose to refer to as an end can be seen as a means to a higher end, we might find ourselves on a slippery slope indeed, because then LN’s framework could be applied to the consideration of any paternalistic policy whatsoever, however far up the hierarchy of ends it is targeted. In some sense we would have defined true ends paternalism out of existence.

In fact, this is no objection at all. It is exactly my point: when an individual chooses to pursue what LN refer to as an end, they are essentially choosing a strategy for achieving whatever they define as success in life, whether it be “life satisfaction,” “subjective well-being,” or simply “being a mensch.” And if the individual is manifestly wrong about the efficacy of their chosen strategy, then they are going to fall short of achieving the highest degree of success for themselves, as defined by themselves. I have chosen to maintain the distinction between means and ends for the same reason that I believe LN chose to do so. It is my belief that in practice, in any given policy situation, there will be broad agreement among parties to the policy debate as to what should be thought of as means and what should be thought of as ends, and thus broad agreement about what constitutes means paternalism and what constitutes ends paternalism. The goal of my paper, and I believe of LN’s book, is to provide practical guidance to people in the real world who are charged with making the determination of whether there is a valid case to be made that some paternalistic policy or other is in society’s best interest, all things considered, whether it is thought of as means paternalism or ends paternalism. And because their framework actually imposes a very high standard for justifying paternalistic policies, even when extended to ends paternalism, I take their book, and my paper, to be steps towards keeping society from sliding down the slippery slope.

The remainder of the paper proceeds as follows. In Section 2, I summarize the LN framework for means paternalism, and elucidate its key principles, in order to be able to extend them to ends paternalism. In Section 3, I examine the evidence on limited imagination in greater detail, and tease apart the cases in which it leads to means-related reasoning failure, from those in which the reasoning failure is actually with respect to the identification of the individual’s ends themselves. Section 4 systematically applies the principles of the LN framework to the case of ends-related reasoning failure, and shows that the
entire framework can be extended directly to the consideration and justification of ends paternalism. Section 5 concludes.

2 Summary of the LN Framework

LN begin by taking a stand on the definition of paternalism. Existing definitions, they explain, typically consist of three parts. A policy is paternalistic if, (1) it interferes with individuals’ freedom, (2) is intended for the good of those individuals, and (3) is implemented without the consent of those individuals. In some cases interference with freedom is replaced by interference with autonomy, which some argue is a more comprehensive principle, inherently subsuming interference with freedom while additionally including things like subsidies for museum attendance — intended to help people find greater satisfaction by pursuing “higher” ends than they otherwise would — which some argue interfere with individuals’ autonomy (or “self-rule”) even while increasing their options, and thus not strictly interfering with their freedom. LN argue that this approach to defining paternalism is problematic, because it is based on concepts (freedom and autonomy) that are irresolvably ambiguous, and thus does not provide a reliable way to distinguish between what is and what is not paternalism.

Instead, LN base their definition on the idea that at its core, paternalism is about questioning the judgment of the individual, on the presumption that the paternalist has a better judgment of what will be good for him or her, or how to pursue it. Their definition is that a policy is paternalistic if 1) it is intended to replace the judgment of the individual with the judgment of the policy maker, and 2) it is intended to do so for the good of the individual. Clearly any policy that interferes with freedom for the individual’s benefit satisfies the judgment-replacement criterion. But more generally, any policy intended to influence individuals’ choices or outcomes for their own good in any way satisfies the criterion, whether or not the policy interferes with freedom. Subsidies for museum attendance are unambiguously covered by this definition, as their intention is to induce people to engage in an activity that they otherwise would not, and which the paternalist thinks they should. True also for policies or actions with even subtler impacts on freedom and autonomy, such as the canonical case of the doctor who, having been asked by a dying patient whether

---

4LN dispense with the “without consent” stipulation on the basis that it is obviated by their judgment-replacement criterion. If an individual’s judgment is defined as what the individual believes to be in his or her best interest, it would be “logically incoherent,” according to LN, for the individual to consent to having that judgment replaced with the paternalist’s, because the replacement would necessarily go against what the individual believed to be in their best interest in some way, a belief upon which consent is presumably conditioned.
his daughter has survived a car accident, withholds the information that she
has not. The ambiguity about whether the doctor has infringed upon the
patient’s autonomy has left scholars in doubt as to whether this should be
considered paternalistic. LN’s definition rules it unambiguously in.

