RICE UNIVERSITY

Choice, manipulation and wellbeing: On the nature and ethical significances of nudging

By

Kerry Vaughan

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

George Sher
Herbert S. Autrey Professor of Philosophy
Chair

Baruch Brody
Andrew W. Mellon Professor in Humanities and
Professor of Philosophy

Rick K. Wilson
Professor of Political Science

Jennifer Blumenthal-Barby
Cullen Associate Professor of Medical Ethics
Baylor College of Medicine

HOUSTON, TX
MAY 2017

ABSTRACT

Recent work in behavioral economics has led to startling conclusions about the limits of human rationality. Contrary to the rational maximizer of utility assumed by traditional economics, actual decision makers make choices that are inconsistent with their own ends and are powerfully influenced by the context in which decisions are presented. Recently, some writers have argued that we ought to use the power of decision making context to offset the inconsistent choice phenomenon. Positions of this kind go alternatively under the banners of "Libertarian Paternalism," "Choice Architecture," and "Nudging." The central idea is that people who shape the context of choices (Choice Architects) should opt to frame choices so all choices remain available (Libertarianism), but should ensure that the choosers are more likely (Nudged) to make choices that make them better off (Paternalism).

Despite an explosion in discussion of and use of nudges, philosophers and ethicists have in large part been missing from the conversation. The discussion that has taken place among philosophers has mostly been about whether there is something objectionable about nudges *in general*. However, as I will argue later in the dissertation, this discussion is of limited use because nudges vary widely in their ethical features.

This dissertation advances in five chapters. In chapter 1, I discuss what precisely a nudge is and what it is not. In chapter two I outline a three-factor model for analyzing whether a nudge is morally acceptable. In chapter three I discuss the question of when a nudge makes a chooser better off. I finally defend a new version of the informed desire account which avoids difficulties with the standard informed desire account in the literature. In chapter four I discuss the question of when a nudge is the best available choice, comparing it to rational persuasion, libertarianism

and paternalism as possible alternatives. Finally in chapter 5 I discuss two real world applications of the nudging and how the ideas developed elsewhere in the dissertation are used to evaluate these nudges.

ACKNOWLEDGMENTS

Although my name appears on the cover of this dissertation, it would not have been possible were it not for the support and encouragement of many people.

I'd like to thank those who first kindled my interest in philosophy. This includes Andrew Kania and Steven Luper who mentored and taught me while an undergrad at Trinity University. Without their love for the subject and facilitation of lively debates, I likely would have done something else entirely. I'd also like to thank Melanie Brashear Lawrence who helped me develop a love for argumentation and the world of ideas through debate.

I'd like to thank those who contributed directly to the work of this dissertation. This includes Jennifer Blumenthal-Barby who first suggested the topic to me and who worked most closely with me in developing my ideas on the topic. Without her help I likely would have had neither the willpower nor the intellectual content to see this project through to completion. I'd like to thank Baruch Brody and George Sher for their consistently insightful feedback on this dissertation at each stage of its development. I'd also like to thank Rick Wilson for his participation on the dissertation committee.

Finally, I'd like to thank my family for their support and guidance. I'd like to thank my mom and dad for their continual love and support. I have always charted my own course through life, and they have been there to support me even when it was not clear to them where my path may lead. This requires a special kind of faith that I have been very lucky to have. I'd also like to thank my loving wife, Lauryn. I have never met a more generous, caring or supportive person as her. This dissertation would not be possible without her in my life. May our next dozen years together be as adventurous as the first.

Chapter 1: What is a nudge?

1.1 Introduction

This is a dissertation about choice and wellbeing. Specifically, it deals with nudges, which are an attempt to apply the lessons of behavioral economics to improve the wellbeing of choosers without restricting or burdening their choices. Nudges are applied by so-called choice architects, who are people responsible for designing the context in which choices are made. Nudges promise to "save money, improve people's health and to lengthen lives.[1]" The nudge approach has been used by policymakers to design the 401(k) pension scheme,[2] suggest a tax refund system,[3] improve compliance in tax reporting[4] and lower youth alcohol consumption[5] among many other things. The seemingly endless promise of nudges has led to an explosion of popularity both as an academic topic and as a tool in public policy. In fact, Cass Sunstein, one of the authors of the seminal work on the subject, Nudge, served as the head of the White House Office of Information and Regulatory Affairs in the Obama administration during which he worked with President Obama to use nudges in the federal government. On the other hand,

---

[1] Cass Sunstein, *Simpler: The Future of Government*, (Simon & Schuster, 2013), 2.
[2] Edmund L. Andrews, "Obama Outlines Retirement Initiatives," *New York Times*, 2009, quoted in Pelle Guldborg Hansen & Andreas Maaløe Jespersen, "Nudge and the Manipulation of Choice: A Framework for the Responsible Use of the Nudge Approach to Behavioral Change," *European Journal of Risk Regulation* (2013).
[3] James Surowiecki, "A Smarter Stimulus," *New Yorker*, 2009, quoted in Hansen and Jespersen, "Nudge."
[4] Cabinet Office and Behavioral Insights Team, "Applying Behavioral Insights to Reduce Fraud, Error and Debt," 2011, quoted in Hansen and Jespersen, "Nudge."
[5] Cabinet Office and Behavioral Insights Team, "Applying Behavioral Insights to Health," 2011, quoted in Hansen and Jespersen, "Nudge."

nudges are also attacked for "manipulating people's choices,[6]" of violating liberal principles,[7] and of being a "threat to choice, freedom and democracy.[8]"

Despite the explosion of interest in nudges, philosophers and ethicists have been less prominent in the conversation. This is puzzling given that nudges touch on many important questions in ethics. They sometimes involve manipulation, always involve attempts to influence behavior through means other than rational persuasion and touch on a myriad of other important ethical issues such as libertarianism, paternalism, accounts of wellbeing, informed consent, public health ethics and many others. The discussion that has taken place among philosophers has mostly been about whether there is something objectionable about nudges *in general*. However, as I will argue later in the dissertation, this discussion is of limited use because nudges vary widely in their ethically important features. That is, nudges can involve different behavior-altering mechanisms, can involve different choosers and choice architects, can aim at different ends and so on. It is unlikely that much can be said about the ethics of nudges *in general*.

The goal of this dissertation is ultimately practical. I will define the circumstances under which a nudge is acceptable and morally preferable, and the circumstances under which it is not acceptable or morally preferable. But it also aims to provide guidance to real life Choice Architects who want to use nudges to alter people's behavior in ethically responsible ways.

---

[6] Luc Bovens, "The Ethics of Nudge," in *Preference Change: Approaches from Philosophy, Economics and Psychology*, eds. Till Grüne-Yanoff and Sven Ove Hansson, vol. 42, (Berlin and New York: Springer, Theory and Decision Library A, 2008), quoted in Hansen and Jespersen, "Nudge."

[7] Till Grüne –Yanoff, "Old Wine in New Casks: Libertarian Paternalism Still Violates Liberal Principles," in *Social Choice & Welfare*, eds. J. Duggan, B. Dutta, M. Fleurbaey and C. Puppe, vol. 38, (Springer, 2012).

[8] Brendan O'Neill, "A Message to the Illiberal Nudge Industry: Push Off," *Spiked*, 2010.

1.2. What is a Nudge?

To understand the ethics of nudging, we first need to be clear on what a nudge is. Some heuristics have been used to explain the idea and a few possible definitions have been suggested. In this section, I briefly review the history that led to the nudge idea, discuss and reject some proposed definitions. I then discuss a recent definition of the term put forward by Yashar Saghai and accept it with some minor revisions.

Nudging is perhaps the unavoidable result of two lines of observation that have occurred during the past four decades of research in the social sciences, especially in the field of behavioral economics. The first observation is that decision makers make decisions that are inconsistent with their own ends or that are not in their best interest. For example, a common investment vehicle is an index fund, which is designed to replicate the movements of an index for a specific financial market. Because the goal of the index fund manager is merely to replicate the index, one should find no differences in average return between different funds that track the same index. Despite this fact, people are willing to pay higher fees to well-advertised index funds instead of finding the index fund with the lowest fee structure.[9] If we assume that the goal of the investor is to make money, the willingness to pay higher fees for the same product is inconsistent with the investor's goals. However, we need not go into anything as esoteric as personal investing to see this: people smoke, drink, overeat, fail to save enough, take out loans they cannot afford and so on. Everyone could use some help in making better choices. Following Blumenthal-Barby,[10] let us call this the "Bad Choice Phenomenon."

---

[9] James Kwak, "Improving Retirement Savings Options for Employees," *John M. Olin Center for Law, Economics and Business Fellows' Discussion Paper Series* (2014).

[10] J. S. Blumenthal-Barby, "Choice Architecture: A Mechanism for Improving Decisions While Preserving Liberty?" in *Paternalism: Theory and Practice*, eds. C. Coons and M. Weber (Cambridge University Press, 2013).

The second observation is that choosers are powerfully influenced by their environments in seemingly illogical ways. Examples include messenger effects, the power of social norms, the use of default options, what information is salient, how information is framed and ordered and so on.[11] One classic example is in setting the default (i.e., what will happen if the chooser does nothing) in 401(k) plans. Research shows that when you automatically enroll employees into a retirement plan the participation rate goes up dramatically whereas if employees begin by being opted out of the 401(k) plan and they must complete paperwork to opt in, the participation rate goes down dramatically. If we assume that the decision to participate in a 401(k) is important, it is hard to explain rationally why asking employees to fill out some paperwork ought to have such a large influence on their decisions. Let us call these subtle choice context effects the "Influence Phenomenon."

Nudges arise from the idea that we ought to use the Influence Phenomenon to counteract the Bad Choice phenomenon. That is, we ought to change and shape the context in which people make choices to help make them better off without restricting their available options. In what follows I discuss several possible interpretations of what it means to nudge. The goal of this section is to arrive at a definition that captures the use of the concept in the literature as closely and consistently as possible. A successful definition should include all the paradigmatic cases of nudges and should exclude most of the tools of behavioral change that are not usually thought to be nudges including incentives, disincentives, coercion and so on.

[11] Paul Dolan, Michael Hallsworth, David Halpern, Dominic King, and Ivo Vlaev, "Mindspace: Influencing Behavior through Public Policy," *Institute for Government* (2010); Sunstein and Thaler, "Nudge: Improving Decisions"; Daniel Kahneman and Amos Tversky, "Choices, Values, and Frames," *Cambridge University Press* (2000).

1.2.1. Nudges as factors that influence 'Humans' but not 'Econs'

In their book Nudge, Richard Thaler and Cass Sunstein (from now on, T&S) offer a useful heuristic that helps to explain the concept of a nudge. They make a distinction between Humans and Econs. An Econ is a rational maximizer of utility assumed by traditional economics. Econs have perfect information, perfect calculation abilities and unlimited self-control. Econs always make decisions that are in their best interest. Humans of course do not always do this. They act on partial information, fail to evaluate situations properly and sometimes lack self-control. A nudge is any influence attempt that might change the behavior of a human, but would not change the behavior of an Econ.

To see how this heuristic works, let us return to the opt-out 401(k) enrollment case discussed above. As noted, research has consistently shown that automatically including employees in the 401(k) plan creates large increases in the rate of retirement savings. On the humans and Econs heuristic, such a scheme is an example of a nudge because an Econ's retirement decisions would not be influenced by the presence of such a simple form to fill out. An Econ always calculates what is in her best interest and acts accordingly, and presumably, the presence of such a form would not factor into the calculations. The heuristic is also useful for distinguishing between nudges and incentives. One reason employees enroll in the 401(k) plan is to get employer matching funds. If the employer increased the match, employees might choose to save more. However, this is not a nudge because, presumably, this *would* influence the choice of the Econ.

While the human and Econs distinction provides a useful way of thinking about nudges, unfortunately, it is not helpful with many of the difficult cases. This is because it relies heavily

on what the Econ would care about, which itself relies on an undefended theory of rational choice. For example, in 401(k) enrollment, it seems clear that an Econ would not care about the paperwork needed to opt in or out of the plan. However, if instead we required paperwork to buy candy from a vending machine, this may or may not matter to the Econ. Therefore, to distinguish nudges from other means of altering behavior, we will need a more precise definition.

1.2.2. Nudges as Factors that Influence 'System 1' but not 'System 2'

A second heuristic used by Sunstein in his most recent book Simpler: The Future of Government is to borrow from Kahneman's work in his book, Thinking, Fast and Slow and distinguish between two cognitive systems, called System 1 and System 2. System 1 is immediate, fast, emotional, intuitive, and automatic whereas System 2 is calculating, purposeful, considered and logical. The idea, according to Sunstein is that "System 1 is much moved by features of choice architecture that would seem irrelevant to System 2." Borrowing from our 401(k) example again, System 2 would probably not care whether the 401(k) plan is opt-in or opt-out because System 2 would calculate that neither alternative should change the decision. However, System 1 does not do such calculating, it tends to go with the flow and is heavily influenced by the option it perceives to be the most common. Accordingly, so the heuristic goes, because opt-in policies impact System 1 but not System 2, they are nudges.

This heuristic suffers from similar deficiencies as the humans and Econs heuristic. It is not clear what things are of interest to System 2 and what things are of interest to System 1. For example, choosing the option that seems most commonly chosen by others is usually a tactic taken by System 1, but not System 2. However, in a situation where the chooser does not have

complete information or is unable to do the necessary calculating, System 2 may reasonably choose the option that appears most common on the assumption that others have done the necessary calculating. Additionally, this heuristic leans heavily on distinguishing precisely what kinds of behaviors and processes are involved in System 1 versus System 2. But, not all processes can be so easily delineated. For example, a patient may listen to advice that is presented by a doctor but ignore similar advice given by a stranger. This could be because System 1 trusts authority, but it could also be because System 2 calculates that the doctor is likely to have good information whereas the stranger is not. Therefore, for philosophical discourse, these heuristics will not be sufficient.

1.2.3. Nudges as Influences that Result in Predictable Behavior but Preserve Open Choice Sets (Thaler and Sunstein's Definition)

For their part, T&S provide the following definition of nudges:

> Nudge: "… any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives.[12]"

Here, the "choice architecture" is understood to be the context surrounding a decision. It includes anything that is impactful in making a particular decision. For example, say you need to decide if you would like to undertake a particular medical procedure, so you discuss this with your doctor. Many things about that interaction may constitute the choice architecture. The fact that the *doctor* presents the information, that the information is presented at a doctor's office,

[12] Sunstein and Thaler, "Nudge: Improving Decisions," 6.

how that information is framed (e.g., as losses vs. gains), the ordering of the available options, even the dress of the doctor all constitutes the choice architecture among many other factors.

There are a few problems with the definition proposed by T&S. For example, the definition only excludes two kinds of cases: "forbidding option" and "changing economic incentives;" however, there are many kinds of cases that are not nudges, but that do not fall into either of these cases. First, there are many cases where one can alter the behavior of a chooser in a way that is not a nudge by changing their nonmonetary incentives. For example, imagine if we made it illegal to not enter your company's 401(k) plan, punishable by jail time. This would not directly affect the chooser's economic incentives, but it is clearly not a nudge. One might argue that such a case is excluded because it "forbids" options. However we might instead imagine a case where one is punched if they fail to participate in the company 401(k) plan. This would alter the behavior of choosers in a predictable way, does not forbid options or change economic incentives, but is also clearly not a nudge. Therefore, a better definition needs to exclude nonmonetary means of changing the incentives of choosers. Another problem is that the term "forbidding options" is too narrow. As a result, it does not exclude certain kinds of cases that are not nudges. To better match up with our intuitions, we might want to exclude certain means of burdening options that do not forbid them. Returning to the candy bar example, if we required an hour of complex paperwork and waivers to buy a candy bar, this would not affect the chooser's economic incentives and would not forbid the option, but also seems to be at odds with the spirit of the nudge concept.

1.2.4. Nudges as Influences that Preserve Both Open Choice Sets and Low Cost Alternatives

An alternative definition proposed by Hausman and Welch is as follows:

> "Nudges are ways of influencing choice without limiting the choice set or making alternatives appreciably more costly in terms of time, trouble, social sanctions and so forth.[13] [14] [15]"

This definition solves some of the problems of the original definition proposed by T&S, but it suffers from some additional problems. The most obvious is that it does not specify what means of influencing choice are considered nudges. Consider rational persuasion as an example. If I try to get an employee to save for retirement by convincing him that doing so is in his best interest through argument, this would count as a nudge under the letter of Hausman and Welch's definition because it would influence choice without making options more costly. However, rational persuasion is usually not thought to be a nudge. A second problem is that this definition does not say anything about the resistibility of such nudges. For example, imagine I try to encourage employee 401(k) savings by brainwashing them to have a strong preference for saving money. This follows the letter of the Hausman and Welch definition, but violates the spirit of the liberty-preserving nature of nudges.

A final problem with this definition is that it leaves it ambiguous as to whether there can be unintentional or accidental nudges. Imagine, for example, that I force employees to choose

---

[13] Hausman and Welch, "Debate," 123-136.

[14] Hansen and Jespersen suggest that "choice" is the wrong word here because nudges often do not operate on choice as the term is understood in the psychological literature. They suggest that we discuss influences in behavior instead. I am open to this question, but use choice here for citation consistency.

[15] To be maximally charitable to Hausman and Welch, they foresee some of these arguments and include that nudges are "called for because of flaws in individual decision-making, and work by making use of those flaws." However, I think this aspect of nudges should be built into the definition itself.

between opting in or opting out of the 401(k) plan so I distribute a form where the employee must indicate if they wish to be in the 401(k) plan or not. However, it just so happens that I place the opt-in option first and my employees are far more likely to choose the first option they see from a list of options. Does this count as a nudge? Opinions on this question differ. On the definition proposed by T&S, this counts, as they want to include anything that alters a behavior as a nudge. However, others like Hansen and Jespersen[16] and Sahai[17] prefer an understanding of the concept that requires it to be an intentional act.

I think the distinction I made earlier between a nudge and the choice architecture makes it clear that nudges should be understood as intentional acts. Recall that the choice architecture consists of all the circumstances surrounding a decision. They can include obvious factors like how a decision is framed, but they can also include subtle features like the color of the room the decision is made in or the physical appearance of the person framing the choice. The choice architecture surrounding a decision is sometimes intentional, but it is sometimes unintentional. I think the easiest way to explain the distinction between the choice architecture and a nudge is with reference to intentionality. A nudge is [in part] an intentional attempt to alter the choice architecture. Therefore, I think nudges are intentional.

1.2.5. Nudges as Influences that Trigger Shallow Cognitive Processes, Preserve Choice Sets and are Substantially Non-Controlling

With these issues in mind, I turn now to a definition of nudge put forth by Yashar Saghai that I think (with a few changes) solves these problems.

---

[16] Hansen and Jespersen, "Nudge."
[17] Yashar Saghai, "Salvaging the Concept of Nudge," *J Med Ethics* (2013).

> Nudge: A nudges B when A makes it more likely that B will φ, primarily triggered by B's shallow cognitive processes, while A's influence preserves B's choice-set and is substantially noncontrolling (ie, preserves B's freedom of choice).[18]

Several components of this definition require further explanation. The first is the term "shallow cognitive processes." According to Saghai, shallow cognitive processes are: 1) fast; 2) consume few cognitive resources and; 3) yield responses that are not the result of full-blown deliberation.[19] They are related to the System 1 thinking discussed by T&S, but Saghai does not require such processing to be automatic in the way that System 1 thinking is sometimes described. One of the benefits of framing nudges in terms of "shallow cognitive processing" is that it excludes certain kinds of cases, like rational persuasion, that preserve choice sets and alter behavior, but are not nudges, but solving the issue with Hausman and Welch's definition.

The next important term is "substantially noncontrolling." Here, Saghai intends the following:

> Substantially Noncontrolling: A's influence to get B to φ is substantially noncontrolling when B could easily not φ if she did not want to φ.[20]

The idea is that influences can be placed on a continuum that ranges from fully controlling (compulsion, coercion, choice elimination etc.) to fully noncontrolling (rational persuasion). Nudges are substantially noncontrolling, meaning that they are more controlling than rational persuasion, but less controlling than coercion and are relatively easy to resist. The key idea here is the idea of easy resistibility. That is, nudges can be easily resisted whereas

[18] Saghai, "Salvaging," 5.
[19] Saghai, "Salvaging," 3.
[20] Ibid., 2.

coercion cannot. Saghai identifies three criteria for easy resistibility of nudges. Imagine that A is trying to nudge B to φ. The nudge is easily resisted if:

1. B has the capacity to become aware of A's pressure to get her to φ (attention-bringing capacities); and
2. B has the capacity to inhibit her triggered propensity to φ (inhibitory capacities); and
3. B is not subject to an influence, or put in circumstances that would significantly undermine the relatively effortless exercise of attention-bringing and inhibitory capacities.[21]

If influences are not easily resistible, then they are not substantially noncontrolling and therefore, are not nudges according to Saghai. Saghai also introduces a new term for influences that are like nudges, but are substantially *controlling* instead of substantially noncontrolling – a behavioral prod. Behavioral prods preserve the choosers choice set (like nudges), but do not make it easy for the chooser to resist the choice architect's influence (unlike nudges).

An example of an influence attempt that is a behavioral prod but not a behavioral nudge is given in the following (real life) example:

> Asparagus-Lovers. An investigator suggests to research participants that they liked or loved asparagus during childhood the first time they tried it, creating a false memory, and a false belief about the taste of asparagus. Subsequently, participants reported increased general liking of asparagus, greater desire to eat it, and willingness to pay more for it.[22]

Asparagus-Lovers is not fully controlling because research participants could not eat asparagus if they did not want to. That is, there is some degree of resistibility. However, it is also not a nudge

[21] Ibid., 3.
[22] Ibid., 1.

on Saghai's definition because the degree of resistibility is too low. Asparagus-Lovers involves a large degree of deception which makes it very hard for the participants to become aware of the researchers influence attempts and makes it difficult for participants to inhibit their triggered propensity to like asparagus.

Saghai's definition makes a very important contribution to the literature base. However, in the next section I offer an alternative definition that I think better captures the common parlance of the term than Saghai's definition.

1.2.6. Nudges as Intentional Influences that Trigger Shallow Cognitive Processes, and Preserve Choice Sets

I offer an alternative definition that includes two qualifications. First, I think nudges are, at their core, intentional acts. That is, I do not think one can nudge another unintentionally. However, Saghair's definition technically leaves the possibility of unintentional nudges open. For example, imagine I just happen to place the potato chips in the back of the pantry instead of in the front and as a result it becomes more likely that my wife will eat healthy. On Saghai's definition this is a nudge. Perhaps not much turns on this because from an ethical standpoint the relevant cases are those where the choice architect's behavior is intentional and therefore morally evaluable, but I prefer to remove any possible ambiguity from the situation.

Second, while I think the differentiation of substantial control from substantial noncontrol is a very important contribution to the literature, I do not think it is a *definitional* contribution to the literature. Definitions should ideally match the common parlance for the term, and it seems to me that the common usage of the term includes cases like Asparagus-Lovers as examples of

nudges. So, while I think it is important to distinguish between substantially controlling and substantially noncontrolling nudges for purposes of assessing their ethical status (as I will do in Chapter 2), I do not want to exclude substantially controlling nudges from the concept

For his part, Saghai seems to foresee and welcome this move. As a footnote to this definition he says the following: "of course, the reader does not have to adopt the technical terminology I suggest. If some prefer to use the term 'nudge' to refer very broadly to influences activating shallow cognitive processes, they could distinguish controlling from noncontrolling nudges."[23]

With these qualifications in mind I prefer the following definition of a nudge:

Nudge: A nudges B when A intentionally makes it more likely that B will φ, primarily triggered by B's shallow cognitive processes, while A's influence preserves B's choice set.

This definition seems to include only the kinds of cases that are usually discussed as examples of nudges. It excludes paradigmatic cases of rational persuasion because such cases usually do not involve shallow cognitive processes. It excludes paradigmatic cases of coercion, and compulsion because such cases do not preserve choice sets. It excludes incentives and disincentives because such cases are usually not primarily triggered by shallow cognitive processes.[24] When I refer to nudges in this dissertation, I will intend this definition.

1.3. Libertarian Paternalism

[23] Ibid., 5.

[24] Some incentives could, arguably, invoke shallow cognitive processes. For example, one possible nudge discussed in the literature is to charge shopper five cents to use plastic instead of paper grocery bags. This is an incentive, but it is unlikely that choosers pick paper bags because they calculate that the cost is too high.

In Nudge and in many subsequent and antecedent articles, T&S defend a position related to the nudge concept called Libertarian Paternalism (hereafter LP) and often use the two terms in seemingly interchangeable ways. However, as we will see, the two terms are not co-extensive. Indeed, as I will argue, the best way to think about the relationship between LP and nudges is that LP is a justificatory strategy for the use of certain kinds of nudges. Unfortunately, the literature creates an ambiguity between at least two different possible roles of LP in justifying nudges. In this section I will differentiate between these two senses and arrive at the meaning of LP that I intend.