The attempt to establish an unambiguous and comprehensive definition
of paternalism is an important contribution in itself, but perhaps more im-
portantly, LN’s emphasis on judgment provides the basis for their approach
to determining when paternalism can be considered justified. Simply put,
their view is that replacing another’s judgment with your own for their own
good can only be justified if you have rigorous social-scientific evidence that
their judgment is flawed, and that they are failing or will fail to maximize
their own wellbeing, as defined by themselves as a result. LN review the
kind of flaws in judgment that psychologists and behavioral economists have
been exploring and documenting for the past several decades and sort them
into four categories of “reasoning failure”: limited technical ability, limited
imagination or experience, limited willpower, and limited objectivity. The
key point in each case is that social scientists have found clear evidence that
these reasoning failures result in choices that are not in the best interest of the
individual, as defined by the individual. As Thaler and Sunstein (2003) put
it, “in some cases individuals make inferior choices, choices that they would
change if they had complete information, unlimited cognitive abilities, and
no lack of willpower.” In effect, LN have salvaged what might be considered
the essence of consumer sovereignty by requiring that the individual’s own
preferences are what matters, even if we reject that they themselves know
best how to satisfy those preferences. And since ends paternalism, by its very
definition, seeks to replace the individual’s judgment about which ends to
pursue, i.e. their preferences, LN conclude that only means paternalism can
be justified.

LN’s framework is completed by the observation that to say that it can be
justifiable to replace an individual’s demonstrably flawed judgment with one’s
own is only to establish necessary conditions. To establish sufficient conditions
for any particular paternalistic policy to actually be justified in any particular
case, LN return to the consideration of autonomy. Scholars have debated
whether there can be any form of paternalism, or any specific application of
paternalism, that does not infringe upon autonomy in some harmful way. LN
take the position that there cannot, even under the currently popular idea that
Libertarian Paternalism, with its “nudge” approach to changing individual
behavior without formally restricting choices, successfully avoids infringing
autonomy. Having concluded that paternalism inevitably comes at a cost to
autonomy, the next question LN address is whether there can be any degree
of infringement of autonomy that can be deemed acceptable in the case of
paternalism. Philosophers since Mill have argued that infringement upon an in-
dividual’s autonomy can only be justified if the individual’s actions cause harm
to others. This is the so-called “harm principle.” If the intention of a policy is simply to make individuals themselves better off, then autonomy should be treated lexicographically: there is no amount of benefit to the individual that can outweigh the slightest infringement upon his or her autonomy. LN demur, and the essence of their case can be thought of as simply that if reasoning failure can be shown to “lead people not to behave in their own interests” then we should think of their behavior as causing harm to themselves, and extend the logic of the harm principle to include the case of self-harm as well as harm to others. Thus, for LN, for a paternalistic policy that satisfies the above-stated necessary conditions for justification, the sufficient condition for justification is the same as in the case of policies intended to address harm to others: we simply ask, “Is the cost in terms of infringement on autonomy outweighed by the benefit in terms of reducing harm, which is to say, in the case of self-harm, the benefit of making the individual better off, according to their own definition of “better off?”” In the next section I present the psychological and behavioral-economic evidence upon which I base my argument that LN’s justification framework can be extended from means paternalism to ends paternalism.

3 Ends-Related Reasoning Failure

Though LN may not have settled the matter of how to define and when to justify paternalism, from the standpoint of behavioral economics and policy analysis I consider LN’s definition and framework to be sound with respect to means paternalism. However, where they err is in thinking that they have settled the means versus ends question more fully than they have. In fact, ambiguity remains regarding the validity of their requirement that in order to be considered potentially justifiable, a paternalistic policy must be intended to address the judgment of individuals only with respect to their means, and not with respect to their ends. LN believe that they have drawn a bright line between the two. First, echoing proponents of asymmetric and libertarian paternalism, they correctly cite, “an increasing volume of evidence from behavioral economics and psychology of what we term ‘reasoning failure’: the fact that individuals, in trying to achieve the end of improving their well-being, often make mistakes and do so in a systematic way.” Then, in defense of their rejection of ends paternalism, they add, “However, there is no similar accumulation of evidence that individuals make mistakes over their ends; that is over the factors that contribute to their well-being. Indeed, since such ends are essentially value-driven, it is hard to see what form such evidence might take.”

\(^5\)Camerer et al. (2003).
\(^6\)Le Grand and New (2015, p. 3).
In fact, however, there is increasing evidence that people have reasoning failure with respect to their ends as well as their means. Indeed, without recognizing it as such, LN themselves cite some of this evidence, under the category of reasoning failure they call limited imagination, which they define as “people’s limited ability to imagine or predict their utility in alternative situations or at different times.” The evidence comes largely from the study of affective forecasting, much of which is reviewed by Wilson and Gilbert (2003), hereafter WG. As WG put it, the question of whether people pursue the right ends in their ultimate pursuit of happiness or satisfaction “depends on whether people can predict accurately which events will make them happy, by how much, and for how long.” LN themselves point in this direction by paraphrasing Burrows (1993) who argued that “thinness of experience”—by which he meant limited direct experience of certain outcomes—can, as LN put it, “constrain a fully rounded preference formation.” They write, “[p]references, or the ends that an individual prioritizes, would be different if the individual had experience (rather than simply abstract knowledge or information) of all the possible outcomes of a certain activity.” This more or less explicitly acknowledges that limited imagination can cause people to have the wrong ends (within which I include both the complete omission of some end from an individuals’ preferences, and the wrong balance of different ends among their preferences, which are very much the same thing to an economist⁷). I argue that the evidence on affective forecasting demonstrates conclusively that people do, in fact, pursue the wrong ends, in the sense that they pursue ends that do not provide them with as much happiness and life satisfaction, as defined by themselves, as they could achieve if they pursued a different set of ends, and that this provides a robust case for the justification of certain kinds of ends paternalism, on precisely the grounds that LN use to justify certain kinds of means paternalism.

To set the stage: in their discussion of limited imagination, LN refer to four examples of systematic failure to accurately imagine (which is to say predict) the future, that may cause people to undermine their own well-being, as defined by themselves. Those examples are: 1) failure to predict how one

---

⁷Economists use the term “preferences” to mean simply an individual’s rank ordering of the desirability of the various outcomes available to them (or more abstractly, their inclination to choose one over another). Consider two individuals with different preferences over watching TV and reading books. The first has no desire whatsoever for reading books, and does not consider it an end worth pursuing, while the second has some interest in reading books, but more interest in watching TV. Now, suppose that both individuals, if they were induced to spend a considerable amount of time reading books, would discover that reading, for them, became much more rewarding with experience, and wound up being a more desirable end than watching TV. To an economist, there is no meaningful distinction between the ways in which the two individuals’ preferences were ex ante not “fully formed.” They both should have been pursuing, with greater fervor than they initially were, the end of reading books.
will respond to temptation in the future — so-called “naiveté” with respect to future self-control problems; 2) failure to predict how one’s valuation for an object may change once one possesses it; 3) failure to predict how one will adapt to, or get used to, positive or negative outcomes; and 4) failure to predict what it would be like if one were either addicted to (for the non-addict) or not addicted to (for the addict) some behavior or substance. All of these can be thought of as failures of affective forecasting, discussed in detail by WG, who enumerate the categories of ways that individuals incorrectly forecast their affective responses to future outcomes or conditions. What LN miss is that in some of their examples, the failures of affective forecasting they cite are likely to cause people to have the wrong ends, not just the wrong means. And there are many more that LN do not cite. A systematic investigation of some of WG’s categories of affective mis-forecasting will clarify and differentiate between the cases in which limited imagination is likely to lead to means failure, and those in which it may cause ends failure. I begin by discussing means failure and then move on to cases of ends failure.

3.1 Limited Imagination as a Form of Means-Related Reasoning Failure

LN correctly explain how some forecasting errors can lead to means failure. I will add others, though this is by no means intended to be an exhaustive list. First, as LN correctly explain, if an individual with self-control problems systematically fails to foresee that their current lack of willpower will in all likelihood persist into the future, they may believe that they will act differently in the future than they are prepared to act in the present. Effectively they say to themselves, “I don’t have the willpower to do (or not do) such-and-such activity now, but I will tomorrow,” when in fact, as they later discover, they are no more likely to be able to overcome their self-control problems in the future than they are in the present. As a result of this naiveté, they may procrastinate activities with immediate costs and long-term benefits (such as signing up for a retirement plan), do too much of activities with immediate benefits and long-term costs (such as eating too much over the weekend under the belief that they will work off the calories in the gym during the work-week), or under-invest in self-control mechanisms (such as opening a retirement-savings account with penalties for early withdrawal, or buying, at greater per-unit cost, only enough ice-cream for a modest weekend binge). When soberly considering the long-term big picture of their preferences and goals, such a person might very well, if they had no other failure of imagination, be able to articulate their over-arching preferences for current versus future consumption, or for tasty food versus diabetes and heart disease. We would not think of their inability to correctly forecast their future self-control problems as causing them to pursue the wrong ends, or the wrong balance of ends. Rather, we would think of it as causing them to fall short of successfully achieving their own self-identified
ends, by using the wrong means to pursue them.\textsuperscript{8} Means-paternalistic policies intended to help people overcome these problems, such as forced retirement savings and taxes on sugary foods, would be potentially justifiable within the LN framework, provided the autonomy cost was outweighed by the benefit to the individual.

The same is true for a second, closely related category of affective forecasting failure which LN do not mention, the so-called “hot-cold empathy gap,” sometimes framed as “projection bias” with respect to “visceral influences.”\textsuperscript{9} A person who is currently hungry may mispredict, while making food choices for the future, how much fatty or sugary food they will want in the future when they are not hungry. A person who is currently not sexually aroused may mispredict how likely they will be to use a condom (or just say no) when they are aroused. A person who is gearing up for a job interview in which they may be subjected to sexually inappropriate comments or questions may mispredict how likely they will be to get up and walk out.\textsuperscript{10} Again, there’s no reason to think this kind of misprediction would cause people to pursue the wrong ends. In a cool state, they could presumably correctly appreciate the effect on their well-being of excessive junk food, unwanted pregnancy, and working under conditions of sexual harassment. But they might use ineffective means to pursue their desired outcomes in those domains. Thus for example, according to LN’s framework, paternalistic policies to restrict the food choices of nutritional-assistance recipients, subsidize long-term birth-control alternatives such as intrauterine devices, or require video-recording and random review of job interviews, would be potentially justified on the basis of means-related limited imagination.