1.3.1. The Narrow Interpretation of LP

In some writings it appears that LP is the name given to a very particular justificatory strategy for the use of nudges. Here is how the Nudge Blog discusses the term:

> It's important to point out that nudging complements a libertarian paternalism outlook about public policy, but the two are distinct concepts. Libertarian paternalism is intended as a means to help people make decisions that make them better off as defined or judged by themselves – not by a government or private authority. While the nudges cited in the book are intended to do exactly this, nudging takes place in [a] variety of realms where the nudger's explicit goal is to promote [the nudger's] own welfare.[25]

On this interpretation, LP justifies the use of nudges through two very specific ethical commitments. First, under this interpretation of LP, by "benefit" we mean that nudges promote what the chooser would want if fully informed and rational. Put another way, LP appears to

[25] John Balz, "A nudge on a hot button issue: Abortion," *The Nudge Blog*, 2008.

endorse an informed desire criterion of welfare. Second, this view is committed to the idea that when a choice architect changes the behavior of a chooser, it must be for the chooser's benefit. If we interpret this commitment strictly, it means that, on the narrow interpretation of LP, we could not nudge more people to sign up for organ donation because doing so does not benefit the chooser (although it greatly benefits others).

1.3.2. The Wide Interpretation of LP

Another kind of interpretation of LP is as any of a class of theories that work to improve the wellbeing of choosers (paternalism) while preserving their choice set (libertarianism). This interpretation leaves the precise definition of wellbeing open such that many possible account of wellbeing are possible. As an example, one could intend an informed desire account of wellbeing, or one could intend a hedonistic interpretation of wellbeing or one could intend a perfectionist account of welfare (as will be discussed in Ch 3). This interpretation provides a nicely straightforward interpretation of the title itself; after all, there is nothing in the name 'Libertarian Paternalism' that ought to suggest an informed desire account of wellbeing.

1.3.3. Why I Prefer the Wide Interpretation of LP

For my part, I prefer the wide interpretation of LP. It seems clear that T&S have a particular predilection for the informed desire account of wellbeing, and it is true that others have discussed it in those terms as well, but we need not read anything *fundamental* into this fact. Instead, I would argue that the usage of the term seems to endorse the narrow interpretation because the discussion has not yet matured. To briefly explain: many of the examples of potential nudges that

have been discussed involve cases where the decisions of choosers seem to not be in the chooser's best interest. The easiest way to pump this intuition is to use cases where, had the chooser had better information, it seems clear the chooser would have made a different decision. The most paradigmatic case of this is 401(K) enrollment. It seems clear that if most people took the time to calculate the various uses of their money, they would conclude that enrolling in a 401(k) plan is the superior option. Yet, many people do not choose to enroll. Because it seems clear that choosers *would* want to enroll in the 401(k) plan if they considered it carefully, it provides an easy example case for writers like T&S. However, we need not make use of an informed desire account of wellbeing to justify this intuition. Indeed, having more money rather than less can be justified because it provides benefits in happiness on a hedonic account or it can be justified because it makes it easier for one to get the good things in life on a perfectionist account. I think that, as the discussion matures, it will become clear that one need not make use of the informed desire account to justify nudges.

Therefore, when I refer to libertarian paternalism I will refer to any viewpoint that attempts to make choosers better off while preserving their choice set. However, for the most part, in this dissertation, I will not talk about libertarian paternalism and will instead be discussing the ethics of nudges and choice architecture. This is mostly a semantic decision. While I agree with T&S that LP is not a contradiction in terms[26], I think it is a very confusing term because it does *appear* to be contradictory at the outset. So, I will avoid it. However, the goal is this dissertation is to provide an understanding of the ethical evaluation of nudges that will be useful to libertarian paternalists.

---

[26] Sunstein and Thaler, "Libertarian Paternalism is Not an Oxymoron," *Social Science Research Network*; cf. Mitchell, "Libertarian Paternalism is an Oxymoron."

Chapter 2: When is a nudge morally acceptable?

2.1 Introduction

In this section of the dissertation I attempt to accomplish two objectives. First, I think that so far the discussion of nudges in the literature has focused on whether nudges are permissible or impermissible *in general*. I hope to show that this dialectic is far too simplistic as the specifics of the nudge in question will determine whether the nudge is permissible. Second, I argue that, in general, nudges are a *pro tanto* moral wrong. That is, they are a moral wrong that can be outweighed by other considerations especially the positive benefits that the nudge might achieve. I then consider a number of exacerbating and mitigating factors that determine whether the *pro tanto* wrong of the nudge is in fact mitigated by the positive benefits of nudge implementation. Exacerbating conditions include manipulation, whether the subject endorses or would endorse the nudge, whether the relationship between nudger and nudgee is one of high trust and whether the state is the one implementing the nudge. Mitigating conditions include whether the nudge is noncontrolling, situations in which the agent is not reasons-responsive, where the nudge only benefits the chooser and when the nudge makes the chooser better off as defined by themselves. I divide these claims into three primary categories: nudge type/mechanism considerations, agent-relative considerations, and ends-based considerations.

2.2 Nudge type/mechanism considerations

2.2.1 Introduction

In this section I review considerations relative to the *type* of nudge being used or the mechanism of action for the nudge. By type of nudge, I refer to how the nudge alters the choice context. For

example, take the classic 401(k) enrollment nudge where employers automatically opt employees into the company-sponsored 401(k) plan to produce greater enrollment rates. Here, the type of nudge is altering defaults. However, we might also produce greater enrollment rates by framing the decision to not enroll in terms of money lost (from failure to get employer matching) to take advantage of loss aversion. Different nudge types and mechanisms have different ethical considerations as I will explain.

2.2.2 Nudges are morally problematic because they are manipulative and therefore they subvert rationality and threaten personal autonomy

In this section I consider a cluster of claims concerning the idea that many nudges are morally problematic because their mechanism is manipulation and this causes them to subvert rationality and therefore to impair personal autonomy among other related harms. It is very likely that this cluster of claims forms the core of the concern that many authors have about the implementation of nudges. Below I will first outline some of the concerns raised by a number of authors about the influence nudges have on reason and autonomy, then I will ultimately conclude that this claim is essentially true and that as a result nudges are *pro tanto* morally problematic. However, this does not imply that they are always or even usually all-things-considered morally problematic. I will argue that while one ought to prefer rational persuasion in cases where the likely outcomes are the same, the wrongness of nudges is minor and easily outweighed by other considerations.

2.2.2.1 Why nudges are manipulative

In explaining why nudges are manipulative and the ethical implications of such manipulation, one needs to discuss the relationship between manipulation, reason, and autonomy. One way to think about this relationship is outlined by Blumenthal-Barby and Burroughs in their paper "seeking better Health Care Outcomes: The Ethics of Using the 'Nudge.[27]'" They argue that the connection is as follows:

> Manipulation occurs when one influences another by *bypassing their capacity for reason*, either by exploiting nonrational elements of psychological makeup or by influencing choices in a way that is not obvious to the subject. By virtue of it bypassing a person's capacity for reason, manipulation bypasses the exercise of autonomy as well. It blocks the consideration of all options and threatens the agent's ability to act in accordance with her or his own preferences (as opposed to someone else's).[28]

Therefore, on this interpretation we might be concerned about nudges because they may bypass the ability of persons to act in accordance with their own preferences. Similar concerns are raised by Luc Bovens who writes that "there is something less than fully autonomous about the patterns of decision making that *Nudge* taps into. When we are subject to the mechanisms that are studied in 'the science of choice', then we are not fully in control of our actions.[29]" Finally, Hausman and Welch argue that nudges threaten autonomy if they undermine the nudgee's "control over [her] own evaluations and deliberations[30]" where this control is identified with autonomy.

---

[27] Jennifer Blumenthal-Barby and Hadley Burroughs, "Seeking Better Health Care Outcomes: The Ethics of Using the 'Nudge,'" *The American Journal of Bioethics* 12, no. 2 (2012).
[28] Ibid., 5.
[29] Luc Bovens, "The Ethics of Nudge."
[30] Daniel Hausman and Brynn Welch, "Debate: To Nudge or Not to Nudge," *The Journal of Political Philosophy* 18, no. 1 (2010): 125.

I agree with these concerns. In fact, given that I define nudges as necessarily relying on shallow cognitive processes, and given that many nudges work best if the nudgee does not know she is being nudged, it seems clear that nudges are manipulative and subversive of reasons. The question then is what normative weight this fact has. That is, why is it wrong (if it is wrong) for nudges to subvert reasons and how does this wrongness compare to the supposed benefits of using nudges? To explain the wrongness of subverting reasons, I offer two kinds of accounts. On the first, ensuring that agents are autonomous and reasons-responsive is good extrinsically because it prevents them from being influenced to achieve ends that are not in their interest. On the second, being reasons-responsive is intrinsically good because it gives the actions of agents moral worth.

2.2.2.2. Extrinsic Account of the Wrongness of Nudges

The first account of the moral value of letting agents act autonomously and for their own reasons essentially holds that autonomy and acting for reasons are valuable because they lead to better outcomes. A common way of arguing for this conclusion is to hold that decision makers generally have much better access to their interests than others do and thus are in a better position to maximize those interests. For example, imagine I am trying to decide what to eat for dinner. An external agent like a Choice Architect might be able to guess that I have some interest in the price, health quality and taste of the food. However, itr would be very difficult for a Choice Architect to know how these interests trade off in my particular case. In my case, I am relatively young and a poor graduate student. Therefore I would tend to have large interest in reducing the cost of food, and very little interest in increasing its health value. If one nudged me

towards eating healthier, it might create a resulting outcome that is not in accordance with my interests. Therefore, if we manipulate choosers and disassociate their choices from their reasons, it may create worse choices as a result.

A related concern might be called the "government knows best" critique which surfaced in the public discourse surrounding the possibility of a U.S. "Nudge Squad." Michael Thomas, an economist at Utah State University said: "I am very skeptical of a team promoting nudge policies… ultimately, nudging… assumes a small group of people in the government know better about choices than the individuals making them.[31]" However, this kind of argument need not only apply to the government. Indeed, we might wonder why *any* group of people would be in a better position to know the benefits making certain choices than the people making them. We might extend the argument by defining characteristics that make certain Choice Architects uniquely unsuited to implementing nudges. For example, owners or operators of for-profit companies like those in charge of cafeterias or administering 401(k) plans might be poor Choice Architects because they have a strong profit motive which may ultimately conflict with the needs of the choosers. Similarly, one might argue that the government is a poor choice architect because of a general distrust for the government or because the government already exerts a significant degree of control over our lives. This provides us with a reason to be skeptical of the nudge mechanism in a wide range of particular cases.

2.2.2.3 Against the Extrinsic Account of the Wrongness of Nudges

I do not find the extrinsic account of the wrongness of nudges persuasive as it pertains to properly implemented nudges. It is particularly hard to justify in the face of the robust evidence

[31] "Gov't Knows Best? White House Creates 'Nudge Squad' to Shape Behavior," Fox News, last modified July 30, 2013, http://www.foxnews.com/politics/2013/07/30/govt-knows-best-white-house-creates-nudge-squad-to-shape-behavior.

on various flaws in human decision making. For example, while it is true that choosers have better access to, for example, their financial interests than a choice architect might, it is quite a leap to argue that they are actually *acting* in their best interest when they fail to save or fail to take advantage of matching money on their savings. In any case, nudging is not committed to the idea that some external party knows better than choosers do *in general*, it is only committed to the claim that a choice architect can help choosers make better choices in some cases where there is reason to believe that altering a choice context can improve better outcomes. Additionally, an important aspect of many nudges is that choosers do not bring to bear their full cognitive resources to the problem at hand. For example, it is probably not the case that employees who fail to save in their company 401(k) do so because they calculate the options and determine that it is in their best interest not to save. They probably fail to save because they did not fully consider the choice, or meant to register later and forgot. The choice architect, on the other hand, does bring to bear her full cognitive resources in considering the options in order to determine which is likely to produce a positive outcome. Therefore, the claim is that a choice architect fully considering a decision might make better decisions in general than a chooser using shallow cognitive processing to make a decision. On this interpretation it seems quite plausible that we can discover some particular cases where we would prefer that Choice Architects take advantage of the nudge mechanism.

2.2.2.4 Intrinsic Account of the Wrongness of Nudges

A second kind of argument argues that there is intrinsic value in having agents act autonomously, for reasons, and free from manipulation and argues that because of this intrinsic value, there is an

intrinsic wrongness to nudging. Note that such an account cannot depend on the extrinsic bad effects that nudges may sometimes produce. That is, it cannot lean on cases where the nudges might produce negative outcomes like worse choices by choosers. It also cannot rely on the potential for abuse or misuse in the future. Instead, such an account must support the idea that nudges are wrong even when the manipulation aims at getting the chooser to act in accordance with her own reasons. Following Gorin,[32] this account needs to explain the wrongness of "reasonable manipulation," that is, manipulation done to get a manipulee to act in a way that the manipulator believes the manipulee has reason to act. The following pair of examples from Gorin will help illustrate the kind of manipulation inherent in reasonable manipulation and also the intuition that there is something objectionable about this manipulation:

> *Airport 1*: Larry arranges to ride to the airport with William. At the last moment William becomes ill and thus Larry is in need of a ride from someone else. He knows that Katherine very likely will give him a ride if he explains to her both the importance of his getting to the airport on time and the circumstances leading up to his asking her for a ride on such short notice. He also knows that Katherine will very likely give him a ride to the airport if he simply makes his request with no explanation, but only if he asks her immediately after subtly reminding her of the time she inadvertently embarrassed him at a faculty meeting, an occasion about which Larry knows Katherine feels much guilt. Larry has long stopped being bothered by what Katherine did at the faculty meeting and does not believe that her embarrassing him provides her with a good reason to give him a ride. Larry describes his situation to Katherine, explaining the importance of his getting to the airport, the suddenness of William's illness, and so on, and then asks her for a ride. Katherine agrees to drive Larry to the airport and does so as a result of recognizing the reasons that speak in favor of her doing so, the very reasons that motivated Larry to make his request.
>
> *Airport 2*: This case is identical to *Airport 1* except that here Larry offers no explanation of why he needs to get to the airport or why he is making his request on such short notice. Instead, he first elicits Katherine's feelings of guilt and then simply asks her for a ride. Katherine agrees to drive Larry to the airport and she does so because she feels guilty about embarrassing him at the faculty meeting.[33]

---

[32] Moti Gorin, "The Nature and Ethical Significance of Manipulation," (PhD diss., Rice University, 2013).
[33] Ibid., 100.

Intuitively, Larry's behavior in *airport 2* is objectionable but not in *airport 1*. According to Gorin, the explanation for this intuition is in the way Larry's influence tracks or fails to track reasons. In *airport 1*, Larry's reasons for asking Katherine to help and her reasons for helping are the same, namely, to help Larry catch his flight. On the other hand, in *airport 2*, Katherine's reasons for helping are not the same as Larry's reasons for asking. Gorin explains the wrongness of using means of influence that fail to track reasons as follows: "such means of influence leave those whose behavior they target detached from the considerations we believe ought to govern their behavior, consequently rendering their behavior lacking in normative worth.[34]" For my part, I agree with Gorin's analysis that it is *pro tanto* wrong to influence reason-responsive agents through means that fail to track their reasons. This observation can help to explain the common intuition that, given the choice between nudging an agent to engage in some activity and rationally persuading an agent to engage in some activity, one ought to prefer to rationally persuade the agent.

One might object that one needs to know more about Katherine to determine if Larry has done something wrong in Airport 2. For example, imagine that Katherine endorses being motivated by guilt or thinks that inadvertently embarrassing Larry at the faculty meeting provides her with good reason to take him to the airport. In such a case, presumably, Larry has not done anything wrong despite not engaging Katherine's capacity for reason. I would argue that in such a case, the guilt clearly does provide Katherine with a reason for taking Larry to the airport, but that not all ways of using this guilt are equally good at respecting her capacity for reason. One way to think about the distinction is to look at how reminding Katherine of her guilt

[34] Ibid.

is operative in her. If the guilt is operative at the level of cognitive considerations and other reason-directed behavior, then reminding her of the guilt respects her capacity for reason. On the other hand, if the guilt is introduced in such a way that it primarily plays on her emotions or on her shallow cognitive processes, then it is likely manipulative and not respecting her capacity for reason. There are, of course, many other factors that are relevant in this case, but discussing them in full is beyond the scope of this dissertation.

Extending this case further, we might imagine a situation in which Katherine does not care about being reason-responsive and wants to be manipulated. Is there an argument that nudging is an objective wrong in this case? To answer this question, we must divide this case into two kinds of cases. In one case, Katherine evaluates a particular situation and determines that she wants to be nudged. For example, imagine Katherine is on a diet and trying to lose weight. She determines that unless her husband nudges her to eat healthier by hiding all of the unhealthy food, she will have the willpower necessary to continue on her diet. So, Katherine asks her husband to nudge her to continue the diet. This case does not seem morally problematic because Katherine exercised autonomy in initially deciding that she would prefer to be nudges, so her husband's actions respect her autonomous decision.

A second kind of case would be one in which Katherine has a general lack of a concern for her autonomy. In this case she simply does not care if she is manipulated and does not value making autonomous decisions. In this case, Gorin's analysis does imply that manipulating Katherine is *pro tanto* morally wrong. If we manipulate Katherine in this case we still leave her detached from the "considerations we believe ought to govern [her] behavior, consequently

rendering [her] behavior lacking in normative worth.[35]" However, I would argue that the degree of wrongness is smaller than in a case where Katherine wants her behavior to be autonomously driven because, all else being equal, it is worse to frustrate someone's desires. So, in such a case, the wrongness of violating Katherine's autonomy consists in both the violation of autonomy and the violation of her desire to have her autonomy.

Finally, as Gorin notes, it is important to keep in mind that just because something is *pro tanto* wrong does not imply that it is all-things-considered wrong. For example, it is *pro tanto* wrong to lock someone in a jail cell, but it is unlikely to be all-things-considered wrong to lock violent murderers in jail cells. In the next section I will briefly examine some of the considerations that determine if a nudge is all-things-considered wrong. This analysis will be extended throughout this chapter.

2.2.2.5 When do other considerations outweigh the *pro tanto* wrongness of nudges?

As mention in the previous section, I agree that nudges are *pro tanto* morally wrong because they dissociate choosers from their reasons. The intuition that undergirds this claim is helpfully revealed by our differing reactions to *Airport 1* and *Airport 2* above. However, to determine how the *pro tanto* wrongness impacts the all-things-considered wrongness of nudges, we must first investigate the intuitions more carefully. I argue that part of the intuition that Larry did something wrong in *Airport 2* is generated by his triggering negative emotions in Katherine (e.g. guilt, obligation) whereas no such negative emotions are created in *Airport 1*. Therefore, part of the intuition is that it was wrong of Larry to make Katherine feel bad given that he could have

[35] Ibid.

avoided doing so. To help remove this component from the intuition, consider the following example in which the manipulation relies on positive emotions instead:

> *Parent 1*: Kelly has two teenaged sons, Tim and Patrick. Tim has been playing for many hours with a video game that he and Patrick share, and Patrick is upset that he has not received his turn to play with it. Kelly knows that she can get Tim to share the game with Patrick is she simply explains to him the importance of taking into account the interests of others and of sharing with his brother. She also knows that she can get him Tim to share the game with Patrick by subtly reminds him of a time that his friends praised him for his generosity towards another student in class, an event which Tim feels very proud about. Kelly does not believe that having been complimented by his friends provides Tim with sufficient reason to share the game because she instead believes that Tim should share the game because doing so is the right thing to do. Kelly explains the reasons for sharing to Tim and he shares the game with Patrick.
>
> *Parent 2*: This case is identical to *Parent 1* except that in this case Kelly subtly reminds Tim of the praise he received from his friends eliciting his feelings of pride and acceptance. Tim shares the game with Patrick in order to maintain the positive self-image elicited by his friends.

If Kelly has done something wrong in *Parent 2*, it is a very minimal kind of wrong that does not seem comparable to the wrong of Larry in *airport 2*. However, the relevant details of the cases are the same except that Kelly uses positive emotions to manipulate Tim whereas Larry used negative emotions to manipulate Katherine. The point of this example is that, while it is true that *all other things being equal* one ought to prefer other means of influence like rational persuasion to manipulative nudges, this intuition is not hard to override with considerations like the positive outcomes generated by the manipulation. And, since nudges are often intended to create large positive outcomes, it may turn out to be the case that the *pro tanto* wrongness of using nudges is consistently outweighed by these considerations.

2.2.2.6 Do nudges prevent agents from acting for reasons?

Gorin's explanation of the wrongness of manipulation is that "such means of influence leave those whose behavior they target dethatched from the considerations we believe ought to govern their behavior, consequently rendering their behavior lacking in normative worth.[36]" In cases where we can either have an agent act for reasons or act because of the influence of another, this is a relatively straightforward claim. However, in paradigmatic cases of nudges, it is often unclear that agents are in fact acting for reasons. Take the case of 401(k) enrollment as an example. Research indicates that changing a company's 401(k) policy to automatically opt employees in instead of automatically opting them out will result in much higher savings rates. Imagine for example, that at a company of 100 people, 20 employees will participate in the 401(k) if they are automatically opted out and 40 will participate if they are automatically opted in, meaning that 20 employees change their behavior depending on the choice architecture of the situation. Is it plausible that these 20 employees were not saving for retirement for reasons? If so, it is hard to imagine what their reasons would have been. An alternative explanation is that they were not acting for reasons, and that the nudge simply got them to act differently, but with better outcomes. On this interpretation, the actions of the employees lack moral worth either way, but if we nudge we at least generate the benefit of creating better outcomes.

One might respond to this argument by claiming that it is wrong to manipulate someone *even if* their action, absent your manipulation, would not be based on reasons. However, I think the following example will call this intuition into question:

> *Coin flipper:* Jill has the curious habit of flipping a coin for all of the trivial decisions in her life. She does this because she reasons that she ought to spend her powers of reason only on more important pursuits. Jill is trying to decide if she should get her new car in black or in silver, so she decides to flip a coin. Jill will get a black car if the coin is heads and a silver car if it is tails. Bob thinks Jill has reason to get a black car because it is

[36] Ibid.

> easier to clean, and she has expressed that this is important to her, so he rigs Jill's coin so that it always comes up heads. He tells Jill that he has rigged her coin, but does not tell Jill which result the coin will produce. Jill flips the coin and, as a result, Jill ends up getting a black car.

In this example, Jill has the *capacity* for reason, but does not exercise it on trivial decisions. Bob then manipulates Jill to help her reach a better outcome. Despite this manipulation, it is hard to have a problem with Bob's behavior in this case. One might argue that it would be morally superior for Bob to rationally persuade Jill to get the black car, but even if this is true, the moral difference between the two options seems small.

Proponents of the use of nudges would argue that the paradigmatic nudge case is closer to what Bob does to Jill in *coin flipper* than to what Larry does to Katherine in *airport 2*. Nudges are not designed to take one who has strong reasons for action and attempt to override those reasons. In fact, the paradigmatic nudges do not change the behavior of these people. Take the 401(k) case as an example. If I am not saving for retirement because I have a mountain of credit card debt that needs to be paid off, an opt-in strategy will not alter my behavior because I will simply opt out of the 401(k) plan. It will only alter my behavior if my reason for not saving is procrastination or a lack of completely considering the options. Nudges are only designed to influence the behavior of those who have weak or non-existent reasons for their behavior.

However, in many cases, one does not have to make a choice between existing behavior and a nudge because there are alternative options that might help agents be reasons-responsive. I discuss this possibility in the next section.

2.2.2.7 Are nudges the best available alternative?

In some cases the choice architect does not have to choose between a chooser's unaltered unreasonable (i.e. not based on reasons[37]) behavior and nudge-induced unreasonable behavior because there are other options. Take the 401(k) case as an example. We might instead employ what Thaler and Sunstein call "forced choosing." This occurs when the choice architect requires that employees make a choice about opting in or opting out of the 401(k) plan.[38] We might combine this option with rational persuasion such that employees must choose whether to save or not and arguments for choosing to save are presented to all employees. Are nudges preferable to rational persuasion in cases like these?

If rational persuasion can achieve the same outcomes as a nudge, I find it relatively uncontroversial that rational persuasion is to be preferred. And, given that in ethically-justifiable nudges the choice architect thinks the chooser has reason to act in a particular way, rational persuasion will almost always be an available option. This means that choice architects should determine if rational persuasion is an option to achieve their desired outcomes and, if it is, determine if it would have similar efficacy to a nudge. If so, rational persuasion should be used. However, per my argument in *parent 2* (2.2.2.5), I think the wrongness of engaging in manipulation when rational persuasion is available is often outweighed by other considerations. However, the practical upshot of this point is that, on my interpretation, choice architects must justify their nudges as a prerequisite for their implementation. This is distinct from the position

[37] There is a distinction between behavior that would be based on reasons where it not for the nudge and behavior that would not be based on reasons were it not for the nudge. Here I am concerned with behavior that would not be based on reasons were it not for the nudge.

[38] Of course, there will still be choice architecture in this case because the forced choosing must be presented in some way or another. However, because the chooser will be more likely to engage cognitive resources in the forced choosing it might better respect their capacity for reasons.

taken by Thaler and Sunstein who seems to see nudges as almost entirely unproblematic when properly implemented and executed.[39]

A final point worth making is that in many cases the nudge mechanisms itself carries some cost. For example, in a case where the nudge mechanism is to default employees to being in their 401(k) plan, a cost of this nudge is that many employees will have to complete annoying paperwork in order to get out of the plan. These costs must also be taken into account to determine if a nudge is the all-things-considered best option.