As a final case of means-related affective forecasting failure, LN address the issue of addiction and cessation, saying, “People under appreciate how difficult it will be to kick a drug habit, for example.”\textsuperscript{11} If so, they may take up addictive drugs thinking they will be able to control their use, and later find that their \textit{ex-ante} desired balance of drug-induced euphoria and soul-crushing misery is not actually achievable. Conversely, there may be cases in which already addicted individuals who have a clearly self-identified desire to be free of their addiction mispredict in the opposite direction, thinking that the agonies of early withdrawal will last longer than they actually do, and thus through premature discouragement fail to achieve their clearly identified goal of cessation. In both cases, what is at issue is not the individuals self-identified ends, their preferred balance of drug-induced states and addiction-induced

\textsuperscript{8}It is important to note that it is not the lack of willpower that causes the problem here, but rather the failure to forecast that lack of willpower will persist.
\textsuperscript{9}For an overview see Loewenstein et al. (2003).
\textsuperscript{10}These three examples are studied, respectively, in Read and van Leeuwen (1998), Arielly and Loewenstein (2006) and Woodzicka and LaFrance (2001).
\textsuperscript{11}Le Grand and New (2015, p. 92).
suffering, but their failure to effectively achieve those ends as a result of a failure of affective forecasting regarding the process of addiction and cessation.

### 3.2 Limited Imagination and Ends-Failure: Impact Bias and the Misconstrual Problem

In addition to the means-related cases discussed above, there are a number of ways in which limited imagination, in the form of inaccurate affective forecasting, may lead people to identify not the wrong means with which to pursue their ends, but actually the wrong ends, in the sense that, because they fail to foresee how different outcomes will affect them, they pursue a set of ends, or a balance of ends, that even when successfully achieved brings them less satisfaction or well-being than they could have achieved by pursuing other ends, or a different balance of ends. And contrary to LN’s contention, there is a considerable accumulation of evidence from psychology and behavioral economics in support of these ends-related affective-forecasting errors. Two categories in particular are relevant: “impact bias” and “misconstrual.” As I will explain, these two categories have the common characteristic that they may cause people simply to be wrong about how some future outcome or action will make them feel, or more generally, how much it will contribute to their happiness or life satisfaction. If so, it is hard to imagine how such people, pursuing their currently self-identified ends, could reasonably be expected to achieve the highest level of satisfaction potentially available to them. And, to preemptively fend off a foreseeable libertarian objection, even an infinitely comprehensive process of exploring all possible outcomes would not necessarily solve the problem, because, as WG show, people often continue to make significant affective forecasting errors about outcomes they have experienced many times before.

#### 3.2.1 Impact Bias

The first category of ends-related affective forecasting error WG examine is impact bias, which involves mispredicting how quickly, how much, and for how long some positive or negative experience will affect us. LN touch on this topic when they mention evidence that people “derive less enjoyment from increased salaries, status, or even lottery wins than they thought they would, thus under appreciating the way we ‘adapt’ to our new circumstances.” This is a somewhat simplified rendering of the phenomenon of impact bias, which consists of three dimensions of mis-forecasting of an affective response: the initial intensity, the duration, and the rate at which the response arises. In one of the most rigorous studies, Wilson et al. (2000) found that American-football fans overpredicted how happy they would be immediately after their favorite team won a game, and how long the effect would last. On average subjects predicted significant increases in happiness over at least the first three days.
after the win. In fact they were back to baseline within less than 24 hours. In another study, Gilbert et al. (1998) found that assistant professors facing the tenure decision overpredicted how long both the satisfaction of achieving tenure, and the disappointment of failing would last, by a matter of years. People have been shown to overpredict how happy they will be on vacation, after their preferred candidate wins an election, and after receiving positive feedback on their social aptitude or their desirability as a mate. There is also suggestive evidence that people who have never won a big-stakes lottery underpredict how quickly the satisfaction fades, and that those who have not experienced paraplegia overpredict how much and how quickly their happiness level would bounce back if they did. Introspection and anecdote might suggest that this list is not exhaustive.

WG contend that impact bias results from two underlying forms of reasoning failure. The first is under-appreciation of the phenomenon of focalism. When thinking about a specific future outcome, people tend to focus on it in isolation from the plethora of other conditions and activities that will surround that specific outcome when it actually transpires, and thus do not take into account in their affective forecast the way that their response to that outcome will be ameliorated or otherwise affected by the other things going on in their lives and environment contemporaneously. This in itself is a form of limited imagination, but one could argue that focalism alone isn’t enough to cause ends failure, because as Wilson et al. (2000) have shown, it is possible to overcome focalism by requiring people to think through what their lives are likely to look like at the time for which they are being asked to provide a forecast. In their study of football fans, they asked one group of subjects to write out a “prospective diary” for each of the days for which they were being asked to predict their post-win happiness, in which they additionally predicted how much time they would spend on each of ten other everyday activities, as well as what they thought they would most likely be doing during each hour of the day. The prospective-diary exercise significantly reduced overprediction of happiness immediately after the game, and completely eliminated overprediction for each of the subsequent three days. In other words, it was possible to overcome focalism. An individual who understood focalism and the way it can be overcome, could presumably impose the rigors of such prospective cogitation on themselves, at least in highly consequential cases. But if an individual is naïve about the effect of focalism on their affective forecasting, they are likely to fail to overcome it in many cases, and thus overpredict the intensity and in all likelihood the duration of their affective response. As a result, they may invest too heavily in the getting and avoiding of particular outcomes. In other words, they may pursue the wrong ends, or the wrong balance of ends. Strikingly, subjects have been found to exhibit this form of misprediction both with respect to outcomes

13 Brickman et al. (1978).
they have never experienced (winning the lottery) and with respect to those that they have experienced countless times (their favorite team winning a game). Thus, this form of reasoning failure cannot be attributed simply to “thinness of experience,” but must involve some more intractable problem, about which individuals apparently have a hard time learning from experience.

It is important to note that there may be cases in which impact bias actually functions to support individuals in the successful pursuit of their goals. If the achievement of a desired end requires overcoming significant obstacles, including perhaps the kinds of issues surrounding lack of willpower addressed by LN, then believing that the end will be more gratifying than it actually is may help motivate individuals to overcome those obstacles. Thus, though they may have been mistaken about the impact of the end, they may not have been mistaken in pursuing it, and the impact bias may have been part of a successful means to achieve the challenging end. In this case we would not want to categorize impact bias as an ends-related reasoning failure.

The second form of reasoning failure that WG believe drives impact bias is failure to predict the set of processes by which we lower the intensity of our affective responses to specific outcomes and put them behind us, which they refer to as “sense-making processes.” As they point out, it would be maladaptive indeed if we allowed the psychological and physiological impact of the never-ending stream of salient outcomes we experience each day to reverberate within our minds and bodies over any extended period of time. The result would be an indecipherable cacophony of thoughts and feelings. Instead of allowing such a cacophony of affective responses to build up in our minds and bodies, we bring each one, positive or negative, down to a manageable level quite quickly, so that we can pay attention to our response to the next salient experience. WG cite research suggesting that we accomplish this by semi-consciously “making sense” of the events and outcomes we experience, incorporating them into our baseline schema of how things are and how they are supposed to be. There may be other factors at work, including things as basic as the parasympathetic nervous system. But here again, what matters in terms of ends failure is not that we have these mechanisms for putting our affective responses behind us quickly, but that we systematically fail to predict the functioning of these mechanisms, or how quickly and effectively they work, despite the fact that we experience the process in action over and over again. As a result, we may persist in pursuing a set of ends that brings us less satisfaction (or at least less durable satisfaction) than we believe they will, while avoiding another set of ends that brings us less dissatisfaction than we believe they will.

3.2.2 Misconstrual

WG’s second main category of ends-related affective forecasting error is misconstrual. According to WG, in the context of affective forecasting,
construal is the process of constructing a mental representation of what a particular event or outcome will entail, and the misconstrual problem is when people “mistakenly imagine the wrong event.” Patrick and MacInnis (2006) define it as “when individuals consider one way in which an outcome might turn out, and fail to consider others.” Patrick and MacInnis present qualitative evidence of misconstrual of various future outcomes, including what a vacation will be like, what other people’s reactions to a fancy new watch will be, and what it would be like to wear a particularly striking garment in various social settings. Woodzicka and LaFrance (2001) found that when women were asked to predict how they would feel in a hypothetical sexual-harassment scenario, they predicted they would feel anger, when in fact subjects who were actually placed in an enactment of the scenario reported feeling fear. This misprediction may have happened because the women in the study had an incorrect mental representation of the external scenario of the interview, or because they failed to foresee what their emotional response to a correctly represented scenario would be. In either case, the result of their misconstrual was that they misforecast their affective response.

There could be any number of similar examples of misconstrual of specific outcomes, some of which could result in potentially catastrophic ends failure. There is the familiar trope of the person who construes the event “having a baby” as including the outcome “resolution of marital strife.” There is the person who construes the event “getting a face tattoo” as consisting primarily of the outcome “getting lots of admiration and positive comments from friends and strangers.” And then there is the endless list of consumer durables legendary for bringing with them outcomes far less rewarding than individuals predict.

There is evidence of misconstrual with respect to life goals. Sheldon et al. (2010) categorized subjects on the basis of their “relative extrinsic value orientation,” or REVO, a measure of how much value an individual places on extrinsic goals such as financial success, attractive appearance, and social recognition relative to the value they place on intrinsic goals such as self-acceptance, community feeling, and affiliation. They then randomly assigned subjects to two groups and asked the subjects in the first group to go out and take some action (of the subjects’ own choosing), over the course of a three-week period, in pursuit of one of the three extrinsic goals, and the subjects in the second group to take an action in pursuit of one of the intrinsic goals. They then asked subjects in each group to predict their subjective wellbeing subsequent to successful achievement of their assigned goals, and then measured their actual subjective wellbeing at the end of the three weeks. They found that subjects with relatively high REVO predicted higher wellbeing subsequent to achievement of extrinsic goals than those with relatively low REVO, and were wrong, in the sense that their predictions were not borne out by their actual post-achievement wellbeing. The setup of the experiment made it explicit that the ends about which subjects were being asked to make affective forecasts
were not the ends of completing the actions they had chosen, but the higher extrinsic or intrinsic ends themselves, and that the outcome they were being asked to make an affective forecast of was not the completion of the task, but the achievement of the goal.