2.2.2.8 Concluding thoughts on nudges and manipulation

In this section I considered the claim that nudges are morally problematic because they are manipulative, subvert rationality and threaten personal autonomy. I ultimately conclude that this claim is essentially true and that as a result nudges are *pro tanto* morally problematic. However, this does not imply that they are always or even usually all-things-considered morally problematic. On this point I argue that it is relatively easy to outweigh the *pro tanto* wrongness of nudges with other considerations and that many nudges do not interfere with reasons-responsiveness because agents who are receptive to nudges are not acting for reasons. However, it is nevertheless true that if one can achieve similar outcomes with rational persuasion as with a nudge, one ought to prefer rational persuasion. This means that choice architects ought to consider the possibility of rational persuasion before implementing a nudge and they ought to be able to explain why rational persuasion will not achieve similar outcomes when nudges are

[39] Of course, Thaler and Sunstein think that choice architecture is inevitable and unavoidable. As a result, since it cannot be avoided, utilizing it is not morally problematic. However, even if it is true that choice is unavoidable, that does not imply that all choice architectures are equally respectful of autonomy and the ability for people to act for reasons.

used. For the remainder of this chapter I will consider potential complicating and mitigating factors that can help determine if a particular nudge is in fact morally acceptable.

2.2.3 A nudge is problematic to the degree that it is substantially controlling and not easily resistible

In this section I consider the idea that whether a nudge is morally problematic is related to whether the nudge is easily resistible. The central claim is that as a nudge becomes less easily resisted and more controlling, it becomes more morally problematic. Here I borrow heavily from the work of Yashar Saghai in defining the nudge concept. As I discussed in chapter 1, Saghai defines a nudge as follows:

> Nudge: A nudges B when A makes it more likely that B will φ, primarily triggered by B's shallow cognitive processes, while A's influence preserves B's choice-set and is substantially noncontrolling (ie, preserves B's freedom of choice).[40]

In chapter 1, I remove the specification that nudges must be substantially noncontrolling on the grounds that I did not think this matched up with the common parlance for the term. However, I do think that the degree of control is an ethically relevant consideration when evaluating a nudge. Recall that by "substantially noncontrolling," Saghai intends the following:

> Substantially Noncontrolling: A's influence to get B to φ is substantially noncontrolling when B could easily not φ if she did not want to φ.[41]

The idea is that influences can be placed on a continuum that ranges from fully controlling (compulsion, coercion, choice elimination etc.) to fully noncontrolling (rational persuasion).

---

[40] Saghai, "Salvaging," 5.
[41] Ibid., 2.

When an action is substantially noncontrolling, it is more controlling than rational persuasion, but less controlling than coercion and is relatively easy to resist. The key idea here is the idea of easy resistibility. In a case where A is trying to nudge B to φ, A's influence is easy to resist if the following are true:

1. B has the capacity to become aware of A's pressure to get her to φ (attention-bringing capacities); and
2. B has the capacity to inhibit her triggered propensity to φ (inhibitory capacities); and
3. B is not subject to an influence, or put in circumstances that would significantly undermine the relatively effortless exercise of attention-bringing and inhibitory capacities.[42]

If influences are not easily resistible, then they are not substantially noncontrolling. As an example of the difference between a substantially noncontrolling and substantially controlling influence consider the following (real life) example:

> Asparagus-Lovers. An investigator suggests to research participants that they liked or loved asparagus during childhood the first time they tried it, creating a false memory, and a false belief about the taste of asparagus. Subsequently, participants reported increased general liking of asparagus, greater desire to eat it, and willingness to pay more for it.[43]

Asparagus-Lovers is not fully controlling because research participants could not eat asparagus if they did not want to. That is, there is some degree of resistibility. However, it is also not easily resistible. This is because Asparagus-Lovers involves a large degree of deception which makes it

[42] Ibid., 3.
[43] Ibid., 1.

very hard for the participants to become aware of the researcher's influence attempts and makes it difficult for participants to inhibit their triggered propensity to like asparagus.[44]

Saghai's definition also gives us some tools to deal with the harm of others kinds of manipulation. For example, imagine if a company encouraged participation in the 401(k) plan by threatening to dock employee pay by 10% if they did not participate.[45] In this case Saghai can turn to condition 3 and argue that such a threat makes it hard to exercise the relevant capacities effortlessly. Thus, the threat is controlling.

2.2.3.1 Saghai's analysis of why control matters morally

In addition to analyzing the idea of substantial control and substantial noncontrol, Saghai offers a useful framework for determining when healthcare nudges are wrong (if they are wrong) and for understanding the ethical importance of noncontrol. It relies on the idea of a self-determining life, which is a "life that is, in its main contours, free from the exercise of power by other individuals and by social and political institutions.[46]" To the degree that an action exercises a degree of control that interferes with the ability to lead a self-determining life, it is morally problematic. Applying this to the use of nudges, we can see two ways that a nudge might be morally problematic.

The first is that a nudge may be problematic if the decision in question is of very large importance in determining the main contours of the life of the chooser. In fact, Saghai contends

---

[44] One might argue that all nudges involve some kind of deception and so are not easily resistible. However, in the case of Asparagus-Lover, the use of false memories makes it very difficult for participants to have the capacity to become aware of the manipulation. On the other hand, most paradigmatic nudges are more easily brought to awareness of choosers.

[45] This would technically not be a nudge because it does not make use of shallow cognitive processes,

[46] Ibid., 7.

that "some health-affecting choices are so fundamental for leading a self-determined life that they ought to be as fully noncontrolled by others as possible.[47]" However, we need not restrict this framework to healthcare nudges. Many things shape the contours of one's life. For example, using this framework it might be especially problematic to nudge one towards attending a particular college, choosing a particular college major or choosing a particular career path because these can have dramatic weight in determining the contour of one's life. This means that nudges must only be used in relatively trivial matters, dramatically undercutting the usefulness of the tool. A second way that nudges or other types of influence might be problematic is if they are overly controlling. That is, a nudge that is easily resistible in the way that Saghai discusses is unlikely to determine the contours of a chooser's life in any meaningful way. However, a more substantially controlling or fully controlling nudge might have this power. Therefore, one way of spelling out Saghai's account of why control matters morally is that it can sometimes remove the ability of choosers to determine the contours of their own life.

2.2.3.3 Gorin's account of why control matters

Gorin's account of why the degree of control matters differs from Saghai's account in a number of key respects. One way to understand this difference is to look at his account of the difference between coercion and persuasion. According to Gorin, "[t]he crucial difference between coercion and persuasion is that only coercion involves a threat of serious harm, which harm it is in the power of the coercer to bring about.[48]" The primary reason for this difference is that persuaders are constrained by reasons whereas coercers are not. For example, a coercer might use a threat to

[47] Ibid.
[48] Gorin, "The Nature," 143.

generate new reasons and such a threat can be used to generate any reasons whatsoever. In this way, coercion subjects the "will of the coercee to the will of the coercer, whatever the content of the coercer's will.[49]" In this way, coercion is arbitrary in the sense that it is not supported by reasons and it is "open-ended" in the sense that the coercer can use her threat to compel any action she wants.

To extend this argument to the case of nudges, we might argue that nudges are similarly open-ended in the sense that a nudge mechanism can be used to increase the probability of all kinds of behaviors subject to the will of the choice architect. For example, one could use defaults to produce increases in savings or decreases in savings depending on the goals of the particular choice architect. On the other hand, a rational persuader cannot do this. One cannot use arguments about the dramatic increases in quality of life that one gets through increased savings to influence a chooser *against* saving money. So, nudges are morally problematic to the degree that they tie the behavior of choosers to the will of the choice architect instead of to their reasons.

This account provides us with an easy way to see why one might prefer easily resistible or substantially noncontrolling nudges to those that are hard to resist and substantially controlling. In short, the more easily resisted a nudge is, the less the chooser is tied to the arbitrary will of the choice architect. This ensures that the choice architect is different from the coercer because easily resistible nudges only work on certain people with certain kinds of reasons whereas coercion works on most people with most kinds of reasons. For example, if a coercer places a gun to someone's head and tells them to save money for retirement, we do not need to know much about that person to guess that they will do it. However, if a choice architect

[49] Ibid.

designs a nudge to encourage employees to save, we need to know quite a bit about the individual employees to guess if they will save or not. In fact, we probably need to know something about their reasons for saving or not saving currently in order to know if the nudge will be successful on a specific person. If those reasons are strong, the nudge will be unlikely to have an influence, but if those reasons are weak or nonexistent, it is far more likely to work. It is preferable that the nudge be receptive to reasons the nudgee currently has instead of being receptive mostly to the will of the choice architect.

2.2.3.4 Control matters because it reduces the probably of harmful nudges

For my part, I agree with the general thrust of Gorin and Saghai's account that there is something valuable in the self-determined choices of choosers. However, I also think that choice architects ought to prefer substantially noncontrolling nudges because they are less likely to produce substantial harms. We can see this in two respects. First, substantially noncontrolling nudges are less likely to result in abuse of the nudge mechanism. For example, recall the "government knows best" critique that surfaced during the public debate over the creation of a nudge squad in the U.S. According to this criticism, it is unreasonable to think the government knows better than choosers about what choice they ought to make. The implicit concern is that the government will design nudges that make choosers *worse* off instead of better off by getting the interests of choosers wrong. However, easily resistible nudges help to avoid this concern. If a choice architect designed an ineffective nudge but ensured that such a nudge was easily resistible, then choosers would be easily able to ignore the nudge and do what is in their best interest instead. In this way, easy resistibility helps to ensure that the nudge mechanism is not abused.

Similarly, easily resistible nudges can sometimes help to increase the aggregate good that a nudge achieves by preventing the nudge from negatively affecting certain choosers. Returning to the 401(k) example, getting employees to participate in a 401(k) is good for the majority of employees, but it might not be good for all of them. Imagine for example that an employee has a lot of high-interest credit card debt. Such an employee might be better served by paying of that debt rather than saving for retirement even given the employee match. If the nudge is easily resistible as it is when one changes the default options, then the employee can easily resist the nudge and pay off the credit cards. However, imagine if instead we did something similar to what is done in Asparagus-Lovers and implanted a false belief memory in the chooser that led her to save in the 401(k) plan. This would be harder to resist and therefore makes it more likely that the chooser would do fail to pay off the credit card debt. So, one benefit of easily resistible nudges is that it makes it less likely that the nudges will have unintended negative consequences.

Of course, difficult to resist nudges may also *increase* the probability that the nudge accomplishes its intended consequences, so, choice architects must carefully weigh the costs and benefits of various nudges to determine whether the nudge should be easier or harder to resist. However, given the inherent uncertainty of the outcomes, other things being equal, easily resisted nudges should be preferred.

2.2.3.5 Concluding thoughts on nudges, control and resistibility

In this section I discussed two kinds of concerns that choice architects might have about the use of nudges both of which can be mitigated by using nudges that are easily resistible. The first is that there is something intrinsically valuable about having choosers make choices free from

nudges. I presented two accounts of why this might be the case. The first is Saghai's account according to which nudges can sometimes affect the ability of choosers to determine the shape of their own lives. The second was Gorin's account according to which nudges are "open-ended" in the sense they can subject the chooser to the will of the choice architect instead of allowing the chooser to act for her own reasons. This concern can be mitigated by making the nudges easily resistible and thus giving the chooser more control over her actions. A second concern is that nudges might results in unintended consequences and harms. This too can be mitigated by making the nudges easier to resist. Therefore, nudges that are not easily resistible are going to be harder to justify morally than nudges that are more difficult to resists.

2.2.4 Nudges that persons do not endorse or would not endorse if aware of are morally problematic

In this section, I consider the claim that whether a nudge is morally problematic depends on whether choosers would endorse the nudge and the behavior the nudge attempted to create if they became aware of the nudge. As discussed in chapter 1, nudges rely on shallow cognitive processes and, as such, the existence of a particular nudge is not always clear to the chooser. However, there are some nudges that choosers might endorse if they were aware of them and others that choosers may not endorse. For example, if I found out that the reason I was defaulted in to my company's 401(k) plan is because they intended to encourage more employees to save for retirement and research indicates that opting them into a savings plan will encourage this, I would probably endorse this nudge. However, if I found out that I was automatically opted in to a website's emailing list so that they could send me spam, I would probably not endorse this

nudge. In what follows, I first try to defend this intuition in a more robust fashion. Next, I some possible justifications for this intuition.

2.2.4.1 Are unendorsed nudges problematic?

In the previous section I considered the hypothesis that nudges that the chooser would not endorse are morally problematic. In this section, I aim to motivate that hypothesis further. Consider the follow cases:

> *Saver 1:* Plex-electronics is attempting to implement an employee 401(k) program. They believe there is good reason to think that saving for retirement would be good for their employees. They also know that their employees would prefer to save if they carefully considered the implications of the decision. As a result Plex-electronics nudges employees to participate in the company 401(k) program.
>
> *Saver 2*: This situation is the same as saver 1 except that Plex-electronics knows they prefer not to save (even though doing so would be best for them) and would object to being nudged towards saving. Nevertheless, Plex-electronics nudges employees to participate in the company 401(k) program.

My claim is that it seems that Plex-electronics has acted *pro tanto* wrongly in *Saver 2* but has not acted *pro tanto* wrongly in *Saver 1*. This is because the knowledge that the employees would not endorse the program in *Saver 2* seems to give Plex-electronics some reason not to implement the nudge. Put another way, it seems worse to know one's preferences and violate them than it does to comply with one's preferences (or violate them accidentally).

2.2.4.2 Endorsed nudges lead to better outcomes

One possible justification for this claim would be that nudges that choosers would endorse lead to better outcomes. When a chooser does not endorse a nudge, it may often be because the chooser does not think the nudge makes her better off. Take the email spam case mentioned above. Getting more spam in my inbox does not make me better off and so it is not a nudge I would endorse. However, if I were automatically opted in to an email list that sent me free gift cards, I would like endorse such a nudge. Similarly, restricting nudges to those that choosers would endorse helps to ensure the nudge mechanism is not abused. Such a criterion makes it more difficult for choice architects to justify nudges that do not actually benefit the choosers but instead benefit someone else. For example, this criterion would prevent companies from nudging employees towards choosing investment funds that have higher fees to benefit the investment company. It also makes it less likely that choice architects will engage in nudges that they *think* will benefit choosers, but instead do not, as such policies would likely not be endorsed by choosers if they became aware of them.

While any of these claims may be true, it reduces the idea that nudges must be endorsed by choosers to a claim about the *outcomes* of nudges. That is, it depends on the more general idea that nudges ought to produce positive outcomes for choosers and becomes a useful heuristic for ensuring that this occurs. However, if this is the correct way to understand the claim, then it is not problematic to use a nudge that choosers would not endorse if such a nudge produces large positive outcomes. In the next section I will consider arguments that the value of nudges that choosers would endorse does not depend on the outcomes of the nudge.

2.2.4.3 Endorsed Nudges and Autonomy

As mentioned previously, one of the standard concerns about nudges, manipulation and other forms of nonargumentative influence is that use of these tools reduces the autonomy of the chooser. Autonomy involves "shaping one's own life in ways that one finds valuable or important, as opposed to going through life mindlessly or based on other people's agenda."[50] Many have argued that influences like nudges, which utilize shallow cognitive processing threaten autonomy because they do not give choosers an opportunity to exercise the capacity to decide how they would like their life to be shaped. For example, if a doctor presents a patient with arguments for choosing a particular medical intervention, the patient can choose the intervention that best accords with her values and helps her achieve her life goals. On the other hand, if the doctor nudges the patient towards a particular intervention in a certain kind of way, she no longer has this ability.

However, not everyone has agreed that all forms of nonargumentative influence are threatening to autonomy. For example, Sarah Buss[51] has argued that nonargumentative influences like nudges are entirely compatible with autonomy. To see how this could be the case, consider that all of our choices are influenced by nonargumentative influences constantly. Recall the example of a patient deciding on a surgical intervention. In addition to being influenced by the doctor, the patient is influenced by the location of the decision, the physical appearance of the doctor, the patient's past experiences with doctors and similar decisions, the patient's mental state at the time, how recently the patient has eaten, the time of day, and so on. Thus, there are

---

[50] Gerald Dworkin, *The Theory and Practice of Autonomy* (Cambridge University Press, 1988), 164, quoted in Jennifer Blumenthal-Barby, "Between Reason and Coercion: Ethically Permissible Influence in Health Care and Health Policy Contexts," *Kennedy Institute of Ethics Journal* 22*,* no.4 (2012): 352.

[51] Sarah Buss, "Valuing Autonomy and Respecting Persons: Manipulation, Seduction, and the Basis of Moral Constraints," *Ethics* 115 (2005).

numerous nonargumentative influences on the patient's decision *even if* the doctor attempts to utilize rational persuasion. Because there is no reasonable distinction between these environmental influences and nonargumentative influences like nudges, then either all of these influences threaten autonomy or none of them do. Because we presumably do not want to say that we have no autonomy due to these environmental influences, we ought to conclude that nonargumentative influences like nudges are not threatening to autonomy.

At the same time, we probably do not want to conclude that *nothing* threatens autonomy. For example, imagine that a doctor presents information in such a way that I am more likely to choose a particular surgery because my choosing the surgery will make the doctor more money. This *does* seem like a threat to autonomy. So, what distinguishes this case from environmental nonargumentative influences? One plausible answer is whether I would endorse the particular kind of influences or not. I might be willing to endorse environmental influences and the influence they have on my life even if I was aware of them, but I would probably not be willing to endorse the outcomes of the doctor's manipulation.

Whether one thinks this kind of endorsement is sufficient for autonomy depends on the particular view of autonomy that one holds. While, adjudicating between different accounts of autonomy is beyond the scope of this chapter, I will briefly highlight three competing views and show what a proponent of each view might say about whether endorsement implies autonomy.

The first view is what we might call the *coherentist* view. On this account "an agent governs her own action if and only if she is motivated to act as she does because this motivation coheres with (is in harmony with) some mental state that represents her point of view on the

action."[52] On this account, the cause of the motivation is not important, only its harmony with other mental states. So, as long as the chooser would endorse the behavior induced by the nudge, where "endorse" means "recognize as harmonious," then endorsed nudges are fully autonomous.

A second account is what we might call the *reasons-responsive*[53] account. This account emphasizes that for an agent to be autonomous she must be sufficiently responsive to the reasons that stand behind or back up her motives. So, the agent must be capable of understanding what she has reason to do and be capable of acting on those reasons. This also seems compatible with the autonomy of endorsed nudges. As an example, we can imagine an agent saving for retirement because she was nudged to do so, but we can also imagine her endorsing saving for retirement because of the reasons in favor of doing so. On this account, the agent would be autonomous.

The final account is called the *responsiveness-to-reasoning* account. On this account, "the essence of self-government is the capacity to evaluate one's motives on the basis of whatever else one believes and desires, and to adjust these motives in response to one's evaluations.[54]" This reflects the intuition that "someone whose education consisted of a method of indoctrination that deprived her of the ability to call her own attitudes into question would, in effect, be governed by her 'programmers,' not by herself.[55]" This is different from the reasons-responsive account because it only requires *reason capacity*; it does not require that one actually consider the reasons for a particular action. On this account, an endorsed nudge probably do not entail autonomy.

---

[52] Sarah Buss, "Personal Autonomy," *Stanford Encyclopedia of Philosophy* (2013).
[53] Ibid.
[54] Ibid.
[55] Sarah Buss, "Personal Autonomy," *Stanford Encyclopedia of Philosophy* (2013).

This account of autonomy involves adjusting one's motives based on beliefs and desires, but, endorsing a nudge does not necessarily ensure that one has this capacity. If I successfully nudge you to save and then you become aware of and endorse this nudge, but do not alter your future motivations in favor of saving, then you are not autonomous on this view. However, the *responsiveness-to-reasoning* account also entails that typical cases of nudges do not harm autonomy. The account only requires *reason capacity* and does not require the actual exercise of reason. On this view, if a nudge is performed on a person with reason capacity then the person is still autonomous. Thus, while this view does not entail that an endorsed nudge means the person is autonomous, it also does not entail that typical causes of nudges threaten autonomy.

In this section I have shown that the fact that a nudge is endorsed is sufficient to demonstrate that the agent is autonomous on some views of autonomy. This was not true on the *responsiveness-to-reasoning* account of personal autonomy, but as I argued, nudges are not threatening to autonomy on this to begin with. Thus, it seems plausible that either nudges do not threaten autonomy or, if they do threaten autonomy, endorsed nudges avoid this concern.

2.2.4.4 Against Endorsed Nudges

In this section I discuss an argument one might have for rejecting the claim that a nudge is not morally problematic if the agent does or would endorse the nudge.

First, we must break the argument into two claims. The first is that a nudge is morally problematic if a chooser does not endorse the nudge (after becoming aware of it). Call this the Actual Awareness Claim. The second claim is that a nudge is morally problematic if a chooser *would not* endorse the nudge *if she became aware of it*. Call this the Hypothetical Awareness

Claim. As already discussed, because nudges utilize shallow cognitive processing, most of the time choosers will not be aware of the existence of nudges used on them. This means that the Hypothetical Awareness Claim is more important for deciding if a nudge is appropriate. It is also more problematic.

The problem with the Hypothetical Awareness Claim, is that there may not be an endorsement or rejection of a nudge for choosers independent of how that nudge is brought to their attention. The idea here is related to Thaler and Sunstein's claim that the anti-nudge position is a nonstarter because there is no neutral choice architecture. That is, for most decision, the decision must be framed in *some* context, so there can be no threat to freedom or autonomy to merely choose one framing over another. Extending this argument to this case, there is no neutral way for choosers to become aware of the nudge, and how that awareness is brought to attention may have a profound impact on their endorsement of the nudge. This may mean that there is no real answer to the question of what a chooser would endorse if they became aware of a nudge. If the so-called preference on the part of the chooser is so weak that a nudge can affect their decision, it might not make much sense to label the chooser as endorsing or not endorsing the nudge.[56]

This problem with the Hypothetical Awareness Claim works in some cases, but I do not think it works in all cases. To take an extreme example, imagine I nudge you to give money to a political party whose policies you strongly disagree with. It seems reasonable to think that I can

[56] One might argue that we could bring the nudge to the attention of the chooser with some kind of neutral framing that avoids these problems. While Thaler and Sunstein would argue that the idea of a truly neutral framing is a non-starter, it might be possible to mitigate framing effects to some large degree. However, in practice, I do not think this would be effective. As I argue elsewhere, we should expect nudges to work mostly in cases where the preference on the part of the chooser is very weak. In such cases, any remaining framing effect would remove the practical utility of trying to bring the nudge to awareness.

predict how you would react if you become aware of this nudge even without knowing anything about the framing used to reveal the nudge itself. There may be some framing according to which I can cause you to endorse this nudge, but this is a small minority. It seems to miss the point to say that because there are some possible framings according to which you will endorse the nudge your hypothetical endorsement is indeterminate. It seems more plausible to say that you do not endorse the nudge because, when we consider the space of possible framings, we predict that you will not endorse the nudge on most of them. On this account, we can say that a person would endorse a nudge if they became aware of it if we reasonably predict that the person would endorse the nudge on most possible framings of the nudge. This allows us to avoid the theoretical problem in some, but not all cases.

2.2.4.5 Concluding thoughts on nudges and endorsement

In this section I consider the idea that a nudge is morally problematic if a chooser would not endorse the nudge is she became aware of it. I first consider the idea that this claim might be true because nudges that are endorsed by choosers lead to better outcomes. I then consider the claim that nudges that choosers would not endorse are problematic because they interfere with autonomy. I ultimately note that note that this claim is true on most conceptions of autonomy and that, on the account where this is not true, nudges are not a significant threat to autonomy. I conclude that in general nudges that choosers endorse are not threatening to autonomy.

2.3 Agent-Relative Considerations

2.3.1 Introduction

In this section I review considerations relative to the agents involved in a particular nudge. The basic idea of this section is two-fold: 1) certain kinds of relationships between choice architect and chooser can determine if a nudge is acceptable; 2) the use of a nudge by certain kinds of choice architects on certain kinds of choosers can determine if the nudge is acceptable. In this section I will consider a number of possible claims that consider the agents involved and will ultimately arrive at a set of considerations that are relevant for the choice architect.

2.3.2.1 A nudge is morally problematic if it is performed in the context of high-trust relationships

In this section I consider the idea that nudges are morally problematic if they are performed in the context of a high-trust relationship. Examples of such a relationship include those between a physician and a patient, a husband and wife, or two very close friends. The basic idea is that, whereas I might expect certain choice architects such as advertisers to nudge me, I do not expect those that I trust to nudge me, and therefore their nudges are morally problematic. An example will reveal the intuition:

Amazon: Frank, a manager at the online company Amazon.com wants me to buy a set of cooking knives for $50 from their online store because he claims a commission on the sale. In order to get me to buy the knives he nudges me by displaying the price of the knives as $100 and saying that they are 50% off for a limited time. While I want to buy new knives eventually, I would prefer to wait and buy them next month when I expect to have more money. However, because the knives are on sale, I decide to buy the knives.

Best Friend: My best friend, Sam, wants me to buy a set of cooking knives for $50 because she will claim a commission on the sale if I do. In order to get me to buy the knives, she tells me that the knives are on sale for $50 instead of $100 for a limited time only. She does not disclose that the sale is running for many months and that the knives are almost always on some kind of sale. While I want to buy new knives eventually, I would prefer to wait and buy them next month when I expect to have more money. However, because the knives are on sale, I decide to buy the knives.