It is possible that people may misconstrue certain spiritual or self-improvement practices, and the outcomes they lead to, and thus fail to foresee how much satisfaction and happiness those practices, and the life experience they lead to, could bring. A considerable body of experimental evidence has found that simple practices such as writing letters that express gratitude, counting one's blessings, and performing kind acts, can improve subjective well-being at very low cost to the individual. These are all practices promoted by various religious denominations and self-help movements. A particularly thorough study by Fredrickson et al. (2008) randomized participants in a corporate wellness program into either a seven-week loving-kindness meditation class or a waitlist condition (in which they waited until after the end of the study period to enter the meditation class). Subjects in the meditation program experienced an increase in positive emotions, leading to increases in a remarkable array of what the authors refer to as “personal resources,” such as ego-resilience, positive relations with others, and sleep duration, which in turn led to increases in self-reported happiness and decreases in depressive symptoms. They did not ask their subjects to make affective forecasts, but one can imagine a person construing a period each day of sitting still with eyes closed, doggedly repeating a mantra of loving-kindness towards others as boring or even aversive, without at all considering that it might in fact be an enjoyable practice leading to a significant uptick in many of the basic ingredients of a happy life at virtually no cost. It is plausible that such a person, upon being induced by some paternalistic agent to take an introductory meditation course, might discover that meditation, and the states of mind and being that it induces, was an end far more worthy of their time and effort than some of the ends they had previously been pursuing. Given the extensive literature on the positive effects of this and other forms of meditation, such a person might discover that they had previously had entirely the wrong ends, as judged by themselves.

4 Justification of Ends Paternalism

LN’s framework for analyzing whether any given paternalistic policy is potentially justifiable rests on two basic pillars, first that government paternalism can only be justified if the intention is to help individuals to achieve their own self-identified ends, rather than those of the paternalist, and second that

---

the paternalist’s claim that individuals need the help of government policy in achieving those ends, and that the specific policy they are analyzing will provide that help, must be supported by an accumulation of rigorous social-scientific evidence. They claim that these two criteria definitively reject all forms of ends paternalism. According to LN, replacing a person’s judgment about their ends can only move them further away from getting what they themselves want, and doing so purely to satisfy the paternalist’s ends almost certainly does not increase social welfare as commonly understood by economists, and is thus an unambiguously inexcusable interference with autonomy. And, they claim, there is no accumulation of evidence that people make judgment errors regarding their ends, so that claims of such errors can only be based on opinion and speculation, and are thus unlikely to be right in any predictable or trustworthy way.

I have now shown that excluding ends paternalism on the basis of these pillars is unfounded. There is, in fact, a large and growing body of evidence that people pursue ends that they themselves, under plausible conditions, could reasonably be expected to identify as suboptimal, which is to say, for the purposes of maximizing wellbeing, the wrong ends, as defined by themselves. Thus LN’s framework for determining whether government paternalism can be considered potentially justifiable is directly applicable to ends paternalism, in the case of policies intended to help people pursue and achieve ends that are superior, as (eventually) judged by themselves, to those they are currently pursuing. Under the LN framework, thus extended to ends paternalism, the relevant questions are these: 1) is there a reasoning failure that may cause people to fail to correctly identify the ends that they themselves would, if they could, identify as most effective in enhancing their ultimate happiness and satisfaction, and which this paternalistic policy can reliably be expected to help them overcome; 2) is there an accumulation of social-scientific research in support of these claims of reasoning failure; and 3) is the harm caused by the inevitable interference with individuals’ autonomy outweighed by the benefit in terms of well-being and satisfaction, taking into consideration any unintended harm that might be caused to any members of society? In other words, the analysis under the LN framework is exactly the same in the case of ends paternalism as in the case of means paternalism.

What kinds of ends paternalism, then, might we apply the extended framework to? An obvious category, motivated by the phenomenon of misconstrual, would be to subsidize or otherwise induce people to explore experiences and

15A supplementary question that paternalists would be wise to add to the analysis would be, “is there an alternative paternalistic policy that would achieve qualitatively the same welfare gains while interfering less with autonomy?” It is on this question, as LN explain, that libertarian paternalists have something really important to contribute, not, as libertarian paternalists themselves sometimes claim, the question of how to achieve blanket justification of a whole category of interventions on the basis of some comprehensive claim about preserving freedom of choice.
activities that social science suggests may increase their happiness and satisfaction more than they realize. Subsidizing the arts might be included in this category, or even requiring exposure to the arts in some way. But there may also be important examples in the domain of positive psychology, as suggested by the studies cited above on meditation, gratitude practice, and acting kindly. Tax incentives for firms that proactively encourage or even require their employees to participate in classes on these kinds of practices could cause people to revise their self-identified ends and thus achieve greater life satisfaction on their own recognizance after participating in the classes.