Amazon does not seem objectionable whereas *Best Friend* does.[57] Put another way, if I found out about the nudge performed by Frank, I would be unlikely to be particularly upset about it, but if I found out about the nudge performed by Sam, I would likely be upset about it. I think the different intuitions are mostly explained by differing expectations for how Frank and Sam will behave and for what intentions they will posses. Because Frank is representing a for-profit company, and because Frank's job is to make that company money, I expect Frank to behave in ways that will get me to purchase things from his company and to only intend that I provide his company with profit. This may involve nudges in the online arena or a variety of sales tactics if I dealt with Frank in person. Because I expect that this will occur, it does not seem objectionable. On the other hand, I do not expect Sam to behave similarly. While I do expect Sam to try to sell me knives in the context of discussing knife sales, I expect her to have additional motivations and considerations as part of her motivational set. For example, I expect her to value our relationship and to refrain from doing things that might jeopardize that relationship. These

[57] One might correctly point out that the relevant moral difference here is in intentionality. It is OK for Frank to merely intend to make money off of me, but it is not OK for Sam to do the same. I agree with this assessment. In fact, if Sam nudged me to buy knives due to a belief that buying knives would make me happy, I would not find that nudge problematic. This implies that we must consider the motivations behind the nudge in addition to the contours of the nudge itself.

include violating our shared trust by attempts to get me to do things that are not in my best interest or by failing to disclose relevant information. However, the shared history that I have with Sam in *Best Friend* need not be the only explanation for the intuition. For example, imagine I find out that a doctor was nudging me to opt for a particular surgery because he makes money performing the surgery. Even though I may not have a long-standing relationship with the doctor, I may find this objectionable because I expect doctors to act in ways that are focused on maximizing my health instead of acting in ways that are focused on maximizing their profit.

Some might object that nudging in high-trust relationships might be more acceptable, not less because in such relationships we expect people to look out for each other and nudging is a way of doing so. As I will explore in 2.3.2.2. this depends a bit on what the norms are in the relationship in question. If the norms are such that non-argumentative influence is not surprising, then the nudge will be more easily resistible and therefore less problematic. My guess is that most high-trust relationships do not have norms compatible with easily resistible nudges, but this may well be false in many cases.

In the next section I will explain why I think high-trust relationship are a relevant moral consideration by arguing that high-trust relationships pose greater difficulties in bringing attention to the nudge and are therefore harder to resist.

2.3.2.2 High-trust relationships and easy resistibility

As I mentioned in the previous section, it seems to be the case that nudges in the context of a high-trust relationship are more objectionable than nudges where no such relationship exists. This fact can be mostly explainable in terms of the expectations that the chooser has about how

the choice architect will interact with them. In this section I will argue that whether one expects to be nudged is a morally relevant factor because unexpected nudges are harder to resist.

The idea of easy resistibility was introduced in chapter 1. Recall that a nudge is easily resisted if:

1. B has the capacity to become aware of A's pressure to get her to φ (attention-bringing capacities); and
2. B has the capacity to inhibit her triggered propensity to φ (inhibitory capacities); and
3. B is not subject to an influence, or put in circumstances that would significantly undermine the relatively effortless exercise of attention-bringing and inhibitory capacities.[58]

The idea is that high-trust relationships violate criteria 3 from Saghai's analysis. When I am dealing with someone that I do not expect to nudge me, I am put in a circumstance that inhibits my attention-bringing capacities. That is, I am not looking for the high-trust person to nudge me, so my guard is down. Contrast this with a case where I am interacting with a salesperson at a store. In that situation, I expect that the salesperson is going to attempt to sell me something and, accordingly, I can react with proper skepticism and bring my attention to potential nudges. And, as already discussed, all things being equal, it is preferable that a nudge be easy to resist instead of hard to resist.

2.3.2.3 Concluding thoughts on nudges and high-trust relationships

[58] Saghai, "Salvaging," 3.

In this section I considered the claim that nudges where the choice architect—nudgee relationship is high-trust are particularly problematic. I argue that a high-trust relationship is a specific instance of a case where one cannot engage "relatively effortless exercise of attention-bringing and inhibitory capacities" and so may be morally problematic. However, the actual norms implicit in the relationship in question determine whether the nudge is in fact problematic. There may well be many cases of high-trust relationships where a nudge would not be a departure from the established norms.

2.3.3.1 A nudge is not morally problematic if the agent is not reasons-responsive or autonomous

In this section I will consider the claim that nudges on agents who are either not acting autonomously or who are not reasons-responsive are not problematic. This claim stems in large part from the work in section 2.2.1 in which I discuss the claim that nudges are problematic because they are manipulative, subvert rationality and threaten personal autonomy. While I do not agree with that claim in full, I do argue that nudges are manipulative and so, it is preferable to avoid nudges if one can achieve the same results through other means. However, not all agents act autonomously and not all agents act for reasons. For example, some agents may be practically irrational and not all agents are engaged in "shaping one's own life in ways that one finds valuable or important, as opposed to going through life mindlessly or based on other people's agenda"[59] as their lives may be shaped by another or they may be adhering to another's

[59] Dworkin, *The Theory and Practice of Autonomy*.

agenda. In what follows I will consider the idea that it may not be morally problematic to nudge people of this type.

2.3.3.2 Why it might be good to nudge agents that are not reasons-responsive or autonomous

I have two arguments for why it might be less problematic to nudge agents of this type. First, an agent who does not act for reasons is relatively unlikely to achieve positive results through her actions. That is, if an agent is acting at random, or for no reason at all, or against reason, she is unlikely to get the things that are valuable to her through her actions. On the other hand, a choice architect that attempts to frame choices for the benefit of the chooser is framing the decision intentionally and for reasons. Such a choice architect ought to carefully considered which outcome is likely to benefit the chooser and will have carefully considered the methods by which she can encourage the chooser to take the appropriate action. If this is the case, it seems likely that the choice architect will produce much better outcomes than will the chooser.

A second argument is that there is no cost to such an action. Manipulations like nudges, are usually thought to be bad because they interference with autonomy. Recall Saghai's analysis of when healthcare nudges are wrong (if they are wrong). He argues for the importance of a self-determining life, which is a "life that is, in its main contours, free from the exercise of power by other individuals and by social and political institutions.[60]" However, the obvious question for this account is "free to do what?" and any reasonable answer would center on the freedom to engage in the activities of one's life as one sees fit. However, if an agent is not engaging in the

[60] Saghai, "Salvaging," 7.

activities of one's life as she sees fit, then a nudges would not seem to interfere with their capacity. Alternatively, we can recall Gorin's account according to which nudges are open-ended in the sense that they subject the chooser to the arbitrary will of the choice architect. But, this can only be problematic if subjecting the chooser to the will of the choice architect would interfere with the goals and projects of the chooser. But, agents that are not autonomous do not fall under this category. Therefore, this is not problematic.

2.3.3.3 An argument against nudging agents that are not autonomous or reasons-responsive

A few arguments can be made against this account. The most forceful is that intentionality matters. That is, it might be a different thing for an agent to fail to act for reasons than it is for a choice architect to intentionally manipulate the chooser towards particular actions. As an example, we know that the order that information is presented in can affect decisions. Imagine that I research retirement on the internet and I am presented with an argument against saving for retirement first, and then I am presented with an argument in favor of saving for retirement and, as a result, I elect not to save for retirement. Now imagine that someone wants to manipulate me by nudging me not to save for retirement. Accordingly, they present me with the arguments against saving before presenting the arguments in favor of saving and I decide not to save for retirement. Even though both cases have the same outcomes, the second case seems to be objectionable whereas the first case does not. So, manipulating people who are not autonomous or who are not acting for reason might be objectionable when it is done intentionally. Against this argument, it seems that intention matters as well. *Intending* to harm me by manipulating me to not save for retirement might be problematic because you intend to harm me and it is wrong to

intend to harm another. However, we can rule out nudges that are intended to harm another for reasons that will be discussed in the next section. So, it is unclear that the intuition persists in cases where one intends to help another through manipulation.

2.3.3.4 Concluding thoughts

In this section I argued that, if nudges are sometimes problematic because they interfere with the autonomy and reasons-responsiveness of agents, then it is less problematic to nudge agents who are currently not autonomous or reasons-responsive. Agents might have or fail to have autonomy or reasons-responsiveness in degrees and accordingly, it would be less problematic to nudge those who fail to have these traits to a high degree. Another practical upshot of this discussion is that it might not be problematic to nudge agents whose decisions are practically irrational. For example, imagine someone is attempting to choose an auto loan from among a number of providers. The terms on all of the loans are the same, but one has a lower interest rate and comes from an unknown provider. There can be no reasonable argument for choosing the loan with the higher interest rate and so, choosing it would be practically irrational. In such a situation, nudging choosers to choose the lower-cost loan would not be problematic.

2.3.4.1 A nudge is morally problematic if it is performed by the state

In this section I consider the claim that a nudge is morally problematic when it is performed by the state. Since nudges can be used in many contexts, showing that it is impermissible to use them in the state context does not provide an argument against nudges *in general*. However, much of the traction in using nudges has occurred in government. For example, there is the

Behavior Insights Team or "nudge unit" in the UK which works to implement nudges in the UK government. In the US, Cass Sunstein, one of the authors of the seminal work on the subject, Nudge, served as the head of the White House Office of Information and Regulatory Affairs in the Obama administration during which he worked with President Obama to implement nudges in the federal government. In addition, the US has considered creating its own Nudge Unit. If it can be shown that it is wrong for the state to implement nudges, then it would have far-reaching implications for the potential of the nudge tool. In the remainder of this section I discuss I use the concept of state neutrality about the good to offer some arguments against nudges that are performed by the state. I ultimately conclude that such arguments are not successful and that state nudges are acceptable.

2.3.4.2 Nudges by the state and the principle of state neutrality

One argument against the use of nudges by the state centers on the principle of state neutrality which holds that the state should be neutral among rival understandings of the good. Such a position can be problematic for the use of certain kinds of nudges because in implementing a nudge, the choice architect must remove herself from this neutrality, determine that a particular decision is best, and nudge choosers towards that decision. This principle might not rule out all nudges performed by the state, however. Many formulations of the principle of state neutrality would only prevent nudges that would nudge choosers on the basis of a controversial conception of the good or towards conceptions of the good that do not have societal consensus. Many of the paradigmatic nudges, like opt-in 401k plans, do not promote any such controversial position. The principle of state neutrality can be spelled out in three different ways:

1. The state should not promote the good, either coercively or non-coercively, unless those who are subject to the state's authority consent to its doing so.
2. The state should not aim to promote the good unless there is a societal consensus in support of its doing so.
3. The state should not justify what it does by appealing to conceptions of the good that are subject to reasonable disagreement.[61]

A full exploration of the neutrality principle is beyond the scope of this dissertation. Below I offer some brief remarks to motivate the claim that nudges by the state are acceptable.[62]

Arguments in favor of state neutrality can be divided into roughly three camps: concerns about autonomy, concerns about abuse of power, and concerns about knowing the good. Concerns about autonomy have already been discussed in section 2.2 of this dissertation, and concerns about knowing the good will be addressed later, so I want to focus on concerns about the abuse of power by the state. I will focus on this in the next section.

2.3.4.3 Nudges by the state and abuses of power

One kind of argument in favor of state neutrality is that the power of the state is so vast and those who are in control of that power are so flawed that we should not trust the government to promote the citizen's good even through nudges. Sher identifies three concerns that one might have about the government's use of power: oppression, instability and error. By oppression he means the abuse of the vast amount of power usually controlled by the state. By instability he means that if a state promotes a particular conception of the good it might cause the exacerbation

[61] Steven Wall, "Perfectionism in Moral and Political Philosophy," *Stanford Encyclopedia of Philosophy* (2012).
[62] For a detailed discussion of the principle of state neutrality, see George Sher's excellent work Beyond Neutrality: Perfectionism and Politics. Much of the discussion in this section borrows from his work.

of destabilizing influences and could cause the state to collapse. And, by error he means that we have no reason to think that the state will be better at ascertaining the good than the individual.

Sher argues that each of these are very real concerns, but that state neutrality is not the most promising way to address these issues. Instead, we are protected from oppression not by neutrality, but by *rights*. As Sher points out, our current system of government is nonneutral, yet:

> "Despite two centuries of nonneutral laws and policies, the long slide into authoritarianism simply has not occurred. Indeed, the long-range tendency seems, if anything, to be in the other direction. Our civil rights were significantly expanded by the Miranda decision, the exclusionary rule, and the series of court decisions that drastically extended the sphere of privacy … all things considered, the sphere of protected activities seems significantly larger now than fifty or a hundred years ago.[63]"

So, there does not seem to be a large reason to suspect that nonneutral governments will lead to the abuses of power.

One could respond to this argument in two ways. First, one could argue that nudges violate rights and so erode the defense we have against nonneutral governments. We might define a right as follows:

> Rights are entitlements (not) to perform certain actions, or (not) to be in certain states; or entitlements that others (not) perform certain actions or (not) be in certain states.[64]

One could imagine many plausible stories according to which the right to pursue one's life according to whatever conception of the good one has and free from substantial government interference is a right afforded to citizens. Nudges by a nonneutral state may violate this right

---

[63] George Sher, *Beyond Neutrality: Perfectionism and Politics* (Cambridge University Press, 1997) 114.
[64] Leif Wenar "Rights" *Stanford Encyclopedia of Philosophy* (2015) <http://plato.stanford.edu/entries/rights/>

and therefore weaken the use of rights to prevent the authoritarian slide that a nonneutral government may cause. Alternatively, one could argue that rights will not be an adequate defense from nonneutral governments in the future. The state's ability to non-argumentatively influence its citizens without their awareness is stronger now than it has been in the past. We could imagine that as the state's abilities grow it may be possible to descend into a government that is authoritarian in the sense that it uses mechanisms that encourage obedience to authority but does not violate rights in the sense that citizens are free to do otherwise (but are manipulated not to choose to do so).

One way to avoid these arguments is to accept a restricted principle of state neutrality. This would provide a principled defense against authoritarianism while allowing certain forms of nonneutrality by the state. Wall offers the following restricted neutrality principle (RNP):

> RNP: If two or more ideals of the good are eligible for those who live in a particular political society, and if these ideals have adherents in that political society, and if these ideals cannot be ranked by reason as better or worse than one another, then the state, to the extent that it aims to promote the good in this political society, should be neutral between these ideals in its support of them.[65]

This principle would likely allow most nudges that a choice architect might wish to employ while also avoiding the concerns about nonneutral states and authoritarianism.

2.3.4.4 Conclusion on nudges and the state

[65] Steven Wall "Neutralism for Perfectionists: The Case of Restricted State Neutrality," *Ethics*, 120:232-256.

In this section I considered arguments against the use of nudges by the state by focusing on the concept of state neutrality. I consider Sher's argument that nonnetural governments are not at any increased risk of abusing their power and some possible responses to Sher's argument. I then offer a reduced neutrality principle that would both allow nudges and avoid concerns about nonneutral state's becoming authoritarian.

2.4 Ends-based considerations

2.4.1 Introduction

In this section I review considerations relative to the ends that a nudge produces or that the choice architect intends to produce. As already discussed, nudges involve some cost insofar as the choice architect chooses to use nudges instead of rational persuasion. Accordingly, they must accomplish a positive end in order to justify this cost. However, many have argued that nudges cannot just aim for positive ends. For example, Thaler and Sunstein stress that nudges ought to aim for positive ends *for the chooser*, which excludes nudges that aim for positive ends for society. In what follows I review a number of possible claims that consider the aims of the nudge and will ultimately arrive at a set of considerations that I believe are relevant for the choice architect.

2.4.2. A nudge is only morally acceptable if it is for the benefit of the chooser

In this section I consider the idea that nudges are only acceptable if they are intended to benefit the chooser (as opposed to benefiting society as a whole). This claim derives primarily from Thaler and Sunstein's understanding of libertarian paternalism in Nudge. There, they argue "in

our understanding, a policy is "paternalistic" if it tries to influence choices in a way that will make choosers better off, *as judged by themselves*.[66]" Notice that the relevant criteria is that it makes the *chooser* better off as opposed to making society better off or producing positive consequences. Elsewhere they distinguish between libertarian paternalism and libertarian benevolence which they define as "an approach that attempts to promote benevolence, and to assist vulnerable people, without mandating behavior in any way.[67]"

It might be important in this section to distinguish between two kinds of scenarios that this claim might apply to. The first is a scenario where the nudge has little or no cost to the chooser, but can benefit others. An example of this would be organ donation. Whether one's organs are donated or not has no effect on the donor because the donor is dead.[68] Therefore, a nudge that increases the rate of organ donations is certainly not done for the benefit of the chooser, but it also is not done to their detriment. A second kind of scenario is one that harms the chooser in exchange for a larger benefit to others. For example, say I implement a nudge that encourages more rich westerners to donate large percentages of their money to the most effective nonprofits. This arguably does not benefit the one who donates, but produces a quite large benefit for others. It might be possible to argue in favor of the first kind of nudge, but argue against the second kind of nudge.

2.4.2.1 The extrinsic argument for why nudges must benefit the chooser

---

[66] Sunstein and Thaler, "Nudge: Improving Decisions," 4.
[67] Sunstein and Thaler, "Libertarian Paternalism is Not an Oxymoron," 25.
[68] One might hold the view that it is possible for a person to be harmed after they are dead and thus organ donations can harm a dead person. Since this is the minority view and not the issue under analysis, I will presume that post-mortem harms are not possible.

One type of argument for why nudges must benefit the chooser is that these nudges will be problematic in terms of their consequences. This argument can take two forms. The first is that nudges of this type might be seen as more intrusive or offensive by choosers which may cause them to reject the nudge concept or choice architects. Take the charity donation nudge mentioned above as an example. Imagine the IRS instituted a nudge that opted taxpayers in to a plan to donate tax refunds to nonprofits instead of returning them to the taxpayer. This would undoubtedly create a large amount of good in the world, but it would likely be received extremely unfavorably by taxpayers, especially those who did not realize they were donating their tax refunds to charity. It might, for example, cause taxpayers to distrust the IRS and the fairness with which they administer the tax system; it might cause choosers to have greater skepticism and distrust of the government in general. These are responses that taxpayers would be unlikely to have if a nudge benefited the chooser.

A second argument is a kind of slippery slope argument that the use of nudges that do not benefit the chooser may lead to abuses of the nudge mechanism. The idea here is that it is much easier for choice architects to rationalize abusive or harmful nudges if they can justify a nudge by claiming that it is for the benefit of others. Returning to the IRS nudge mentioned above, we might imagine that the IRS extended the nudge even further. Instead of just opting taxpayers into sending their tax refund to charity, we might imagine that the IRS automatically attempts to take taxes in excess of what taxpayers owe and donate that to charity as well. While this may make the world better off, it certainly cannot be justified by reference to the benefits it creates for the chooser. So, it might be the case that allowing nudges that do not benefit the chooser can lead to abuses of the mechanism.

This argument is far more plausible in cases where a nudge harms a chooser for the benefit of others than in cases where a nudge is neutral for a chooser but done for the benefit of others. In the case where the nudge is neutral for the chooser, there does not seem to be any obvious opening up of nudges for abuse or for possible slippery slope considerations. Nudges that are neutral for the chooser represent a fairly small range of potential options, whereas nudges that are harmful for the chooser represent a wide range of potential options. Opening up this small range of options provides all of the same clear guidelines for choosers; it just includes a slightly wider range of cases.

However, cases where the nudge harms choosers are harder to justify. The answer to this question rests in whether these nudges will in fact have negative consequences. This is difficult to determine but we can make some educated guesses. These kinds of nudges seem to violate common sense morality according to which we should not attempt to harm people for the sake of someone else. So, it does seem plausible that there could be negative consequences if nudges of this type were used. However, the particular nudge mechanism in question matters a great deal in these cases. For example, automatic opt-ins like the IRS case mentioned above are probably not acceptable because a large number of people will end up making a decision that they did not realize they had made, leading to a lot of frustration and anger. However, other kinds of nudge might be acceptable. For example, imagine the IRS creates a forced choice between either donating or not donating the money to charity, but they frame the charity donation in such a way that it makes donating to charity much more likely than it would have been otherwise. For example, the IRS might put a picture of the people you would be helping on the tax form or make the option to choose to donate more obvious. Because the people that choose to donate to

charity under these alternatives will likely understand their choice, this kind of nudges seems far less likely to result in backlash from choosers. The principle at work here is the idea of substantial noncontrol as introduced in section 2.2. Nudges that do not benefit the chooser but are extremely noncontrolling seem to be more palatable than those that do not benefit the chooser but are more controlling.

That said, if all we consider are the consequences of a nudge, in some cases we will probably be forced to endorse some cases of nudges that harm the chooser regardless of the particular nudge mechanism employer. Returning to the IRS charity donation scheme imagined above, the amount of good this scheme could do is quite enormous. Imagine the IRS used this scheme to cause people to donate to one of the most effective charities in the world, the Against Malaria Foundation (AMF). According to the charity evaluator GiveWell, for every $2,838 donated to AMF a child under 5 will not die that would have died from malaria.[69] In 2015 the average tax return in the US was around $2,800.[70] This means that, in expectation, for every person successfully nudged by the IRS, a child will no longer die. It is hard to imagine a consequences-based argument that could reasonably show that the loss of $2,800 to a US taxpayer is worth more than saving a child's life. Thus in some cases the consequences may be sufficiently strong that any kind of nudge is justified.

2.4.2.2. The intrinsic argument for why nudges must benefit the chooser

[69] GiveWell "Against Malaria Foundation (AMF)" <http://www.givewell.org/international/top-charities/AMF#Costperlifesaved> retrieved 12/12/2015
[70] Jeanne Sahadi "Average tax return tops $2,800" *CNN Money* April 9, 2015 <http://money.cnn.com/2015/04/09/pf/taxes/tax-refund/>

One way to argue that nudges that do not benefit the chooser are wrong in their own right regardless of the consequences is from a Kantian point of view. One could argue that nudging a person to do some activity that benefits others and not the person and treats the person as a means and not as an end, and therefore violates the humanity formulation of Kant's categorical imperative. The core of this argument is that it is wrong to directly and intentionally harm someone for the benefit of others. Many classic counterexamples against utilitarianism tend to lean on this intuition as a mechanism for arguing in favor of considerations beyond utility.

To see whether this is an objection against nudges that do not benefit the chooser, I will first need to unpack some of the details behind Kant's second formulation of the categorical imperative. First, it is important to note that this does not rule out using people as a means to our end as this would be a burdensome demand. We use people as a means to our end all the time when we pay for their labor or buy goods from them and so on. Instead, what we must avoid is treating someone *merely* as a means. In addition, the formulation does not apply to humanity but to the "humanity" in people. By humanity, Kant means the collection of features that make us distinctively human. This includes the capacities we need to engage in self-directed rational behavior. Finally, the idea of an end has three different meanings for Kant. The first meaning is "a thing we will to produce or bring about in the world."[71] If gaining money is my end, then money is a thing that I am aiming to produce. The second meaning is as something to:

> [R]ealize, cultivate or further by my actions. Becoming a philosopher, pianist or novelist might be my end in this sense. When my end is becoming a pianist, my actions do not, or at least not simply, produce something, being a pianist, but constitute or realize the

[71] Robert Johnson "Kant's moral philosphy" *Stanford Encyclopedia of Philosophy* (2008) <http://plato.stanford.edu/entries/kant-moral/#HumFor>

activity of being a pianist. Insofar as the Humanity in ourselves must be treated as an end in itself in this second *positive* sense, it must be cultivated developed or fully actualized.[72] The third sense of the term is as "something that limits what I may do in pursuit of my other ends, similar to the way that my end of self-preservation limits what I may do in pursuit of other ends.[73]"

In what sense might a nudge that does not benefit the chooser violate Kant's second formulation of the categorical imperative? On the first definition, the categorical imperative suggests that we must not aim to bring about a violation of one's rational capacity. A nudge that does not benefit the chooser probably does not violate this formulation because the nudge does not violate one's rational capacity. It may decrease the probability that one engages their rational capacity, but this is not the same thing as violating it. On the second definition we must cultivate one's rational capacity. Nudging when rational persuasion is an available option probably does violate the categorical imperative in the sense that it does not cultivate rational capacity when an option to do so is available. Finally, on the third definition we are limited to actions that do not decrease one's rational capacity. Again, it does not seem plausible that a nudge decreases one's rational capacity. Instead it only seem plausible that a nudge makes the exercision of rational capacity less likely.

So, on some ways of understanding Kant's second formulation of the categorical imperative it seems likely that nudges that do not benefit the chooser violate the categorical imperative. Furthermore, it seems plausible that all nudges violate this interpretation of the categorical imperative if rational persuasion was an available option. What does this say for

[72] Robert Johnson: Ibid.
[73] Robert Johnson: Ibid.

proponents of the nudge mechanism? I'm not sure. For a committed Kantian this may spell doom for the nudge mechanism. However, I do not think the arguments in favor of Kantianism are persuasive enough to commit oneself to following any implication of the view. Instead, I find a view which takes our normative uncertainty into account much more plausible. MacAskill has argued that decision makers should "maximize expected choice-worthiness, treating normative uncertainty analogously with how they treat empirical uncertainty.[74]" Although exploring normative uncertainty is beyond the scope of this dissertation, I think that an account that takes normative uncertainty into account is unlikely to rule out nudges in the general case and is likely to endorse some instances of nudges that harm the chooser for the benefit of others.

2.4.2.3 Concluding thoughts on nudges that do not benefit the chooser

In this section I considered the claim that nudges should always benefit the chooser. I considered two kinds of arguments for this claim. The first is an extrinsic harm argument that holds that nudges that do not benefit the chooser will lead to abuses of the nudge mechanism or will push the use of nudges down a slippery slope. The second is an intrinsic harm argument which says that nudges that do not benefit the chooser use the chooser as a means instead of an end, in violation of the second formulation of the categorical imperative. After considering these arguments, I conclude that some kinds of nudges that do not benefit the chooser may be acceptable on the basis of their consequences. However, some nudges of this type may not be acceptable on the second formulation of the categorical imperative. I conclude that committed Kantians may have a reason to reject the nudge mechanism entirely but that one a reasonable

[74] William MacAskill "Normative Uncertainty" Dissertation, Oxford University (2014)

interpretation of decisions under normative uncertainty nudges in general are likely to be permissible and some nudges that harm the chooser are also likely to be permissible.

2.4.3. Nudges must make chooser better off as judged or defined by themselves

In this section I consider the claim that nudges must make chooser better off as judged or defined by themselves. This claim is endorsed by Thaler and Sunstein in much of their writing on libertarian paternalism. Thaler and Sunstein distinguish between nudges and libertarian paternalism with libertarian paternalism serving the role of being a particular justificatory scheme for the use of nudges. On this interpretation, for something to be a nudge, it does not need to be normatively valuable, but for something to be libertarian paternalism it does. They lay out the distinction as follows:

> It's important to point out that nudging complements a libertarian paternalism outlook about public policy, but the two are distinct concepts. Libertarian paternalism is intended as a means to help people make decisions that make them better off as defined or judged by themselves – not by a government or private authority. While the nudges cited in the book are intended to do exactly this, nudging takes place in [a] variety of realms where the nudger's explicit goal is to promote [the nudger's] own welfare.[75]

So, on this interpretation, a nudge justified by libertarian paternalism must make the chooser better off as judged or defined by themselves. In this section I will consider this as a general principle to be applied to nudges and reject this principle.

[75] Balz, "A Nudge."

2.4.3.1 What does "as judged by themselves" mean?

In this section I try to tease apart the claim that nudges should benefit choosers "as judged by themselves" more clearly. I see three possible interpretations of this claim. The first is that a nudge is only good if the chooser actually reflects on the choice and judges it to be good. This seems to be the most natural interpretation of what it means for one to judge something as good. However, this interpretation is also problematic for a large number of paradigmatic nudges. On this interpretation, nudges must cause chooser to not only bring the choice to conscious attention, but to reflect on the choice and judge it positively. But, many nudges work in the background and do not engage conscious choice or reflection. Recall, for example, the definition of a nudge that I outlined in chapter 1:

> Nudge: A nudges B when A intentionally makes it more likely that B will φ, primarily triggered by B's shallow cognitive processes, while A's influence preserves B's choice-set.

Because nudges are generally defined as involving shallow cognitive processes, it would be highly problematic if some kind of conscious deliberation was needed to determine if a nudge was good. Therefore, I do not think this is a plausible interpretation and it is unlikely to be an interpretation intended by Thaler and Sunstein.

A second possible interpretation is that "as judged by themselves" does not require actual judgment of the nudge, but instead requires only hypothetical ascent to the outcome of the nudge. The idea on this interpretation is something like "the chooser would approve of the outcomes if the chooser considered them." Thus, choosers do not actually need to bring the nudge to conscious attention; it only needs to be true that *if* they brought it to conscious attention,

they would agree. The problem with this interpretation is that many nudges are designed to help choosers get what they *would* want if they understood the situation properly, which is not accounted for on this interpretation. As an example, imagine we wanted to implement a nudge to get choosers to invest in the lowest-cost index funds instead of investing in index funds that offer the same product at a higher fee. Many choosers pick higher-cost index funds because they are familiar with the brand name of the company offering them and they do not realize that they can get an identical product for cheaper. If I nudged such a chooser to pick a different fund, it seems highly likely that, *if* they were to reflect upon this choice, they might judge that they had picked wrongly. So, on this interpretation, I would have engaged in an impermissible nudge. Indeed, because nudges rely on shallow cognitive processes, they often do not change what choosers *judge* to be the correct choice, because such judgment is not a shallow cognitive process. Instead, they only change the likelihood that choosers select a particular option. So, on this interpretation, we are again at risk for ruling out a large number of paradigmatic nudges.

A final interpretation is that nudges are to be evaluated on an informed desire account of the good. Informed desire theories hold that what is valuable, and what the Choice Architect should aim for, is what the person *would* choose if they were properly informed. Recall Thaler and Sunstein's Humans and Econs example as discussed in chapter 1. Econs are perfect maximizers of their own wellbeing who "can think like Albert Einstein, store as much memory as Big Blue, and exercise the willpower of Mahatma Gandhi.[76]" Using this heuristic, Thaler and Sunstein define a nudge as anything that would not change the decision of an Econ, but might influence the decisions of humans like us. Informed desire theories hold, essentially, that the

[76] Sunstein and Thaler, "Nudge: Improving Decisions."

good is what an Econ would choose. Put another way, a nudge is good if it gets a person to choose what they would have chosen if they were an Econ. Informed desire accounts also share the distinction of being the predominant accounts in the literature. In Nudge, Sunstein and Thaler adopt the informed desire account explicitly[77] and the majority of authors seem to have followed suit. Indeed, the majority of the examples of nudges concern cases where actual choices appear to be inconsistent with what we imagine fully informed choices would be. Examples include unnecessary credit card debt, lack of 401(k) savings, payday loans, rent-to -own establishments, lottery tickets,[78] overeating, alcohol abuse, smoking[79] and failure to take necessary drugs.[80] Given the explicit endorsement of this view by Thaler and Sunstein and the common usage in the discussions of nudging, I think this is the most plausible interpretation of the claim.

In the next section I suggest a problem for the idea of informed desires as making choosers better off "as judged by themselves."

2.4.3.2. Wellbeing and the Informed Desire Satisfaction Account

While I think the informed desire account of wellbeing offers the most plausible interpretation of what Thaler and Sunstein mean by "as judged by themselves," I think it is a problematic way to ground that intuition for a variety of reasons. I outline those concerns here. I do not, however,

---

[77] Ibid.
[78] George Loewenstein and Emily Haisley, "The Economist as Therapist: Methodological Ramification of 'Light' Paternalism," in *Foundations of Positive and Normative Economics*, eds. Andrew Caplin and Andrew Schotter (2008).
[79] Katherine Flegal, Barry Graubard, David Williamson and Mitchell Gail, "Cause-Specific Excess Deaths Associated with Underweight, Overweight, and Obesity," *JAMA: Journal of the American Medical Association* 298 (2007), quoted in Loewenstein and Haisley, "The Economist."
[80] Cynthia Jackevicius, Muhammad Mamdani, and Jack Tu, "Adherence With Statin Therapy in Elderly Patients With and Without Acute Coronary Syndromes," *JAMA: Journal of the American Medical Association* 288 (2002), quoted in Loewenstein and Haisley, "The Economist."

discuss the philosophical virtues of the informed desire satisfaction account in this section instead I leave that discussion for chapter three.

One problem with thinking that a person's informed desire is what makes them better off "as judged by themselves" is that in some cases a person's informed desire can be diametrically opposed to a person's actual desires. An example will reveal the intuition.

Imagine a person, Frank, who is uninformed, lacks self-control and is relatively irrational. Frank has a set of actual desires to eat unhealthy food and refrain from exercise, preferring instead to watch TV and play video games. If Frank were an Econ with more knowledge, rationality and self-control he would instead desire to be physically fit. On the informed desire account we ought to nudge Frank towards a healthy lifestyle, so, being the good Choice Architects that we are, we begin to work to make Frank more physically fit. We make it more difficult for Frank to get to the McDonald's and easier to get to the gym. We make the salad menu more prominent and the hamburger menu harder to find. Let's imagine we are successful in this endeavor and as a result Frank grows healthier each day, but Frank does not develop any adaptive preferences and retains his original desires to be lazy and eat unhealthy food. We can imagine that Frank grows more miserable as his desire to sit around the house and eat Big Macs is consistently frustrated. No actual desire of Frank's is ever satisfied by living his new healthy lifestyle.

On the informed desire account what we have done for Frank is tremendously valuable because we have satisfied Frank's hypothetical desires. Yet, we have completely frustrated Frank's actual desires. It seems somewhat implausible to think that Frank is better off "as judged by himself."

Of course, this will be no surprise to the proponent of the informed desire account. On the informed desire account, what is good is what a fully informed version of Frank would judge to be good. But, fully informed Frank need not agree with actual Frank on what is good. Indeed, the informed desire account takes this to be a virtue of the account. The account attempts to accommodate two intuitions: *internalism* and *desire improvement*.

Internalism is the view that there should be a link between a person's desires and what is good for her. The basic intuition is that it is hard to believe that something can be good for a person if the person cares nothing about it. This is one weakness of objective list or perfectionist theories of the good. For example, Sher has developed a perfectionist account that includes understanding the world, the formation and execution of reason-based plans, relationships that involve companionship and mutual respect, developing one's abilities, becoming morally better, becoming more aware of beauty, developing decency or good taste, and privacy. We can imagine a person who sincerely and deeply has no desire to become aware of beauty. Sher would hold that despite this, becoming aware of beauty is good for him. While I'm sure Sher would have much to say on the topic at first look this seem implausible. Thus, many have found the internalism intuition appealing.

The second intuition is *desire improvement*. Desire improvement is the intuition that while what is good for a person should be linked to a person's desires, not just any desires should count. In particular, desires that would not survive certain kinds of improvement should count for less. One might think that this intuition can explained away by cases where the satisfaction of one desire leads to the frustration of other desires. For example, if I desire not to go to the dentist because I have a phobia of the dentist, my not going to the dentist will satisfy that desire, but it

will frustrate desires to have a healthy smile and not be in pain when eating. Yet, other cases do not seem to fall into this pattern. This is especially true when the pathway that I choose generates new desires that I would not want to fulfill from my present state. For example, I currently desire to achieve important things and make valuable contributions to society. I could imagine a set of actions that would make me desire things that are easier to obtain like watching TV and playing video games. This pathway might lead to a greater magnitude of satisfied desires but it conflicts with what I would want myself to want.

Unfortunately, the internalism and desire improvement intuitions sometimes come into conflict. In some of these cases, the informed desire account sides with the desire improvement intuition at the expense of the internalism intuition. But, this means that even on the informed desire account what is good for a person will often *not* make them better off "as judged by themselves" if by "themselves" we are referring to the person as current instantiated.[81]

The upshot of this discussion is that the heuristic of making choosers better off "as judged by themselves" probably does not cash out into any sophisticated, coherent philosophical view. It might be best to think of it instead as rhetorical flourish designed to help the nudge concept seem more palatable to those who might implement it. As such, it will be better to rely on a sophisticated theory of value to determine when choosers are better off. Chapter three will explore this issue more fully.

2.4.3.4 Concluding thoughts on making choosers better off "as judged by themselves"

[81] "As judged by themselves" could also refer to the potential future self who undergoes desire improvement. However, I think this would be a very unnatural interpretation of the phrase.

In this section I considered the claim that choice architects ought to strive to make choosers better off as judged by themselves. I considered three possible interpretations of what "as judged by themselves" might mean and ultimately decided that the most favorable interpretation was an informed desire account of the good. I then showed that the informed desire account does not really make choosers better off "as judged by themselves" because sometimes that our ideal selves would want comes into conflict with what we actually want. Accordingly, it will be better to focus on making choosers better off according to a more sophisticated theory of value.

2.5 Conclusion to chapter 2

In this chapter of the dissertation I hope to have advanced two basic claims. First, I hope to have shown that the question of the permissibility of nudges is far more complicated than a question of whether nudges are permissible or impermissible *in general*. Instead, the permissibility depends on a large number of complex situational factors that cannot be properly taken into account without looking at the specifics of the nudge in question. Second, I argue that nudges are a *pro tanto* moral wrong and I discuss a number of exacerbating and mitigating circumstances that determine if the harm of the nudge is outweighed by its benefit in specific cases. In the most general terms, I conclude that nudges that are especially manipulative, where the chooser does not or would not endorse the nudge and where there is a high degree of trust between chooser and choice architect have an increased threshold that they must overcome in order to be all-things-considered morally permissible. On the other hand, nudges where the chooser is not reasons-responsive, where the nudge benefits only the chooser and where the nudge is substantially noncontrolling have a lower threshold to moral permissibility. Ultimately, however,

the choice architect must weigh the benefits of the nudge against the harms to determine if the nudge ought to be enacted.

Chapter 3: What should we nudge people towards?

3.1 Introduction

In chapter 2 of this dissertation I attempted to accomplish two objectives. First, I attempted to show that the discussion of whether nudges are permissible or impermissible *in general* is too simplistic. Instead, the specifics of the nudge, especially the nudge's mechanism of action, the agents involved in the nudge, and the ends that the nudge aims at determine the permissibility of specific nudges. Second, I argued that nudges are *pro tanto* morally wrong, but that the wrongness can be outweighed by other considerations. In section 2.4 I outlined some ends-based considerations that have been raised in the literature as a means of narrowing the kinds of nudges that are morally permissible. In that section I argued that nudges do not need to only benefit the chooser and that nudges need not make the chooser better off as judged or defined by the chooser. I did not, however, discuss what nudges ought to be aiming for in general. In this chapter I will turn to that question.

In the most general sense, the idea of nudging is that choice architects ought to use environmental influences (e.g. framing, ordering, defaults etc.) to help make choosers better off. However, the question of what it means precisely to make a chooser better off is not a question that has received thorough investigation as it pertains to the nudging literature. Instead, many writers on this topic have implicitly assumed that a particular value theory is correct without investigating the reasons for or against this assumption. This is seen most strikingly with the Informed Desire account of value that was briefly discussed in section 2.4.3.

I review the existing literature on accounts of wellbeing. I follow Parfit's categorization of the existing views and divide it up into three alternatives[82]: hedonism, objective accounts and desire satisfaction views. I show that each existing account of wellbeing is objectionable for various reasons. However, the informed desire account of wellbeing can be rescued from objection by creating a new process for specifying how the idealization process should occur. As a result I conclude that choice architects should nudge choosers towards what the choosers would want if fully informed and rational.

One worry in writing this section is that readers may presume that goal here is to "solve" the problem of wellbeing. Those aware of the vast literature on the topic will realize that this project would at least take an entire dissertation and perhaps an entire lifetime of philosophical work. However, my goal is more modest. First, it is worth noting that because my question is a practical one facing real world decision makers, we cannot wait for the question of wellbeing to be solved before taking action. John Arras makes this point with regard to ethical theories:

> If all interpretive activity within the field were to depend upon the selection of a single, superior moral/political theory, practitioners hoping for assistance in dealing with real world clinical or policy problems would have to suffer a very long wait indeed. ("Be right with you, as soon as we resolve the fundamental disagreements between consequentialists and deontologists.")[83]

The goal instead is to identify which account of wellbeing is most plausible given the current state of the literature. With this in mind, I do not expect that proponents of accounts of wellbeing that I do not endorse will find themselves compelled to abandon their favorite account of

---

[82] Derek Parfit, *Reasons and Persons* (Oxford: Oxford University Press, 1984).
[83] John Arras, "Theory and Bioethics," *The Stanford Encyclopedia of Philosophy* (2010).

wellbeing, But, for those without a settled view, I hope that the considerations outlined here provide them with a reason to prefer the account that I outline.

3.2. Hedonism

In this section I will consider hedonistic accounts of wellbeing. Hedonistic accounts hold that what is good or valuable is what makes a person happy or satisfied, and thus it is happiness or satisfaction that the Choice Architect should aim to maximize. In this section I will show why I do not find hedonistic accounts of wellbeing plausible. I will begin by discussing the argument from false pleasure which includes Nozick's famous experience machine thought experiment. I will consider a number of possible rejoinders to this argument and I will show why each fails. I will then advance two additional arguments for rejecting hedonism to motivate the search for alternatives accounts of wellbeing.

3.2.1. Introduction and history of hedonistic accounts of wellbeing

In the discussion of hedonism we must distinguish between two possible claims one could hold. On one view, human action is always in pursuit of increasing pleasure and decreasing pain. This view is called *psychological hedonism* and it is a claim about human psychology. It will not be my concern in this section. Instead, we will consider *evaluative hedonism* according to which wellbeing consist in increasing pleasure and decreasing pain.

Hedonistic accounts have been discussed both by philosophers and by economists. On the economics side, much of the work has been on creating a notion of pleasure and pain which allows for measurement. There are a number of different accounts that fall roughly into this

category. For example, one approach that was popularized by Daniel Kahneman and embraced by a number of economists is the 'experience utility' account. On this account we should maximize "the happiness of an individual during a period of time as the sum of the momentary utilities over that time period; that is, the temporal integral of momentary utility.[84]" This account shares its notion of utility with standard hedonistic utilitarianism in which pleasure is the only good. Other accounts include Kahneman and Kruger's idea of the U-index which "measures the proportion of time an individual spends in an unpleasant state[85]" with the implication being that what is valuable is to spend as little time as possible in such an unpleasant state. A final account is the remembered utility account according to which what we ought to maximize is the quality of an experience as recorded in memory (usually measured by post-experience surveys).

Philosophical accounts have focused less on measuring hedonic wellbeing and more on issues concerning whether wellbeing is reducible to pleasure and pain. One of the earliest articulations of the view comes from the exchange between Socrates and Protagoras in *Protagoras*[86] and is also a view commonly ascribed to Epicurius.[87] The most well known modern articulation of the view comes from Bentham in *An Introduction to the Principles of Morals and Legislation*:

> Nature has placed mankind under the governance of two sovereign masters, pain, and pleasure. It is for them alone to point out what we ought to do, as well as to determine what we shall do.[88]

---

[84] Daniel Kahneman and Alan Krueger, "Developments in the Measurement of Subjective Well-Being," *Journal of Economic Perspectives* 20, no. 1 (2006): 5.
[85] Ibid., 19.
[86] Plato, *Protagoras*, ed. Gregory Crane (The Perseus Digital Library: 4th Century BCE).
[87] Epicurus, *Principal Doctrines*, trans. Robert Drew Hicks (1925).
[88] Jeremy Bentham, *An Introduction to the Principles of Morals and Legislation*, eds. J. Burns and H.L. A. Hart (Oxford: Clarendon Press, 1789).

Other contributors to the debate include Aristotle,[89] Aquinas, Butler,[90] Hume,[91] Mill,[92] Nietzsche,[93] Brentano,[94] Sidgwick,[95] Moore,[96] Ross,[97] Broad,[98] Ryle,[99] and Chisholm.[100] Hedonism has the advantage of being a view that seems quite plausible. It is intuitively appealing that what is good *for* me ought to also be good *to* me. To many authors, it does seem that pleasure is good and so, hedonism is an appealing account of wellbeing. In fact, the converse idea: that something could be good for me without being enjoyable is quite counterintuitive.

A full discussion of all of the various arguments for and against hedonistic accounts of wellbeing is beyond the scope of this dissertation. Instead, I shall focus on what I take to be the most common and forceful kind of argument against hedonism which I call the *Argument from false pleasures*. I turn to this argument in the next section.

3.2.2. Hedonism and the arguments from false pleasures

Arguments from false pleasures are intended to show that there are cases where one can gain pleasure from a thing, but that thing can be bad for them. This is a problem for hedonism because hedonism holds that one's welfare consists only in the maximization of pleasure. The argument from false pleasures is commonly thought of as a defeating argument for the view. Introductory

---

[89]Aristotle, *Nicomachean Ethics*, ed. Gregory Crane (The Perseus Digital Library: 4th Century BCE).
[90]Joseph Butler, *Fifteen Sermons Preached at the Rolls Chapel* (London: James and John Knapton, 1729).
[91]David Hume, *An Enquiry Concerning the Principles of Morals* (London: A. Millar, 1751).
[92]John Stuart Mill, *Utilitarianism* (London: Parker, Son and Bourn, 1863).
[93]Friedrich Nietzsche, "Twilight of the Idols," in *The Portable Nietzsche*, trans. Walter Kaufmann (New York: Viking Press, 1968).
[94] Franz Brentano, *Psychology From An Empirical Standpoint*, ed. Linda McAlister (London: Routledge and Kegan Paul, 1973); Brentano, *The Origin of Our Knowledge of Right and Wrong*, ed. Roderick Chisholm (London: Routledge and Kegan Paul, 1969).
[95]Henry Sidgwick, *The Methods of Ethics*, 7th ed., ed. John Rawls (Indianapolis: Hackett, 1982).
[96] G. E. Moore, *Principia Ethica* (Cambridge: Cambridge University Press, 1903).
[97] W. David Ross, *Foundations of Ethics* (Oxford: Clarendon Press, 1939).
[98] Charlie Dunbar Broad, *Five Types of Ethical Theory* (London: Routledge and Kegan Paul, 1930).
[99] Gilbert Ryle, *Dilemmas* (Cambridge: Cambridge University Press, 1954).
[100] Roderick Chisholm, *Brentano and Intrinsic Value* (Cambridge: Cambridge University Press, 1986).

texts to ethics frequently explain the hedonistic view, provide the argument from false pleasures, and then move on to more plausible alternatives (e.g. Bagani and Fosl;[101] Frankena;[102] Furrow;[103] Rachels;[104] Rosen[105]) and others have treated the argument as an outright rejection of hedonism (e.g. Furrow[106] and Griffin[107]). Turton[108] formalizes the argument as follows:

> P1. Hedonism about wellbeing states that all pleasure, and only pleasure, intrinsically contributes positively to wellbeing and that all pain, and only pain, intrinsically contributes negatively to wellbeing.
>
> P2. Pleasure based on truth, or something like it, contributes more positively to wellbeing than pleasure based on falsity.
>
> P3. Therefore, something other than pleasure (truth of some sort) must contribute positively to wellbeing.
>
> C. Therefore, hedonism about wellbeing is false.

The key premise in this argument is premise two. Premise one is a relatively uncontroversial statement of the definition of hedonism about wellbeing. Premises three and the conclusion both follow straightforwardly from premises one and two. Therefore, resisting the argument requires a rejection of premise two. In what follows I will outline three arguments that have been advanced in favor of premise two.

---

[101] Julian Baggini and Peter S. Fosl, *The Ethics Toolkit: A Compendium of Ethical Concepts and Methods* (Wiley-Blackwell, 2007).
[102] William K. Frankena, *Ethics*, 2nd ed. (New Jersey: Prentice-Hall, 1973).
[103] Dwight Furrow, *Ethics* (New York and London: Continuum, 2005).
[104] James Rachels, *The Elements of Moral Philosophy*, International ed. (McGraw-Hill, 2005).
[105] Bernard Rosen, *Ethical Theory: Strategies and Concepts* (Mayfield Publishing Company, 1993).
[106] Furrow, *Ethics*, 112.
[107] James Griffin, *Well-Being: Its Meaning, Measurement and Moral Importance* (Oxford: Clarendon Press, 1986), 10.
[108] Daniel Michael Turton, "Reviving Hedonism about Well-Being: Refuting the Argument from False Pleasures and Restricting the Relevance of Intuitive 'Evidence,'" (MA thesis, Victoria University of Wellington, 2008).

3.2.2.1. Humans, oysters and the philosophy of swine

The most common versions of hedonism about wellbeing are aggregative in the sense that the total amount of pleasure experienced determines the value of a state of affairs for a person. But, this view does not take into account the nature of the pleasure itself. In fact, it seems that some pleasures, especially those based in higher-order cognition are more valuable than others.

One way to reveal this intuition comes from the Socratic dialogue in *Philebus.*[109] Imagine you are faced with two options. One option is to live a very fulfilling human life. The second option is to live the life of a barely sentient oyster which experiences some very low-level of pleasure. Imagine also that as the oyster you can live as long as you like, whereas the human life will only live for eighty years. If what is good for a person is solely determined by aggregate pleasure, then at some length of oyster life, the life of the oyster is better. This seems implausible. Hedonism also seems to place all pleasures on par from animal pleasures like sex to cognitively complex pleasures like the acquisition of knowledge. This caused Thomas Carlyle to call hedonistic utilitarianism "the philosophy of swine."

This argument is not a defeating argument against hedonism for two reasons. First, a hedonist can simply bite the bullet and accept that some length of oyster life is better than 80 years of human life. Second, hedonists can add factors that determine the value of a pleasure for a person. The most famous example of this comes from Mill[110] who adds the property of *quality* to the judgment of pleasures. By using the quality property, Mill can hold that some pleasures are better by their very nature than others. So, one might choose the life of a human of 80 years over the life of an oyster of any number of years.

---

[109] Plato, *Philebus*, ed. Gregory Crane (The Perseus Digital Library: 4th Century BCE).
[110] Mill, *Utilitarianism*.

We are now in a position to see a problem in the original formulation of the view. Recall that the original formulation of the argument from false pleasures was expressed as follows:

P1. Hedonism about wellbeing states that all pleasure, and only pleasure, intrinsically contributes positively to wellbeing and that all pain, and only pain, intrinsically contributes negatively to wellbeing.

P2. Pleasure based on truth, or something like it, contributes more positively to wellbeing than pleasure based on falsity.