Another category of government paternalism, which comes to mind as a response to affective misforecasting caused by focalism as well as misconstrual, would consist of policies intended to eliminate or counteract the influence on affective forecasting of advertising and marketing. Advertisers make extensive use of representations of future outcomes that arguably, (a) do not adequately take into account the matrix of other activities and conditions within which the advertised commodity or activity will actually unfold, and (b) involve conditions and experiences that are quite different from how many consumers would actually experience the advertised commodity or activity. I am not aware of any rigorous evidence of the effect of advertising on affective forecasting, but anecdotal evidence is easy to come by. Laws restricting certain kinds of advertising, particularly for certain kinds of goods and activities highly prone to misconstrual, and public information campaigns designed to overcome misconstrual with respect to particular products or activities come to mind. Graphic warning labels on cigarettes are an example, but the range of potentially justifiable cases may be much larger.

One could also imagine policies requiring people to complete anti-focalism tasks such as the prospective-diary exercise mentioned above before making high-stakes decisions. For example, behavioral economists have argued that

16 Though the number of specific activities or consumption goods with which some people don’t have enough experience to appreciate could be very large, and we would want to be confident that we were dealing with cases in which we had clear evidence that large numbers of people were making large mistakes. But that is precisely the requirement of the LN framework: strong social-science evidence of unambiguous mistakes that cause significant welfare loss. The ends-paternalist, like any other paternalist, and indeed any other policy maker, ought to be very confident that their policy is unambiguously and significantly worthwhile, which is likely to require a high degree of certainty that the activities and goods in question are ones that people misconstrue, and that their misconstrual matters a lot.

17 We might be tempted to frame cases such as these in terms of Dewey’s theory of the impossibility of distinguishing between means and ends, which he argued was due in part to the fact that as we engage in our chosen means to achieve a pre-determined end, our evaluation of the desirability of the end may change, so that means essentially become part and parcel of ends. (Dewey, 1939) Instead, the cases I am considering here are ones in which it is not the means by which an end is pursued, or the process of implementing those means, that change the individual’s evaluation of the end, but simply the discovery of what the end actually consists of, and of the satisfaction, happiness, or well-being that it brings, regardless of what means are used to achieve the end.
in the heyday of the subprime-mortgage bubble, many borrowers exhibited means-related reasoning failure, including failure to adequately take the long term impacts of their choice into account, or to correctly assess their own risk of default, leading to outcomes far from their most preferred ends.\textsuperscript{18} But in addition it is possible that some of them simply misconstrued what homeownership would be like, or what losing their home and credit rating would be like, with the result that they placed too much value on the end of homeownership, and too little on the end of avoiding negative financial outcomes. How many of them might have been spared the results of this ends-related reasoning failure if they had been required by law to complete some process of clearly envisioning the various possible outcomes of the choice they were making? A similar requirement might spare people from unfortunate outcomes in the case of family planning, educational attainment, or career choice, though it might be harder to target the precise decision-making moment at which the requirement would need to be imposed.

5 Conclusion

I have shown that ends paternalism can be justified in the LN framework on precisely the same grounds as means paternalism. To make use of the extended framework one must ask, first, is there a strong reason to believe, on the basis of an accumulation of well-conducted social-science research, that there is some reasoning failure that can be shown to cause people to fall short of achieving those ends that they would identify as bringing them the greatest wellbeing and satisfaction, according to their own judgment, if they were not subject to said reasoning failure? Second, is there solid evidence that the policy currently under consideration will successfully address the reasoning failure? Third, if so, does the benefit in terms of increased wellbeing from correcting said reasoning failure outweigh any unintended harm the policy may cause, and the harm done by the inevitable interference with people’s autonomy? The essence of my argument for this extension of the LN framework to include ends paternalism can be found in a statement made by LN themselves. With respect to what I have called ends-related reasoning failure, they write,

\begin{quote}
“This kind of failure would involve not acting in our best interests because, in the judgment of someone other than ourselves, either [1] we have not properly identified our best interests or [2] our identified best interests are simply wrong.”\textsuperscript{19}
\end{quote}

\textsuperscript{18}See, for example, Bar-Gill (2009).
\textsuperscript{19}Le Grand and New (2015, p. 104). Note that throughout this paper I have been using the term “wrong ends” to refer to the case in which the individual has not yet identified their optimal ends, rather than in the way LN are using it here, which is to refer to the
LN abjure this entire notion, but in fact with this statement they have precisely distinguished justifiable from non-justifiable ends paternalism. Cases abound that fall into category [1] in LN’s statement above. As I have shown, and as psychologists in the field of affective forecasting have demonstrated, there are numerous important ways in which people should indeed be thought of as having not yet properly identified their best interest. Thus, we should consider as potentially justifiable any paternalistic policy that would have the effect of either revealing to an individual an end they didn’t know they would value, or cultivating an end they did not previously value but have now come to. But additionally we should consider potentially justifiable policies that protect people from the kind of affective forecasting failure that may not be resolved through experience, such as impact bias, whereby people may persist in believing that their positive or negative affective response to some future outcome will be more intense, and/or last longer, than it actually will, even after repeated exposure to similar outcomes. Meanwhile, category [2] in LN’s statement, above, in which the paternalist simply thinks they know better, remain unjustified.