P3. Therefore, something other than pleasure (truth of some sort) must contribute positively to wellbeing.

C. Therefore, hedonism about wellbeing is false.

It seems plausible that P3 does not follow from P1 and P2 and therefore Turton's argument is not sound. The Millian addition of the quality of the pleasure allows us to retain P2 but not endorse P3. For example, it could be that the only thing that contributes to wellbeing is pleasure, but some facts about the pleasure determine how valuable the pleasure is. So, while it is the case that true pleasure contributes more to wellbeing than false pleasure, it is not true that something other than pleasure contributes to wellbeing.

However, I think a revision to the argument that focuses on the experience of pleasure may help to avoid this problem:

P1. Hedonism about wellbeing states that all experiences of pleasure, and only experiences of pleasure, intrinsically contributes positively to wellbeing and that all experiences of pain, and only experiences of pain, intrinsically contributes negatively to wellbeing.

P2. Pleasures that do not differ in their experience may differ in their contribution to wellbeing.

P3. Therefore, something other than the experience of pleasure must contribute positively to wellbeing.

C. Therefore, hedonism about wellbeing is false.

By "experience of" I am referring to what an experience is like from the inside. The idea is that some experiences that are the same form the inside may be different in their value and thus hedonism about wellbeing is false. This reformulation of the argument also allows us to continue to hold that Mill is a hedonist. Mill is claiming that certain kinds of pleasure experiences are more valuable than others (namely the pleasure experience of being a man is more valuable than the pleasure experience of being an oyster). This means that he can still accept P1 and that he can reject P2 (as, presumably, hedonists are inclined to do).

3.2.2.2. The deceived businessman

Mental state theories of wellbeing (of which hedonism is one) hold that a person's good consists in their attaining certain kinds of mental states. For example, hedonism holds that attaining the mental state of pleasure makes a person's life go better. But, this position leads to counterintuitive conclusions.

One way to reveal this intuition comes from a thought experiment by Shelly Kagan[111] called the deceived businessman. Kagan asks us to imagine a businessman who died thinking that he had achieved everything he wanted in life: a loving wife and children, a successful

[111] Shelly Kagan, *Normative Ethics* (Boulder and Oxford: Westview Press, 1998), 34-36.

business and the respect of his community. However, it turns out that the businessman was completely wrong in his assessment of his life. As it turns out, his wife was cheating, his children and the community were just using him for their own ends and his business partner had been stealing for his soon-to-be bankrupt business. Mental state accounts of wellbeing will hold that the businessman's life went just as well given that everything he thought was true was a lie as it would have gone if his perception of the situation was accurate. This is because the fact that he was deceived did not have an impact on his mental states. As Kagan puts it:

> In thinking about this man's life, it is difficult to believe that it is all a life could be, that this life has gone about as well as a life could go. Yet this seems to be the very conclusion mental state theories must reach!… So mental state theories must be wrong[112] ”

It seems that something beyond pleasure, namely that the things the businessman actually attains what he wants in his life, determines whether his life goes well. If so, premise two from above is true.

3.2.2.3. The experience machine

A final argument intended to show that premise two is true is Nozick's experience machine thought experiment. Recall that according to hedonism a person's wellbeing is increased to the degree that she experiences more pleasure and less pain. But, events can cause pleasure and pain without being *actual*. This leads to counterintuitive conclusions as Nozick's example demonstrates:

[112] Ibid., 35.

> Suppose there were an experience machine that would give you any experience you desired. Superduper neuropsychologists could stimulate your brain so that you would think and feel you were writing a great novel, or making a friend, or reading an interesting book. All the time you would be floating in a tank, with electrodes attached to your brain. Should you plug into this machine for life, preprogramming your life's experiences? If you are worried about missing out on desirable experiences, we can suppose that business enterprises have researched thoroughly the lives of many others. You can pick and choose from their large library or smorgasbord of such experiences for, say, the next two years. After two years have passed, you will have ten minutes or ten hours out of the tank, to select the experiences of your next two years. Of course, while in the tank, you won't know that you're there; you'll think it is all actually happening. . . . Would you plug in? What else can matter to us, other than how our lives feel from the inside? Nor should you refrain because of the few moments distress between the moment you've decided and the moment you're plugged in. What's a few moments of distress compared to a lifetime of bliss (if that's what you choose), and why feel any distress at all if your decision is the best one?[113]

The force of the example is that, despite the fact that the experience machine can produce any mental state you desire, plugging in to the experience machine is not good for you. Having the experience of writing a book is not the same as *actually writing the book*, having the experience of having friends is not the same as *actually having friends*. Because hedonism does not allow for these intuitions, hedonism is false.

---

[113] Robert Nozick, *Anarchy, State, and Utopia* (Oxford: Basil Blackwell, 1974), 44-45.

In the next sections I offer three possible responses to the argument from false pleasures and show that these rebuttals fail. I will focus in particular on rebuttals to Nozick's experience machine example as the literature on the example is well developed. However, the argumentation here will also apply to the deceived businessman example.

3.2.3. Biting the bullet on the argument from false pleasures

The simplest way to rest to the argument from false pleasures is to bite the bullet. That is, one could simply maintain that what is best for a person is to plug into the experience machine. Some proponents of this option go further and attempt to explain away the intuition that many have against plugging in. For example, it could be that our intuitive sense of the pleasure provided by the experience machine is mistaken. In general, we expect to get pleasure from interacting with real people and do not expect as much pleasure from interacting with fake or computer-simulated people. It could be that we are simply miscalculating the pleasure from the experience machine.

A related line of defense is to point out particular heuristics and biases that impact people's evaluation of the experience. One such bias is status quo bias which is a general preference for the current state of affairs. One can reverse the experience machine example to help reveal the presence of status quo bias. Imagine that a credible source, we can call him morpheus, tells you that you are in the experience machine right now. You have no idea what actual reality would be like. Imagine Morpheus gives you the option of having your memory of this conversation wiped and going to reality. Would you choose to do so? Presumably more people would say no to this offer than would accept the experience machine. This reveals a

preference for the status quo which indicates that the experience machine example does not demonstrate an intuitive preference for reality.

While these arguments might provide some solace for the committed hedonist, I do not take them particularly seriously as compelling arguments for non-hedonists or those who are undecided. The experience machine is so often cited as an argument against hedonism because, for the non-hedonist, the intuition that plugging into the machine is not good for a person is quite strong. Pointing to mitigating factors or attempting to otherwise reduce the intuitive cost of biting the bullet is insufficient to deal with this strong intuition. At best, arguments of this type can help those who are committed to hedonism for other reasons sleep a little easier at night.

3.2.4. The argument from false pleasure and preferentism

A different argument against the experience machine example is that it begs the question by assuming preferentism. Accordingly, only preferentists should accept the terms of the thought experiment and thus the argument is not an argument against hedonism.

To see why this is the case, we must first clarify exactly what the thought experiment is designed to accomplish. Baber explains the goals of the example as follows:

> The aim is not merely to establish that most people would choose reality over life in a fools' paradise. That is consumer research. It is not merely to determine what subjects "value" (prudentially) or what "matters" to them—what they believe, whether rightly or wrongly, is good for them. That is sociology. The purpose of the thought experiment is to elicit subjects informed, rationally considered preferences as revealed in their choices under epistemically favorable conditions because the assumption is that under these

> favorable conditions most subjects will get it right: the states which matter to them will be the states that really matter—those which in fact contribute to wellbeing.[114]

On Baber's argument, for the experience machine to demonstrate anything about wellbeing, something like the following claim must be assumed:

> P: If a reasonable and informed subject, i, would choose S over S', then S would contribute more to i's wellbeing than S'.[115]

But, this is not a claim held by the hedonist. The hedonist holds that what makes a state of affairs contribute to wellbeing is an increase in pleasure. The pleasure machine relies on a premise that the hedonist need not accept. Therefore, the experience machine does not refute hedonism.

I do not find Baber's argument or the rejection of the interference to P3 compelling. I think it misunderstands the role of the thought experiments in the philosophical process. The best way of thinking about the role of thought experiments is not to think that thought experiments are premises in a valid and sound logical argument. If this were the goal, thought experiment style arguments would be very weak indeed. The goal of a thought experiment is usually to tease out some implication of a view and show that the implication is counterintuitive. If thought experiments were intended to be premises in logical arguments for the falsity of a view, then they verge on question begging. If hedonism entails that pleasure is the only intrinsic good, it simply makes no sense to try to refute hedonism by simply pointing to some example and claiming that pleasure is not the only intrinsic good in the example. The hedonist can simply deny this interpretation of the thought experiment.

---

[114] Harriet Baber, "The Experience Machine Deconstructed," *Philosophy in the Contemporary World* 15, no. 1 (2008): 133.
[115] Ibid.

A different interpretation of the role of thought experiments in contemporary philosophy comes from the idea of Reflective Equilibrium. Reflective Equilibrium assumes a coherentist epistemology in which the goal is to establish a set of beliefs that cohere with one another. There are two versions of Reflective Equilibrium – wide and narrow. In narrow Reflective Equilibrium one attempts to generate ethical principles from the consideration of moral cases without questioning the veracity or utility of the moral intuitions. Alternatively, in wide Reflective Equilibrium one does not take the moral intuitions as immutable data but instead attempts to find a balance or coherence between our intuitions about specific ethical cases, our beliefs about the reliability of those intuitions, the principles that we think govern these intuitions and any theoretical limitations on accepting these considered intuitions or principles.

On this framework, thought experiment cases are designed to make it more difficult for one to achieve coherence while holding the particular philosophical view. The experience machine does not refute hedonism. Instead, it points to a large intuitive cost that one must pay in exchange for holding the view. That is, one must think that certain states of affairs are good when this coincides with our intuitions. If the view has other benefits and virtues, then it would be reasonable to endorse hedonism despite the experience machine. However, for many writers, the virtues of hedonism are insufficient to rise to the challenge posed by the experience machine. Therefore, I do think that experience machine is an argument against hedonism when the argument is placed in the proper epistemic framework.

3.2.5. Acting against our wellbeing

At the outset of the chapter we distinguished between two claims that a hedonist could be making. On one claim, the hedonist position is that human action is always in pursuit of increasing pleasure and decreasing pain. This view is called *psychological hedonism.* The other claim is that wellbeing consists in increasing pleasure and decreasing pain. This view is called *evaluative hedonism*. One possible response to the experience machine is that the example refutes psychological hedonism but not evaluative hedonism. Since psychological hedonism is an empirical and not normative claim, one could argue that the experience machine does not refute hedonic accounts of wellbeing.

To see why this might be the case, consider the following example. Imagine that the superduper neuropsychologist instead offers you the following choice. You can leave your current life and everyone you know and be transported to an extremely blissful life somewhere else. You will be supremely happy and never experience any pain. Would you say yes? Notice that here the world offered by the superduper neuropsychologist is just as real as your current world, so the usual concerns about the experience machine will not work. Yet, I suspect that most people will not choose this new life.

This might be the case because happiness matters so little in our understanding of what makes our lives go well that opportunities to increase it do not represent an increase in wellbeing. I find this doubtful. An alternative explanation is that we are sometimes willing to make choices that do not promote our wellbeing. It could be that we fail to choose to increase our wellbeing in this case because we feel a sense of duty to our family and friends. Leaving them for this new world would fail to honor that duty. Thus, it can both be true that our lives

would go better if we chose the new life, and that most people would not chose the new life. This means that psychological hedonism is false,[116] but evaluative hedonism can still be true.

Unfortunately, I think this argument fails to explain some of our intuitions about the experience machine. For example, if the concern is reneging on a duty to family and friends, we could imagine that they are imported into the experience machine as well and that the experience machine allows Matrix-style interaction with them. Yet still, it does not seem clear I would wish to plug in. Indeed, the most basic intuition is that, even ignoring concerns outside of wellbeing, it simply does not seem good for me to plug into the experience machine. There might be a set of related, but irrelevant intuitions that strengthen this central idea, but the core intuition remains. Therefore, I think the experience machine remains a powerful argument against hedonism.

3.2.6. Three additional arguments against hedonism

I think the argument from false pleasures provides a power reason to reject hedonism. Hedonists have a number of promising lines of rebuttal against this argument, but I do not think any of these option ultimately succeed. To strengthen the case against hedonism, in this section I offer three additional considerations against the view. I do not necessarily intend these arguments to be persuasive to the committed hedonist, but they provide additional reasons to look for more attractive alternatives.

3.2.6.1. Internalism and externalism

Recall that the reformulated version of the argument from false pleasures is as follows:

[116] That is, unless there is a compelling explanation as to why people *think* the new life would be worse hedonically than their current life.

P1. Hedonism about wellbeing states that all experiences of pleasure, and only experiences of pleasure, intrinsically contributes positively to wellbeing and that all experiences of pain, and only experiences of pain, intrinsically contributes negatively to wellbeing.

P2. Pleasures that do not differ in their experience may differ in their contribution to wellbeing.

P3. Therefore, something other than the experience of pleasure (truth of some sort) must contribute positively to wellbeing.

C. Therefore, hedonism about wellbeing is false.

Hedonists will avoid this argument by rejecting P2 and instead insist that if two experiences are the same "from the inside" then they contribute the same to wellbeing. This view is called *internalism* in the philosophy of mind and is distinguished from *externalism.* The difference can be summarized as follows:

> *Externalism* with regard to mental content says that in order to have certain types of intentional mental states (e.g. beliefs), it is necessary to be related to the environment in the right way. *Internalism* (or *individualism*) denies this, and it affirms that having those intentional mental states depends solely on our intrinsic properties.[117]

According to externalism, the content of mental states can be determined by factors external to that mental state. The classic argument for this claim comes from Putnam who asks us to imagine that somewhere there is a place that is entirely identical to Earth called Twin Earth with

---

[117] Lau, Joe and Deutsch, Max, "Externalism About Mental Content", *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), Edward N. Zalta (ed.), <http://plato.stanford.edu/archives/sum2014/entries/content-externalism/>

one exception. On Twin Earth water does not consist of H20 instead it consists of XYZ although all of the macro properties are the same. Putnam also asks us to imagine that it is the year 17550 so no experts on either planet know the chemical structure of water. Putnam argues that when Earthlings on Twin Earth use the term water they are not referring to the same thing as Earthlings on Earth even though the content of their mental states are the same. Thus, the content of one's mental states can be determined by factors other than what those states are like 'from the inside'

If the content of a mental state can be shaped by external factors in this way, why not think that the value of a mental state could be similarly affected by facts about it that I am unaware of?

There is a vibrant debate about internalism and externalism in the philosophical literature and I do not intend to settle that debate here. However, it is worth noting that one may reject hedonism on these grounds.

3.2.6.2. Posthumous harm

Hedonism holds that a state of affairs is good for someone if it entails an increase in pleasure and bad for someone if it entails an increase in pain. However, this claim leads to paradoxes concerning death. Hedonism must claim that no event can impact a person's wellbeing after they are dead because there is no person to experience the harm. But, some posthumous events do seem to make a person's life go better. An example will reveal the intuitions:

The Achievement: Suppose I want to conduct research that will lead to a cure for Lou Gehrig's disease, ALS. Suppose, too, that I gain great pleasure from this pursuit and from

> anticipating the good that my research will accomplish. Unfortunately, I will die before I achieve what I want, but I will still succeed if various events occur, and fail if some other events occur, after I am dead. For example, I will succeed if my research gives another scientist a critical clue which she develops into a cure that she otherwise would not have found. And I will fail if all of the records of my research are destroyed in a fire before they prompt another scientist to devise a cure. Upon reflection, I dread the prospect of the fire destroying my files even though I will be dead at the time it would occur; I judge that it would be against my interests. By contrast I welcome the prospect of my research inspiring a colleague; I judge that it would be in my interests.[118]

It seems extremely plausible that my life will go better if I succeed in my goal of contributing to a cure for ALS and that my life will go worse if my research is destroyed. Yet, because I am dead, there is no one around to experience any pleasure from the success of my research. Hedonism holds that posthumous events have no effect on wellbeing. This is implausible.

The Achievement fits into a general class of paradoxes concerning the harm of death. The most well known general formulation of this paradox is attributed to Epicurus:

> Death, therefore, the most awful of evils, is nothing to us, seeing that, when we are, death is not come, and, when death is come, we are not.[119]

The paradox then, is how death can harm us when there is no longer an *us* for death to harm. A full discussion of the philosophy of death is beyond the scope of this dissertation, but I will

---

[118] Steven Luper, "Retroactive Harms and Wrongs," (2013), quoted in Ben Bradley, Fred Feldman and Jens Johannson, eds., *The Oxford Handbook of Philosophy of Death* (2013).
[119] Epicurus, *Principal Doctrines*, quoted in Jason Saunders, ed., *Greek and Roman Philosophy after Aristotle* (New York: Free Press, 1997).

briefly discuss a plausible answer to this paradox and show why hedonism is unable to take advantage of this answer.

The most common way to assess whether a state of affairs conflicts with our interests is *comparativism.*[120] According to comparativism, something is good for us if it makes our life go better. Luper states the view more precisely:

> Comparativism directs us to assess an event (such as an ear removal) by comparing how well life goes for us, partly as a result of that event, to how well life would have gone had the event not occurred. We are to compare how well life goes for us in the actual world, where the event occurs, to how well it goes in the nearest possible world in which the event fails to occur. In turn, how well life goes, our welfare level, is assessed in terms of the intrinsic goods and evils we possess in a world, assuming we are capable of having any. Intrinsic goods boost our welfare level in a world; intrinsic evils bring it down. If the former outweigh the latter, we are well off in that world: our welfare level in that world is positive. If the latter outweigh the former, we have a negative welfare level there.[121]

Imagine we want to determine how bad it is for me to have stubbed my toe. Imagine also that we adopt comparativism in conjunction with hedonism as our theory of wellbeing. On this account, we take the amount of pleasure that exists in the world where I stub my toe and subtract the amount of pleasure that exists in the nearest possible world where I do not stub my toe. The result determines how bad it is for me to have stubbed my toe.

However, adopting comparativism in conjunction with hedonism does not work for cases involving death. Imagine a child dies at age 4. To determine how bad it is for the child to have

---

[120] Steven Luper, "Death," *The Stanford Encyclopedia of Philosophy* (2014).
[121] Steven Luper, "Exhausting Life," *The Journal of Ethics* (2010): 1.

died, we have to compare the value of the child's current life to the value of the nearest possible world where the child does not die. But, since the child is dead, the first part of the comparison refers to nothing. So, there is no way to determine how much pleasure exists for the child in the actual world. However, this is not the case on other theories of wellbeing. As Luper puts it "However, posthumous events might well be bad for us on other [non-hedonist] accounts of welfare. Suppose that I want to be remembered after I die. Given preferentialism, something could happen after I die that might be bad for me, namely my being forgotten, because it thwarts my desire.[122]

Of course there is much more to say about the philosophy of death as it related to hedonism. The point that I wish to make here is that hedonism has counterintuitive implications when it comes to death and that hedonism cannot avail itself of many of the solutions available to other accounts of wellbeing. For this reason, I think that other accounts of wellbeing are more plausible.

3.2.6.3. Is pleasure the only thing that is good for me?

Finally, I do not endorse hedonism because the basic intuition -- that pleasure is the only thing intrinsically good for me and pain the only thing intrinsically bad for me -- does not seem plausible. There are a number of candidates for things that can be good for me independent of their impact on my pleasure, but I will discuss only a few candidates here.

To my mind, the most plausible candidate is achievement. One plausible view of achievement proposed by Gwen Bradford[123] involves three components. First, achievements

[122] Luper, "Death."
[123] Gwen Bradford, "The Value of Achievement," *Pacific Philosophical Quarterly* 94, no. 2 (2013): 204-224.

follow a process-product structure, meaning that they involve a process which then results in the creation of a product. Second, achievements are difficult. If something is trivially easy, it is not an achievement. Finally, achievements involve competent causation. The product must be properly credible to the efforts of the agent.

A salient example of an achievement is attaining a PhD. Attaining a PhD involves a process and a product, it is difficult and it involves competent causation. It seems to me that my life will go better in a world where I complete my dissertation and attain my PhD then it will in the closest possible world where I do not attain the PhD. The hedonist will claim that the value of this achievement is reducible to pleasure, but this seems implausible. If the value of the PhD was reducible to pleasure, then it would seem to me that pursuing a PhD does not make a person's life go better. The process of completing a dissertation is not pleasurable. Sure, sometimes acquiring new knowledge can be enjoyable in a certain sort of way, but on the whole the process is well known for being isolating, agonizing and frustrating. Some pleasure will come from celebrating receiving the actual diploma or from reflecting with pride on the fact that the PhD has been achieved, but this seems to miss the point. Pursuing a PhD seems to be valuable because it is a significant accomplishment. It is valuable even if I do not gain pleasure from the experience.

Therefore, it does not seem clear that all the things that make my life go better do so because of an increase in my pleasure. Achievement is not the only example of this. Other salient examples include friendship, aesthetic appreciation, developing one's abilities, and moral improvement. Attaining many of these qualities will be associated with an increase in pleasure, but this does not seem to be why they are valuable.

3.2.7. Conclusion on hedonism

In this section I have argued that hedonism is not a compelling account of what makes a person's life go better. The primary reason to accept this conclusion is the argument from false pleasures which includes the famous experience machine thought experiment. I considered a number of rejoinders to the argument from false pleasures, but ultimately concluded that none of these rejoinders are successful. I then considered two additional reasons to reject the hedonistic account. The first is the problem of posthumous harm and the paradox of death, In that section I argued that hedonism cannot account for events harming a person after they are dead and cannot utilize comparativism to determine how bad (or good) it is for a person to have died. This leads to a number of paradoxes for the view. Finally, I argued against the core hedonistic intuition by pointing to some states of affairs that appeal to be valuable independent of their impact on pleasure.

The goal of this section of the dissertation has not been to provide novel arguments that will convince the hedonist to abandon their position. Instead my goal has been to show why one might find hedonism unappealing and to motivate the search for compelling alternatives for those who are not already committed hedonists. I turn now to alternatives to hedonism.

3.3. Objective accounts of wellbeing

In this section I will discuss two accounts of wellbeing that might fall under the category of objective accounts of wellbeing: objective list accounts and perfectionist accounts. Objective accounts are treated as distinct from subjective accounts which are typically thought to include

hedonistic and desire satisfaction accounts of wellbeing.[124] One way to parse the distinction between objective and subjective accounts is provided by Haybron: "Subjectivism about well-being ... tells us that what ultimately benefits a person is determined by subjective psychological states like desires or pleasures.[125]" Thus, an objective account of wellbeing would need to say that at least some of the things that ultimately benefit a person are not determined by subjective psychological states like desires or pleasures.

Yet, this distinction is controversial. Others have wanted to drawn the distinction differently:

> I would prefer to let the contrast between objective and subjective mark the contrast between (1) views which hold that claims about what is good can be correct or incorrect and that the correctness of a claim about a person's good is determined independently of that person's volition, attitudes, and opinions, and (2) views which deny this.[126]

On this distinction hedonism may be an objective account of wellbeing. This is because on hedonism pleasure is good not matter one's attitudes towards the pleasure.

I do not intend to solve the question of how to draw the division between objective and subjective accounts of wellbeing here. For an excellent review of this question, see Bradley (2014).[127] For the purposes of this section of the dissertation, I find it useful to groups objective list and perfectionist accounts of the good together. This follows the three-fold taxonomy of

---

[124] Ben Bradley, "Objective Accounts of Well-Being" quoted in Ben Eggleston and Dale Miller, eds., *The Cambridge Companion to Utilitarianism* (Cambridge: Cambridge University Press, 2014).
[125] Daniel Haybron, *The Pursuit of Unhappiness: The Elusive Psychology of Well-Being* (Oxford University Press, 2010), 13.
[126] Richard Arneson, "Human Flourishing versus Desire Satisfaction," *Social Philosophy and Policy* 16, no. 1 (1999): 115.
[127] Bradley, "Objective Accounts."

accounts of wellbeing that Parfit proposed in Reasons and Persons[128] which has become relatively standard in discussions of wellbeing.[129]

3.3.1. Background and history of objective accounts of wellbeing

Objective accounts of wellbeing generally hold that wellbeing consists in the attainment of a list of goods which increase a person's wellbeing non-instrumentally. One example of such a list is provided by Sher[130] and it includes: understanding the world, the formation and execution of reason-based plans, relationships that involve companionship and mutual respect, developing one's abilities, becoming morally better, becoming more aware of beauty, developing decency or good taste, and privacy. Nussbaum[131] elaborates on a list by Sen[132] which includes bodily health, bodily integration, imagination, thought, emotions, practical reason, affiliation with animals, play, and control over one's environment. The objective account has a philosophical pedigree that includes Aristotle, Aquinas, Spinoza, Hegel, and Nietzsche and includes more contemporary discussions[133] in Parfit 1984;[134] Scanlon 1998;[135] Hooker 1998;[136] Griffin 1986;[137] Griffin 1996;[138] Griffin 2000;[139] Arneson 1999;[140] and Moore 2000[141] among others.

---

[128] Parfit, *Reasons and Persons*, 493.
[129] Roger Crisp, "Well-Being," *Stanford Encyclopedia of Philosophy* (2013).
[130] Sher, *Beyond Neutrality*. It is worth noting that Sher is a perfectionist and intends this list to be a list of the perfections.
[131] Martha Nussbaum, *Women and Human Development: The Capabilities Approach* (Cambridge: Cambridge University Press, 2000).
[132] Amartya Sen, *Commodities and Capabilities* (Amsterdam: North-Holland, 1985).
[133] All of the following citations are found in Christopher Rice, "Defending the Objective List Theory of Well-Being," *Ratio* 26, no. 2 (2013): 196.
[134] Parfit, *Reasons*.
[135] T. M. Scanlon, *What We Owe to Each Other* (Cambridge: Belknap Press of Harvard University, 1998).
[136] Brad Hooker, "Does Moral Virtue Constitute a Benefit to the Agent?," in *How should one live?: Essays on the Virtues*, ed. Roger Crisp (New York: Oxford University Press, 1998).
[137] Griffin, *Well-Being*.
[138] James Griffin, *Value Judgment: Improving our Ethical Beliefs* (New York: Oxford University Press, 1996).
[139] James Griffin, "Replies," in *Well-Being and Morality: Essays in Honour of James Griffin*, eds. Roger Crisp and Brad Hooker (New York: Oxford University Press, 2000).

One reason that many have found objective accounts of wellbeing plausible is that they seem to do a good job of fitting with people's considered judgment about wellbeing. As Rice puts it:

> Many people judge that certain states of affairs contribute to well-being on account of their objective features, and not because people hold positive reactive attitudes toward them. Loving relationships, for example, are judged to be good for people because they involve reciprocal love among them. Similarly, meaningful knowledge is judged to be good for people because it involves appropriately justified beliefs about meaningful truths.[142]

This observation can help explain intuitive reaction to cases like Nozick's experience machine.[143] The experience machine makes a person's life go better only if one's mental states are the sole determinants of wellbeing. Objective accounts offer a view which implies that the actual attainment of the state of affairs is valuable not merely the mental states. In addition, many objective accounts of wellbeing are mutable in the face of criticism. If it turns out that some state of affairs is valuable that is not on the list, the proponent of the objective account can simply add the new item to the list.

One might ask how the items on the list are generated. Why are some items on the list and not others? This is an area where objective list and perfectionist accounts diverge. On the objective list account, "this question cannot, in principle, be answered ... [t]here's just a list;

---

[140] Arneson, "Human Flourishing."
[141] Andrew Moore, "Objective Human Goods" in *Well-Being*, eds. Crisp and Hooker (2000).
[142] Rice, "Defending."
[143] Nozick, *Anarchy*.

that's the end of the story.[144] I will say more about this feature of the objective list theory later in the chapter.

On the perfectionist account, there is a deeper story to be told. According to perfectionism, wellbeing ultimately consists in perfecting human nature. If friendship, gaining knowledge and appreciating beauty are part of human nature then these are on the list. One theoretically sophisticated articulation of the relationship between perfecting human nature and the items on the list of goods comes from Sher.[145] Sher characterizes his perfectionist theory as being pluralistic in that it holds many different things to be of value and worth trying to maximize, but monistic in that the source of the value of these things is the fact that they are all linked to human goals that are fundamental, meaning that virtually all humans possess and cannot avoid trying to achieve them (they are near universal and near inescapable). Accounts like Sher's can avoid the accusation that the construction of the list of goods is worryingly ad hoc.

Indeed, many have found the perfectionist idea attractive. Hurka explains its appeal as follows:

> The goal of developing human nature of exercising essential human powers is deeply attractive. This is reflected in its widespread acceptance. The ideal is implicit in non-philosophical talk of living a "fully human" or "truly human" life and is endorsed by diverse philosophers.... Some value contemplation; others value action. Some value a communal life; others value a life of solitude. If, despite these differences, these philosophers all ground their particular values in a single ideal of human nature, that ideal must have intrinsic appeal.[146]

---

[144] Bradley, "Objective Accounts."
[145] Sher, *Beyond Neutrality*.
[146] Thomas Hurka, *Perfectionism* (Oxford: Oxford University Press, 1993) 32.

Thus, objective list and perfectionist accounts have much to recommend them. In the next section I will discuss some objections to these views and possible responses on the part of proponents.

3.3.2. Cheap Thrills

Another problem for objective accounts is that they may fail to appropriately value goods like pleasure or desire satisfaction in some cases. One such case is a case of cheap thrills.[147] This is a case of pleasure or desire satisfaction with no perfecting aspect. As Arneston puts it:

> Cheap thrills are pleasures with no redeeming social value beyond their pleasantness. The world being as it is, and human nature being what it is, such pleasures seem to me to be important sources of enjoyment that significantly enhance many people's lives in ways for which there is no practical substitute.[148]

Arneston intends cheap thrills to be an objection only to perfectionism and not to the objective list theory. This is because the objective list theory can include goods that are not related to becoming a more perfect human specimen whereas the perfectionist theory cannot. For example, one could accomodate the value of cheap thrills by adding pleasure to one's list of intrinsic goods.

Some perfectionists do attempt to accommodate pleasure in their discussion of the good. For example, Sher says "[w]e can hardly deny that happiness, pleasure, and enjoyment are among life's goods, so any satisfactory unifying theory" must include them.[149] The "affective

---

[147] Arneson, "Human Flourishing."
[148] Ibid.
[149] Sher, *Beyond Neutrality*, 229, quoted in Bradford, "The Value."

capacities" are explicitly included by Kraut[150] and the enjoyable exercise of capacities is included by Kauppinen.[151]

So, many accounts of perfectionism can take into account the value of cheap thrills. Those version of perfectionism that do not allow for the value of pleasure or those version of objective list accounts that do not include pleasure on the list of intrinsic goods will have difficulties with this objection. I do not find this troubling because I do not find accounts of the good that do not include the value of pleasure to be compelling.

3.3.3. Alienation

Another objection to objective accounts of wellbeing is that they will sometimes claim that something is good for a person even if the person has no interest in the thing. Returning to Sher's list, appreciation of beauty is an intrinsic good. It is easy to imagine a person, let's call him Tom, who simply has not desire to appreciate beauty. This person is perfectly happy to only appreciate items for their functional use, does not desire to learn more about aesthetic appreciation and would gain no pleasure from viewing art. On Sher's account we must say that appreciating beauty would make Tom's life go better. We might imagine signing Tom up for an art class where Tom learns more about aesthetic appreciation. Such a class would increase Tom's wellbeing even if he was miserable the entire time. Railton expresses this concern as follows:

> It does seem to me to capture an important feature of the concept of intrinsic value to say that what is intrinsically valuable for a person must have a connection with what he would find in some degree compelling or attractive, at least if he were rational and aware.

---

[150] Richard Kraut, *What is Good and Why* (Cambridge: Harvard University Press, 2007): 137, quoted in Bradford, "The Value."

[151] Antti Kauppinen, "Working Hard and Kicking Back: The Case for Diachronic Perfectionism," *Journal of Ethics and Social Philosophy*, (2007): 1-9, quoted in Bradford, "The Value."

> It would be an intolerably alienated conception of someone's good to imagine that it might fail in any such way to engage him.[152]

We might wonder why such an objection should be convincing. In some sense the objection merely points out that objective views of wellbeing are not subjective. After all, *of course* objective views of wellbeing sometimes propose that things are good for a person even if that person is not motivated by them; that is what an objective account of wellbeing *is*. So, the alienation objection is not likely to be persuasive to the objectivist. Yet, I think there is something to the objection. It does seem plausible that what is good for me ought to be related, in some way, *to me*, to what I want, to what I enjoy. As with the "cheap thrills" objection, I think this view mostly suggests that the most plausible version of objectivism will include the value of pleasure or of desire satisfaction as a component of the account. Those that leave it out entirely not very plausible.

3.3.4. Tradeoffs between values

Another concern is purely practical. For our purposes the goal of a theory of value is to help determine whether a particular nudge improves the chooser's wellbeing. To do this, the account of wellbeing needs to provide a framework according to which nudges can be evaluated. This is relatively straightforward on other accounts of wellbeing. In the case of hedonism, one simply calculates whether the nudge would increase or decrease the amount of pleasure experienced by the chooser. On the desire satisfaction view one simply calculates the total satisfaction of desires as compared to the total frustration of desires. The actual calculations may turn out to be

[152] Peter Railton, "Facts and Values," *Philosophical Topics* 14, no. 2 (1986): 9.

difficult, but the basic framework for using the calculations is straightforward. However, objective theories of value offer multiple different intrinsic goods and they offer no framework according to which the different values can trade off. This means that determining if a nudge improves wellbeing for a subject might turn out to be mysterious.

Take Nussbaum's list as an example. It seems extremely plausible that we might find ourselves in situations in which those values are at tension. For example, the mere decision to place one in a hospital involves a clear conflict of values. On the one hand, it is likely to promote improved bodily health, but on the other hand it decreases the control one has over one's environment. How do we decide for a particular case of hospitalization whether it improves the subject's wellbeing? Nussbaum's framework offers no clues on this point. Similar problems arise for Sher's account. The very use of a nudge as opposed to engaging in rational persuasion appears to tradeoff between Sher's values of executing reason-based plans with the value implicit in whatever the nudge is designed to increase. How do we say of a particular nudge whether it tradeoffs between these values in a favorable way? It is mysterious.

To compound the problem, it is not simply that the tradeoff between these values is mysterious in practice, it is mysterious *in principle*. These theories posit a number of values that *necessarily* have no common denominator against which to compare one another. If there was a common denominator, then each of these values would be valuable only instrumentally and thus we should work to maximize the common denominator and not these values per se. It might be possible in theory to rank some list of values according to importance. That is, one might say that health is more important than privacy or vice-versa, but this option does not solve the problem. It would be absurd to say, for example, that the ranking is categorical in the sense that privacy

always outweighs health. This would imply that a patient ought to refuse to disclose the color of her eyes even if it could prevent her from getting cancer. If there is no common denominator and if one cannot rank the values categorically, there must be cases where we must rank some increase in one value relative to some increase in a different value. It is totally mysterious how this process ought to occur. As such, objective accounts of wellbeing may not be useful for practical decisions like the one's faced in the nudging literature.

One obvious response to this concern is that, if objective accounts of wellbeing are true, then it does not matter if they are unwieldy. It would be absurd to adopt an account of wellbeing that is false simply because it makes for easier calculations. It would be much better to make the correct account of wellbeing more precise. For example, calculating the motion of heavenly bodies using quantum physics is vastly more complicated than calculating them using Newtonian physics. And, due to the Heisenberg uncertainty principle, quantum physics implies that some calculations are impossible *in principle*. Yet, it would be absurd to reject quantum physics because it is unwieldy. All that matters is that it is true. Similarly, if we think that some particular objective account of wellbeing is true, then we must make our calculations fit the truth, not the other way around.

Whether this response is persuasive depends in part on what we want a theory of wellbeing to do. A common ways of cashing out the goal of an account of wellbeing are "to describe what is non-instrumentally or ultimately good *for* a person."[153] It is compatible with this definition to describe only general categories of things but not provide any guidance on individual cases, but this seems to miss the point. It seems to me that the *reason* to be concerned

[153] Crisp, "Well-Being."

about wellbeing is so that one can take actions that make one's life go better. But, being able to take these actions requires that the account provide guidance not just at a general level but also on particular cases. If the account of wellbeing does not provide guidance on the particular cases it does not seem to fulfill the basic purpose of the theory. We might go further and say that an account of wellbeing cannot be true unless it accomplishes the basic goals of such an account. By way of analogy, imagine I had a scientific theory that, *in principle*, made no testable hypotheses. It would do no good to respond to objectors by saying "yes, I know my theory does not accomplish the goals of a scientific theory but *it might still be true.*"

Of course objective accounts of wellbeing *do* provide guidance on particular cases. This can be seen in particular by contrasting it with hedonism or desire satisfaction views. Imagine you determine that getting a PhD is likely to result in a net negative amount of happiness, but would be a significant accomplishment. Would your life go better if you pursue it? A hedonist would say no, a perfectionist would say yes. Imagine you desire to watch TV instead of working to complete your dissertation. A desire satisfaction account might say that your life would go better if you watch TV; a perfectionist would say that your life would go better by working on the dissertation.

So, I do not think that concerns about tradeoffs between values can be a reason to reject objective accounts of wellbeing outright. If objective accounts of wellbeing provided *no* guidance about what actions one should take to make one's life go better, then this might provide a reason to reject the account entirely. However, I do think this concern provides motivation for seeking alternative accounts of wellbeing and can provide more precision. In the next section I will discuss a different line of argumentation that further supports this motivation.

3.3.5. The wrong properties objection

One concern for objective accounts might be called the "wrong properties objection." This concern originates with Hurka:

> A perfectionist concept of nature assigns intrinsic value to certain properties, and these must on their own seem morally worth developing. A concept of nature may fail this test by not including some properties that do seem valuable. This flaw is less serious, showing at most that perfectionism needs to be supplemented by other moral ideas. It is more damaging if a concept of nature includes properties that on their own seem morally trivial—if it gives value to what, intuitively, lacks it. This is a telling objection to the concept. A morality based on the concept will be hard to accept because it flouts our particular judgments about value.[154]

One way to think about this objection is that the objective account of wellbeing faces a dilemma. If there is a fundamental connection between the properties then it will include useless properties. If there is no necessary connection, then it will be needlessly arbitrary. We can see the first part of this dilemma by returning to Sher's account. Recall that according to Sher the source of the value of the items on his list is the fact that they are all linked to human goals that are fundamental, meaning that virtually all humans possess and cannot avoid trying to achieve them (they are near universal and near inescapable). But, if this is the determinant of Sher's list, the list appears to be too short. For example, some goals that seem to be near universal and near inescapable include the desire to have sex, urinate, defecate, sleep, eat, obtain dominance over

[154] Hurka, *Perfectionism,* 9.

others and so on. If any of these examples meet Sher's criteria then either they must be added to the list (which either significantly reduces its intuitive appeal) or the list must be arbitrary.

In Hurka's case he answers this objection by restricting the range of relevant essential human properties. But, the justification for doing this is that the properties he picks are independently valuable of themselves.[155] The problem with this justification is that it removes the explanatory power of the essential component of perfectionism. If properties are added or subtracted from the list of essential human attributes by assessing their independent plausibility, then what work does the idea that they are essential human capacities do? It seems that one can eliminate the perfectionist component of the argument and arrive at precisely the same account of value.

Of course this concern need not sink the perfectionist project. As Dorsey puts it "the appeal to essence can still support perfectionism if it can plausibly be maintained that essence-development is itself a plausible ideal independent of any adjustment in light of objections.[156]" This is Hurka's move as well. Hurka admits that for the essentialist picture to be plausible, it must have appeal independent from any of its implications. Hurka claims that for his part he is merely fine tuning this plausible idea.[157]

Dorsey denies that Hurka is merely fine tuning.[158] To determine if Hurka is engaged in fine tuning or if the items on the perfectionists list are determined by their intuitive plausibility, Dorsey offers the *resistance to recalcitrance* test.[159] Dorsey asks us to imagine that we come to believe that "a disposition to develop hypothermia under cold conditions is essential to

---

[155] Dale Dorsey, "Three Arguments for Perfectionism," *Nous* 44, no. 1 (2010): 66.
[156] Ibid.
[157] Hurka, *Perfectionism*, 16, 31, quoted in Dorsey, "Three Arguments."
[158] Dorsey, "Three Arguments," 67.
[159] Ibid.

humanity."[160] If perfectionism is based on intuitions about essential human capacities, we should feel at least *some* pull to override our recalcitrant intuitions and accept the disposition to develop hypothermia as a source of value. Yet, Dorsey argues that perfections feel no such pull whatsoever. Instead, the perfectionist would simply feel pressured to revise the account of human nature, not the items on the list of valuable human attributes.

Bradford[161] argues that while this is a good objection, it ultimately misses the mark.[162] She distinguishes between the content of a particular version of perfectionism and the general view. Dorsey's hypothermia example shows that versions of perfectionism which focus on *uniquely* human traits as the source of value are problematic. But, this fact does not show that the general claim of perfectionism is false. There are other possible version of perfectionism which would not be hampered by this critique. For example, on one version of perfectionism, the relevant features are those which are essential to human nature. Dorsey's critique does not seem to be a problem for this version of perfectionism.

Bradford acknowledges that if all of the particular conceptions of perfectionism fail Dorsey's test, then the general claim is in trouble.[163] I think the most charitable interpretation of Dorsey's claim is that all particular conceptions of perfectionism are likely to fall victim to the *resistance to recalcitrance* test. Bradford notes two prominent version of perfectionism: those based on characteristics that are *unique* to humans and those based on characteristics that are *essential* to humans. But, both of these versions run afoul of Dorsey's test. For the essential version we can find characteristics like sexual reproduction, defecation, urination and so on that

---

[160] Ibid.
[161] Gwen Bradford, "Problems for Perfectionism," *Unpublished Manuscript* (2014).
[162] Note that Bradford ultimately goes on to develop a stronger version of Dorsey's objection that she calls the "Deep Problem"
[163] Ibid., 13.

seem to fail the resistance to recalcitrance test. Of course, uniqueness and essentialness do not exhaust the logical possibilities for versions of perfectionism, so we can never be sure that all of the particular conceptions of perfectionism fall victim to Dorsey's test. However, I think the larger point is that there appears to be a worrying pattern in how perfectionism generates the list of goods. The goods seem to be assessed by whether they are plausible *independent of the perfectionist theory* before being added to the list. Thus, the perfectionist account is not doing the work in generating the list.

Bradford suggests that this is a problem for other theories of value as well. For example, she points to Parfit's famous "stranger on a train" objection[164] to desire satisfaction views. She claims that proponents of the desire satisfaction view respond to this objection not by feeling compelled to think that the stranger's cure is good for you. Instead, they respond by changing which preferences are relevant for wellbeing. So, the objection is not unique to perfectionism.

However, I do not think this is the case. Take hedonism as an example. One common objection to hedonism is that not all pleasures are good. Take a sociopath who gets pleasure out of murdering others. One might object that surely this sociopath's pleasure is not good. A hedonist might respond by adjusting the theory to specify that pleasures that come in harming others are not good. But, many hedonists also say that the pleasure the sociopath experiences is *good for him* even though the effect on overall pleasure in the world is negative. Similarly, while some hedonists might attempt to adapt the theory to avoid experience machine style objections, others simply bite the bullet and claim that the experience machine is good for the person who enters it.[165] The same can be said about the desire satisfaction view. One argument against the

[164] Parfit, *Reasons*, 494.
[165] See the section on hedonism in this chapter of the dissertation for more details

view is that it is sometimes not good for us to get what we desire. Imagine we desire to drink a glass of gasoline thinking it is water, surely meeting this desire is not good for us. While one response to this objection is to restrict which desires contribute to wellbeing, others have chosen to bite the bullet. If I want the gasoline, giving it to me *is good for me* insofar as it meets that desire. Of course, I also have desires not to die and not to consume harmful substances. In total, I probably frustrate more desires than I satisfy by drinking gasoline, but the basic intuition about the value of satisfying desire remains.

Note that all Dorsey needs in support of his claim is that other accounts of wellbeing feel some pull to override recalcitrant intuitions where the perfectionist does not feel this pull. Of course, many hedonists and desire satisfaction theorists do attempt to reformulate their theories to take counterintuitive cases into account, but biting the bullet is a sufficiently common alternative that it suggests that the general hedonist or desire satisfaction intuition is doing some of the work in determining what is good whereas it seems less clear that the perfectionist intuition is playing a similar role. Of course this does not settle the issue, Bradford also[166] argues that if a sufficiently compelling version of perfectionism were developed, perfectionists might well feel sufficient theoretical pressure to include counterintuitive features. Thus, we cannot be sure that Dorsey's basic claim is correct to begin with.

The upshot for Dorsey if he is correct is that perfectionists should instead become objective list theorists. They ought to acknowledge what does the work in determining what is valuable is the independent plausibility of the thing and stop seeking a unifying account of why items appear on the list. There seems to be a price to pay for endorsing such a view. The lack of

[166] Bradford, "The Value;" Bradford, "Problems," 10.

a unifying explanatory structure for why *those* items appear on the list and not other items is dissatisfying. But, how much should we make of this dissatisfaction?

The worry as I see it, is that if the objective list account claims that there are seven types of intrinsic goods, they ought to be able to explain why there are that many and not more or less. We can see why this might be a concern by analogy with biology.[167] Let's say there are two species of elephants and 28,000 species of fish. Is this fact arbitrary? Does it make sense to wonder why there are not three species of elephants and 30,000 species of fish? Do we need an explanation for why those are the numbers? No. There are two species of elephant and 28,000 species of fish because we counted them and that is how many there were. There is nothing mysterious about this fact and biologists can tell stories about species membership and evolutionary lineage that explain how we arrive at those numbers. But, if the number of intrinsic goods is seven, this seems to be a brute unexplainable fact of the universe. That such a fact could exist seems counterintuitive.

The issue can be brought into sharper focus by considering why there is a list of goods at all. Take Sher's list of goods as an example. Sher's list includes relationships that involve companionship and mutual respect and becoming morally better. But, if there's no unifying feature of the list, why identify the good with categories in this way? Why not say, for example, that there is a separate item on the list for instances of simultaneously developing relationships and becoming morally better? This might work in the way that red and blue are colors but so is purple. So, in the way that there are over 10 million colors, perhaps the list of intrinsic goods is millions long. To take things further, why not say that the list of intrinsic goods includes "talking

[167] Bradley, "Objective Accounts."

to my friend Billy over the phone on Tuesday the 21st at 3pm" or "chatting with Aunt Sally over barbeque on Friday the 2nd at 7pm"? That there ought to be categories of goods instead of mere instances of goods seems arbitrary as does the number of goods that are on the list. Clearly this is counterintuitive, but it is unclear how much of a price this is for the objective list theories.

So, if we wish to endorse the objective account of wellbeing we have two options: either the list of goods forms a connection between human essence and welfare or the list of goods is suggested because they are independently plausible but without a unifying reason. In either case, we arrive at somewhat counterintuitive results. For my purposes, however, it may not matter how the issue is decided. My goal is to determine what we should nudge people towards. The objective list theory provides a list of goods that serve as the answer to the question. The perfectionist theory also provides a list of goods that serve as the answer to this question. Because the major goods will be similar on an objective list and perfectionist account, if objective accounts of wellbeing have merit, it seems that we will arrive at much the same answers either way. But, the difficulties in grounding the theory are problematic. So, while this argument might not provide a reason to reject objective accounts of wellbeing entirely, it motivates the search for alternative accounts.

3.3.6. Conclusion for objective accounts of wellbeing

In this section, I provided four objections to objective accounts of wellbeing. I conclude that the cheap thrill and alienation objections only provide reasons to reject objective accounts of wellbeing that do not include pleasure or desire satisfaction as include in the list of intrinsic goods. Other specifications of the goods are unaffected by these criticism.

However, concerns about tradeoffs between values and the wrong properties objection are more serious. To be fair, neither provides a reason to reject the objective account entirely, but taken together they suggest that we ought to prefer plausible alternatives. In the next section I attempt to provide such an alternative by offering a modified version of the informed desire account of wellbeing that I think can withstand the objections raised against that account.

3.4. Actual and informed desire satisfaction accounts of wellbeing

In this section I consider two desire satisfaction theories of the good: actual desire satisfaction accounts and informed desire satisfaction accounts. As a slogan, desire satisfaction theories of wellbeing say that getting what one wants makes one's life go better.[168] Where the two theories differ is that the actualist account holds that what one actually desires is irrelevant. So long as one gets what one wants, things go well for the person. On the other hand, informed desire accounts of the good hold that what is valuable, and what the choice architect should aim for, is what the person *would* choose if they were properly informed, rational, and not weak of will. In what follows, I consider each account in turn and show that each has serious flaws. I then offer some possible alternative viewpoints designed to attempt to solve for these flaws.

3.4.1 Actual desire satisfaction accounts of wellbeing

[168] Throughout this chapter I will sometimes refer to "what makes one's life go best" and other times to "wellbeing." I intend these terms to refer to the same thing.

Actual desire satisfaction accounts of the good hold that one's life goes better if one gets what one wants (whatever it is that one happens to want). According to Heathwood,[169] the simplest possible version of this view would contain three theses:

1. Every basic desire satisfaction is intrinsically good for its subject; every basic desire frustration is intrinsically bad for its subject.
2. The intrinsic value for its subject of a basic desire satisfaction = the intensity of the desire satisfied; the intrinsic value for its subject of a basic desire frustration = -(the intensity of the desire frustrated).
3. The intrinsic value of a life (or segment of a life) for the one who lives it (in other words, the total amount of welfare in the life (or life-segment)) = the sum of the intrinsic values of all basic desire satisfactions and frustrations contained therein.

Some clarifications are in order. First, Heathwood distinguishes between basic (or intrinsic) desires and instrumental (or extrinsic) desires and only counts the satisfaction or frustration of the basic desires. If I desire to watch an action movie because I want to see fancy explosions then seeing the fancy explosions is the basic desire and watching the movie is the instrumental desire. If I watch the movie but do not see fancy explosions, then my desire is frustrated. Second, Heathwood intends this theory to require *concurrence*. That is, for a basic desire to count as being satisfied the state of affairs desired must obtain at the same time as the desire to obtain it. If I want a new car tomorrow, but when I get the new car I no longer want it, then my desire has not been satisfied. In what follows, I borrow from Heathwood[170] to raise some seemingly

[169] Chris Heathwood, "The Problem of Defective Desires," *Australasian Journal of Philosophy* 83, no. 4 (2005): 489.
[170] Ibid.

promising, but ultimately unconvincing arguments against the actual desire satisfaction account. I then raise some stronger objections to motivate the informed desire account.