Finally, the question of how to determine whether an individual’s currently self-identified ends are suboptimal is no more problematic, conceptually, than the question of whether an individual’s means are suboptimal. First, there must be some way of at least hypothetically eliciting from the targeted individual that at the end of the day they had, prior to the policy, been pursuing a set of ends that they now see as suboptimal. Second, in any specific situation in which we are considering a paternalistic intervention, we should require strong social scientific evidence that this is actually the case. In general this seems likely to be a challenging evidentiary standard to achieve, though the expanding literature on affective forecasting could be seen as a good beginning.

It is important to note that the LN framework, even as extended to include ends paternalism, is not a green light for what critics call “the new paternalism.” Libertarians such as Glenn Whitman (2010) worry that “the new paternalism carries a serious risk of expansion.” In other words, it could carry us down a slippery slope. This may or may not be true, but to the extent that we should be concerned about lax use of psychology and behavioral economics to justify intrusive paternalism for which policy makers or policy advocates have a pre-existing, non-evidence-based predilection, LN’s framework, even as extended here, should be seen as a brake pedal rather than an accelerator. Their conditions, if applied as intended, are stringent. The analysis they call for as a bare minimum for justification of paternalism is extensive and exacting. If their framework is applied with fidelity by policy analysts, and the resulting recommendations are taken seriously by policy case in which the paternalist wishes to impose a set of ends upon the individual which the individual would never identify as optimal for themselves.

20Whitman (2010).
makers, the result could plausibly be a reduction rather than an expansion of paternalism. In particular, as the proponents of asymmetric and libertarian paternalism repeatedly emphasize, if policy makers were to think harder about the actual causes of reasoning failure in specific situations, and design their paternalistic interventions to more effectively overcome those failures, the resulting interventions might interfere with autonomy less, holding effectiveness constant, than more heavy-handed alternatives.

To be absolutely clear about this point, let me reiterate all that the LN framework demands of the paternalist. First, regardless of whether we are analyzing means or ends paternalism, the claim that individuals are making judgement errors that cause them welfare loss must be solidly grounded in rigorous social science research. Obviously, one could take a rather liberal interpretation of words like solid and rigorous, but it is apparent that LN’s intention is for the standard to be high. Next, there must be good reason to believe that the policy under consideration will cause a substantial net welfare benefit to society as a whole. The paternalist must be confident that the aggregate benefit to those who are the target of the policy must outweigh the aggregate of any loss the policy may impose on any of them, or on others, who may not have been the target of the policy, but who may nonetheless be affected by it. This is no different from the benefit-cost test that is frequently imposed upon all types of policy. Finally, the net benefit in terms of welfare must be great enough to offset the negative impact of the inevitable infringement upon autonomy that paternalism causes. Obviously that is not an objective assessment, and because some of the people affected by any given paternalistic policy may place a very high value on autonomy, it should be viewed as a high bar, and there should be clear evidence of the value the relevant individuals place on autonomy, and the degree to which they experience infringement upon it. Indeed, it seems reasonable to conjecture that the degree of perceived infringement upon autonomy caused by a paternalistic policy is greater when the policy addresses ends than when it addresses means, and if so, we probably ought to hold ends paternalism to an even higher standard with respect to all of these criteria than we hold means paternalism.

In addition, ends-paternalistic policies such as the examples I mention in section four, even when found to be justified by the LN framework, should be subjected to the same set of supplementary considerations that LN consider at length in the case of means paternalism. In particular, ceteris paribus, we should prefer paternalism that expands rather than restricts choices. But as with means paternalism, restriction of choice or coercive manipulation of choice should not be ruled out when the benefits of the paternalistic policy under consideration are expected to be large relative to less coercive or manipulative alternatives.\footnote{Sunstein himself makes this point quite explicitly. See Sunstein (2014, p. 142).} Also, as with means paternalism, restrictions and even outright
bans on the activities of firms and other parties who intentionally or unintentionally take advantage of other people’s ends-related reasoning failure may be considered more justified, *ceteris paribus*, than those that restrict the choices of those making the reasoning failure themselves. Finally, as with means paternalism, we should require the paternalist to consider a range of policies that could achieve the same welfare outcomes with less infringement upon autonomy.\(^{22}\) Along those lines, a clear understanding of choice architecture and “nudging” may help to identify opportunities to harness reasoning failures in the service of correcting ends failure. Given all of these considerations, the replete and excellent toolkit that LN have provided for the self-aware and conscientious government paternalist (or anti-paternalist) is applicable to the analysis and design of ends paternalism just as much, and in precisely the same ways, as to means paternalism.

### References


---

\(^{22}\) And to convince us that there are no adequate private-market alternatives.