3.4.1.1. Ill-informed desires

Many writers on the topic of nudges seem to have rejected actual desire satisfaction accounts of wellbeing because they assume that such an account will hold that a chooser's wellbeing is increased in cases where the chooser has made a decision based on incomplete information or ill-formed processes such as weakness of will. Take borrowers from payday lenders as an example. While some borrowers utilize payday lenders when they unexpectedly fall behind on bills, others use payday lenders habitually, borrowing from one or multiple companies each month. If the borrowers calculated how much the habitual borrowing is costing them, they may not longer want to borrow from payday lenders. Some have argued that, despite the fact that they desire payday lending, it surely must be the case that their desire to borrow from payday lenders is not good for them! Indeed, the majority of the examples of nudges concern cases where actual choices appear to be inconsistent with what we imagine fully informed choices would be. Examples include unnecessary credit card debt, lack of 401(k) savings, rent-to-own establishments, lottery tickets,[171] overeating, alcohol abuse, smoking[172] and failure to take necessary drugs.[173] Therefore, one might argue that it must be the case that what makes a person better off is not just the satisfaction of their actual desires.

However, I do not think this is an argument against a sophisticated version of the actual desire satisfaction account of wellbeing. This can be demonstrated with some clarifications.

---

[171] Loewenstein and Haisley, "The Economist."
[172] Flegal, Graubard, Williamson, and Gail, "Cause-Specific."
[173] Jackevicius, Mamdani, Tu, "Adherence with Statin Therapy."

First, it is important to clarify that the action in question and the relevant desire are not identical. In some sense one could say that a borrower who gets a payday loan *desired* to get a payday loan, but recall Heathwood's distinction between basic and instrumental desires above. The most plausible interpretation of getting a payday loan is that getting the payday loan is an instrumental desire, not a basic intrinsic desire. So, getting a payday loan is not, by itself, good (or bad) for the borrower.

The relevant question then is what the basic intrinsic desire is that getting a payday loan is attempting to achieve. Here, the answer is not so clear and it probably depends on the circumstances of individual borrowers. We could imagine that the borrower's intrinsic desire is to achieve financial security and that the borrower desires payday loans as a way of ensuring that they can meet all of their bills. On this interpretation, however, the actual desire satisfaction theory gives us the correct result. Payday loans have hefty fees and make it harder to achieve financial security. As a result, they frustrate instead of satisfying this actual intrinsic desire and so do not increase the wellbeing of the borrower. I think this explains the majority of cases where payday loans seem bad: the borrower is using payday loans to achieve some financial desire, but does not realize that the loans will frustrate that desire. Therefore a sophisticated defender of the actual desire satisfaction accounts can provide the correct answer in these cases by leaning on the widely-acknowledge distinction between intrinsic and instrumental goods.

One might object that the correct description of the basic intrinsic goal of payday loan borrowers is nothing so sweeping as achieving financial security. One might say that when a borrower takes out a $1,000 loan, their basic intrinsic desire is "to have $1,000." Or, perhaps the borrower wishes to take out the loan to pay some bill, then the basic intrinsic desire might be

"pay my bills." In both cases, the loan seems to satisfy the basic intrinsic desire and so, increasing their wellbeing. In that case, the actual desire satisfaction account seems to get the wrong result. Surely, if the borrower really understood the costs of the loan, they would no longer have the basic desire or, at a minimum, they might choose a better strategy to achieve the basic desire. It seems that the actual desire account does not accommodate this intuition.

However, recall Heathwood's third thesis above: "the intrinsic value of a life (or segment of a life) for the one who lives it (in other words, the total amount of welfare in the life (or life-segment)) = the sum of the intrinsic values of all basic desire satisfactions frustrations contained therein."[174] So, if we want to figure out the amount of wellbeing in the borrower's life after taking out the payday loan using an actual desire satisfaction account, we must include the desires that will be frustrated by taking out the loan to get an all-desires-considered result. It seems plausible that payday loans will fare worse than other alternatives when considered from this perspective. Accordingly, the sophisticated defender of the actual desire satisfaction account can get the intuition that taking out a payday loan decreases the borrower's wellbeing correct even if it is true that taking out the borrower has an actual basic intrinsic desire for the payday loan.

3.4.1.2. Irrational desires

Another kind of case that is thought to be a problem for the actual desire satisfaction account of wellbeing is a case where the chooser has all of the information they need but their desires are irrational or the result of weakness of will. For example, many smokers know that smoking is

[174] Heathwood, "The Problem."

bad for them, expensive, and they know that they should quit. Yet, they continue to desire cigarettes. The actual desire satisfaction account says that the wellbeing of the smoker is increased by granting them the cigarette. Surely this is not the case. Here too the actual desire satisfaction account can avail itself of adding up the frustration or satisfaction of desire over time. Smokers may satisfy some basic intrinsic desire for cigarettes by smoking, but they may be frustrating a larger set of basic intrinsic desires for health and money. So, the satisfaction of the smoking desire is outweighed by other considerations.

3.4.1.3. Problems for actual desire satisfaction accounts of wellbeing

In the previous section I showed how a sophisticated defender of the actual desire satisfaction account of wellbeing might respond to some common objections. I showed that many of the most commonly raised objections are not serious objections for the view. However, the actual desire satisfaction account is not without problems. Below I briefly outline three serious problems that face the actual desire satisfaction account and I show why I think this account is deficient.

3.4.1.4. Pointless desires

The actual desire satisfaction account of wellbeing holds that the intrinsic value of a life for the person that lives it is equal to the sum total of all the desires they satisfy less those that are frustrated. But, there seem to be cases where one's life can contain a great deal of satisfied basic intrinsic desires yet does not appear to be high in wellbeing. To reveal the intuitions, consider the pair of cases below:

Alisa: Alisa is a doctor specializing in internal medicine. Alisa is happy and healthy, has a loving partner and is satisfied in all aspects of her life.

*Alisa**: Alisa*'s is identical to Alisa's life in all aspects with one exception. Every morning before work, Alisa* has an overwhelmingly strong basic intrinsic desire to take a sugar pill (commonly used as a placebo in drug trials). Alisa* knows that the pill is a sugar pill and, as a doctor she has access to a nearly limitless supply of such pills at very low cost. Alisa is happy and healthy, has a loving partner and is satisfied in all aspects of her life.

According to the actual desire satisfaction account of wellbeing, Alisa*'s life is going better for her than Alisa's life. This is because, Alisa*'s life is the same as Alisa's but it includes the satisfaction of an additional intense desire every day. I find it doubtful that Alisa*'s life is any better for her or any higher in wellbeing than Alisa's life. I also find it extremely implausible that Alisa*'s life should be *much better* for her with *much more* wellbeing than Alisa's life which is what the actual desire theory seems to imply.[175]

Note that this case is different from the case of irrational desires mentioned above. That is, I do not think that we can posit some all-things-considered reason why Alisa*'s life results in lower desire satisfaction than Alisa's life. Taking the sugar pill does not harm Alisa* in the way that smoking harms the smoker. In fact, it appears to have no effect on her life whatsoever. We might say that Alisa* has some desire not to have seemingly pointless desires and perhaps this desire is frustrated by taking the pill, but we can construct the example to stipulate that this is not the case. Or, we might posit that some combination of the desire for money (some of which she

[175] One could also imagine a view that only counts the satisfaction of basic desires as valuable, but counts some basic desires less than others. This could be done with some 'quality of the desire' addition to the account. However, unless such a modification yields that taking the sugar pill each morning is of *no value* for Alisa*, the counterintuitive result remains.

is spending on pills) and the possibility of frustrated desires from missing future pills outweighs the desire satisfaction from pill taking, but again we can stipulate that this is not the case in the example.

I think the best argument for a proponent of the actual desire satisfaction view is that the thought experiment is somehow unfair. Perhaps it is hard to imagine a person who intensely wants to take a sugar pill for no obvious reason and so, the supposed intuition that the case gets at is distorted. Or, perhaps it is in fact obvious that Alisa*'s life *is* better than Alisa's. I do not find any of these moves plausible, but I think they represent the best defense.

3.4.1.5. Valuable lives

In addition to the problem of pointless desires, the actual desire satisfaction account appears to provide a dissatisfying story about what kinds of lives are valuable for the person who lives them and why. Consider the following excerpt from the *Times* about the lives of female teen-aged crack cocaine users:

> At the crack houses, which are usually decrepit rooms in abandoned buildings, they go on binges that typically last for two or three days.... The girls often perform oral sex in exchange for a smoke. Between binges they sleep in alleyways or abandoned buildings. Adults at the crack houses become the only family the girls have. They often call the older women Ma and the older men Poppy.[176]

To be sure, the actual desire account would also find the girls' lives deplorable. They seem to be satisfying a single desire for crack at the expense of many other important and actual human

[176] *New York Times*, August 11, 1989, A13, quoted in Sher, *Beyond Neutrality*.

desires. However, as Sher[177] points out, we can discuss the unhappiness of being in such a situation, the risks of violence and disease, and the frustrated desires, but to understand the appalling nature of the situation only in these terms seems to miss the point. Such a reaction would:

> ...fail to capture (what I take to be) the standard reaction that sleeping in alleyways and trading oral sex with strangers for intervals of drug-induced euphoria are simply not good ways for humans to live. In addition, it would fail to capture the sense that these are wasted lives, devoid of constructive activity or any meaningful prospects for it.[178]

What I think this line of thought gestures towards is the general idea that something outside of desire satisfaction might better explain why certain kinds of lives seem valuable for the person living them and why other lives do not. In some sense, this is not an argument against the actual desire satisfaction account. The committed actual desire satisfaction theorist has no compulsory reason to change their view based on this idea. However, for those who are undecided about what it means for a person's life to go well for them, I think the intuition can be powerful.

3.4.1.6. The objection from remote desires

The objection from remote desires is a kind of "irrelevant desires" objection to the actual desire satisfaction account of wellbeing. That is, it argues that there are certain kinds of things that the actual desire satisfaction account claims increases a person's wellbeing but which, intuitively, do

[177] Sher, *Beyond Neutrality*.
[178] Ibid. 179.

not. This objection has been raised by Scanlon,[179] Griffin,[180] Kagan,[181] Sumner,[182] and Murphy[183] among others. However, the most famous formulation comes from Parfit:[184]

> *Stranger:* Suppose I meet a stranger who has what is believed to be a fatal disease. My sympathy is aroused, and I strongly want this stranger to be cured. We never meet again. Later, unknown to me, this stranger is cured. On the Unrestricted Desire-Fulfilment Theory, this event is good for me, and makes my life go better. This is not plausible. We should reject this theory.

I see Parfit's argument as part of a class of cases where the satisfaction or frustration of a desire seems to be unrelated to the person's wellbeing. For example, we can imagine that I have a basic intrinsic desire for the total number of atoms in the universe to be a prime number or I have a basic intrinsic desire Napoleon's favorite color to be red. Does the satisfaction of these desires have anything to do with my wellbeing? It seems that the answer is no, and so, the actual desire theory seems implausible.

To address this concern, recall that in section 3.2.1. above I noted that Heathwood's actual desire satisfaction account requires *concurrence*. That is, for a basic intrinsic desire to count as being satisfied the state of affairs desired must obtain at the same time as the basic intrinsic desire to obtain it. This offers a way out for Heathwood. When the stranger is cured, I no longer have the basic intrinsic desire that he be cured, so there is no concurrence. Thus,

---

[179] Thomas Scanlon, "Value, Desire and Quality of Life," in *The Quality of Life*, eds. Martha Nussbaum and Amartya Sen (1993): 186-187; Scanlon, "The Status of Well-Being," (1996): 16-17; Scanlon, *What we Owe to Each Other*, 120-121.
[180] Griffin, *Well-Being,* 16-17.
[181] Shelly Kagan, "The Limits of Well-Being," in *The Good Life and the Human Good*, eds. Paul & Miller (1992): 169-189.
[182] L. W. Sumner, *Welfare, Happiness and Ethics* (Oxford: Oxford University Press, 1996), 132.
[183] Mark Murphy, "The Simple Desire-Fulfilment Theory," *Nous* 33, no. 2 (1999): 269.
[184] Parfit, *Reasons*.

Heathwood can claim that the stranger being cured is not good for me which seems to be the intuitive result.

But if concurrence solves the problem of remote actual desires for actual desire satisfaction accounts of wellbeing, it does so by sacrificing far too much. Concurrence cannot allow for posthumous harms and benefits. Luper provides the following example to show that posthumous harms exist:

> *The Achievement*: Suppose I want to conduct research that will lead to a cure for Lou Gehrig's disease, ALS. Suppose, too, that my desire is essential to my life plan, and that my plan is rational (more about this later). Unfortunately, I will die before I achieve what I want, but I will still succeed if various events occur, and fail if some other events occur, after I am dead. For example, I will succeed if my research gives another scientist a critical clue which she develops into a cure that she otherwise would not have found. And I will fail if all of the records of my research are destroyed in a fire before they prompt another scientist to devise a cure. Upon reflection, I dread the prospect of the fire destroying my files even though I will be dead at the time it would occur; I judge that it would be against my interests. By contrast I welcome the prospect of my research inspiring a colleague; I judge that it would be in my interests.[185]

It seems extremely plausible that my life will go better if I succeed in my goal of contributing to a cure for ALS and that my life will go worse if my research is destroyed. Yet, because I am dead, there is no one around to have these basic intrinsic desires when the event occurs. So, concurrence holds that such events have no effect on wellbeing. This is implausible.

---

[185] Luper, "Retroactive Harms."

In addition concurrence rules out cases where I have a desire about the past. Imagine I receive a big inheritance from my family. I desire that my family not have made the money by harming other people. I discover that my family owned a large number of slaves in the past who they treated extremely cruelly and that my inheritance is the result of this past mistreatment. According to the concurrence requirement, because I do not have the desire that my family not have harmed people at the same time that they harmed people, the frustration of this basic intrinsic desire cannot be bad for me. This too seems implausible. It seems that my life would go better if, in accordance with my basic intrinsic desires, my family's money had been inherited ethically.

Therefore, I think concurrence is too high a price to pay to solve the problem of distant desires. In addition, I do not think that concurrence actually solves the original problem. Consider the following case:

> *Stranger*:* Suppose I meet a stranger who has what is believed to be a fatal disease. My sympathy is aroused, and I have a strong basic intrinsic desire that this stranger be cured. We never meet again. Years late I am riding the same train and I happen to remember my meeting with the stranger. My sympathy is again aroused and I again have a strong basic intrinsic desire for he stranger to be cured. Unknown to me, at that exact moment, the stranger is undergoes a therapy that cures the disease.

Stranger* is the same as Stranger except that I have added a reoccurrence of the desire that just happens to occur at the same time as the stranger is cured. On Heathwood's account, the stranger's cure is good for me and makes my life go better. This seems just as implausible as the original Stranger account. So, concurrence does not solve the problem of distant desires.

Finally, concurrence can not solve for many other kinds of distant desires. For example, if I desire that the number of atoms in the universe be a prime number, I have that desire now and there is a fact about the number of atoms in the universe now. Therefore, if the actual number of atoms is a prime number, my life goes better for me. We can imagine other such cases where I have a desire now and the fact of the matter occurs now, but the concurrence does not seem to make my life go better.

A plausible way to respond to this issue is to restrict the set of desires that count as increasing a person's wellbeing. We will turn to this possibility in the rest of this chapter.

3.4.1.7. Conclusion for actual desire theories

So far in this chapter of the dissertation I have considered the actual desire satisfaction account of wellbeing. According to this account, a person's life goes well when her actual desires are satisfied and it goes poorly when her actual desires are frustrated. Many philosophers have thought that it is obvious that some of the things that we want are not good for us, so the theory is clearly implausible. I consider and reject some of these common objections and show that the theory ought to be taken seriously. However, I then raise three must stronger objections for the theory. I think these objections provide good reason to reject the account and look for more plausible alternatives.

3.4.2.1. Informed desire satisfaction accounts of wellbeing

Informed desire theories are an attempt to rectify the intuitively appealing notion that what is valuable is that people get what they desire with the evidence from social psychology and

behavioral economics suggesting that what people desire does not always appear to be what is best for them. At its core, informed desire theories hold that what is valuable, and what the Choice Architect should aim for, is what the person *would* choose if they were properly informed and rational and not weak of will.

As an illustrative example, Thaler and Sunstein draw a distinction between Humans and Econs. Econs are perfect maximizers of their own wellbeing who "can think like Albert Einstein, store as much memory as Big Blue, and exercise the willpower of Mahatma Gandhi.[186]" Informed desire theories hold, essentially, that what makes a person's life go well is what an Econ would choose. Put another way, a nudge improves a person's life if it gets a person to choose what they would have chosen if they were an Econ. Informed desire accounts also share the distinction of being the predominant accounts in the literature. In Nudge, Sunstein and Thaler adopt the informed desire account explicitly[187] and the majority of authors seem to have followed suit. Indeed, the majority of the examples of nudges concern cases where actual choices appear to be inconsistent with what we imagine fully informed choices would be. Examples include unnecessary credit card debt, lack of 401(k) savings, payday loans, rent-to-own establishments, lottery tickets,[188] overeating, alcohol abuse, smoking,[189] and failure to take necessary drugs.[190]

In this section I will investigate the informed desire account in detail. I will review the history and various formulations of the idea, show why the informed desire account might be appealing and then consider the major objections that have been raised against the account. I will

[186] Sunstein and Thaler, "Nudge: Improving Decisions."
[187] Sunstein and Thaler, "Nudge: Improving Decisions," 5.
[188] Loewenstein and Haisley, "The Economist."
[189] Flegal, Graubard, Williamson, and Gail, "Cause-Specific."
[190] Jackevicius, Mamdani, and Tu, "Adherence with Statin Therapy."

ultimately conclude by showing how the informed desire satisfaction account might be reformulated to deal with common objections.

3.4.2.2. History of the informed desire account

A truly impressive list of of ethicists have found full information accounts of wellbeing to be plausible. Proponents include: Mill,[191] Sidgwick,[192] Brandt,[193] Hare,[194] Griffin,[195] Harsanyi,[196] Darwall,[197] Railton,[198] and Gauthier[199] among others. At their core, full information accounts involve specifying some privileged epistemic standpoint that has enough knowledge about all of the various options on offer such that what one would choose in light of this knowledge definitively determines the good for the non-informed agent. Many have found the use of such a privileged epistemic position plausible.

The intellectual forebearer of this idea is Mill. In Utilitarianism Mill rejects Bentham's identification of the good with the sensation of pleasure because such a view would hold that all pleasures are equally valuable. Instead, Mill substitutes a "competent judges' test" which defines a privileged epistemic position according to which the relative value of two pleasures can be determined. Mill writes: "Of two pleasures, if there be one to which all or almost all who have experience of both give a decided preference, irrespective of any feeling of moral obligation to prefer it, that is the more desirable pleasure."[200] Here, Mill is arguing that the epistemic position

---

[191] Mill, *Utilitarianism*.
[192] Sidgwick, *The Methods*.
[193] Richard Brandt, *A Theory of the Good and the Right* (1979).
[194] Richard Mervyn Hare, *Moral Thinking* (Oxford: Oxford University Press, 1981): 214-216.
[195] Griffin, *Well-Being*.
[196] John Harsanyi, "Morality and the Theory of Rational Behavior," in *Utilitarianism and Beyond,* ed. Amartya Sen and Bernard Williams (1982), 55.
[197] Stephen Darwall, *Impartial Reason* (1983).
[198] Peter Railton, "Facts and Values."
[199] David Gauthier, *Morals by Agreement* (1986).
[200] Mill, *Utilitarianism* .

of having experienced both pleasures is privileged in determining which pleasure is more desirable. Of course, Mill's judge is not idealized in the way that the informed desire account requires, but the general idea that what one would choose in light of knowledge that the agent does not have may determine what is good for the agent.

Sidgwick builds on Mill's account by expanding on the requirements of the position from which the agent is to assess her options. On Mill's account there might be many different epistemic position from which one can assess the available options. As long as the agent is in the epistemic position to have knowledge of both pleasures, it can assess the options. Sidgwick modified this account by specifying a *single* privileged position from which what is good for a person must be ascertained. Sidgwick's account is as follows:

> A man's future good on the whole is what he would now desire and seek on the whole if all the consequences of all the different lines of conduct open to him were accurately foreseen and adequately realized in imagination at the present point in time.[201]

This move from multiple privileged positions to a single privileged position has been consistently emulated by later informed desire theorists. For example, Brandt elaborated on this framework by proposing the informed desire account as a reforming definition of rationality.

> I shall pre-empt the term 'rational' to refer to actions, desires, or moral system which survive maximal criticism and correction by facts and logic.... This whole process of confronting desires with relevant information, by repeatedly representing it, in an ideally vivid way, and at the appropriate time, I call *cognitive psychotherapy*.[202]

---

[201] Sidgwick, *The Methods*.
[202] Brandt, *A Theory*, 10, 113, 329.

Unfortunately, there are some problems with the Sidgwick and Brandt view. The first argument is raised by Railton.[203] Imagine I am considering whether I should gain more information before making a decision. If my good consists in what I would seek "if all the consequences of differ lines of conduct were accurately foreseen," then I would not seek to gain additional information. This is because the version of me who knows the consequences of different lines of conduct does not need to waste time seeking additional information. In addition, the idealization process appears to change the desires of the idealized self in ways that seem irrelevant to the non-idealized version of me. Sobel argues[204] that my idealized self would have a refined pallet from foreseeing all the lines of conduct involved in becoming a sommelier and would not enjoy cheap wine. But, since my non-idealized self cannot tell the difference between expensive wine and cheap wine, it seems strange to claim that drinking expensive wine is what is best for me. Railton attempts to solve these problems with the following account:

> An individual's good consists in what he would want himself to want, or to pursue, were he to contemplate his present situation from a standpoint fully and vividly informed about himself and his circumstances, and entirely free of cognitive error or lapses of instrumental rationality.[205]

A similar view was also taken up by David Lewis.[206] This view solves the problems above by substituting the identification of the interests of our informed selves with our actual selves for a "wanting to want" framework. In the case of gaining information, I *would* want myself to want to gain information so, Railton's account could hold that gaining information is good for me. In the

---

[203] Peter Railton, "Moral Realism," *Philosophical Review* 95 (1986): 174.
[204] David Sobel, "Full Information Accounts of Well-Being," *Ethics* 104, no. 4 (1994): 792. This critique can also be found in Griffin, *Well-Being*.
[205] Railton, "Moral Realism," 16.
[206] David Lewis, "Dispositional Theories of Values," *Proceedings of the Aristotelian Society* 63 (1989): 113-137.

case of drinking expensive wine, I *would* want myself to want to get as much pleasure from the cheap wine as from the expensive wine. So, the theory yields the correct results.

Most recent critiques of the informed desire theory have focused on Railton's view, so I will use it as a starting point for a deeper dive into the account. But first, in the next section, I discuss the motivations one might have to appealing to an informed desire account.

3.4.2.3. Motivations for the informed desire account

In this section I discuss why so many philosophers have found the informed desire account plausible and why one might be motivated to adopt the informed desire account.

The primary motivation for adopting the informed desire account seems to be widespread dissatisfaction with the actual desire satisfaction account. The widespread view is that actual desire satisfaction theories are false because we can desire things that are bad for us. As Loeb explains:

> Thirsty though I am, I would no longer wish to drink the liquid in that glass were I to learn that it is poison and not water. My desire to see that man hanged would vanish upon learning that he is the defense attorney and not accused. And my desire to give up my career as a philosopher and become a lifeguard would diminish if I were to learn that lifeguards earn even less than philosophers.[207]

What these cases attempt to demonstrate is that in situations where additional information would change a desire, the desire with the additional information seems to be more valuable than the desire without the additional information. We might lump these arguments together and call

[207] Dan Loeb, "Full-Information Theories of Individual Good," *Social Theory and Practice* 21, no. 1 (1995): 1.

them arguments from *ill-informed desires*. As I argued above, I think this objection to the actual desire satisfaction account is mistaken and, accordingly, this concern is not a good motivation for adopting an informed desire account.[208] A sophisticated actual desire satisfaction account can acknowledge that such states of affairs are bad for me because the total frustration and satisfaction of desires that they entail is negative, even though they do satisfy the initial desire.

However, I do think there are powerful intuitions behind the the account. Recall that in chapter 2 I discussed two such intuitions in favor of the informed desire account. They are: *internalism* and *desire improvement*. I briefly sketch these intuitions out again below.

Internalism is the view that there should be a link between a person's desires and what is good for her. The basic intuition is that it is hard to believe that something can be good for a person if the person cares nothing about it. This is one weakness of objective list or perfectionist theories of the good. For example, Sher has developed a perfectionist account that includes understanding the world, the formation and execution of reason-based plans, relationships that involve companionship and mutual respect, developing one's abilities, becoming morally better, becoming more aware of beauty, developing decency or good taste, and privacy.[209] We can imagine a person who sincerely and deeply has no desire to become aware of beauty. Sher would hold that despite this, becoming aware of beauty is good for him. I'm sure Sher would have much to say on the topic, but at first look this seem somewhat implausible. Thus, many have found the internalism intuition appealing.

The second intuition is *desire improvement*. Desire improvement is the intuition that while what is good for a person should be linked to a person's desires, not just any desires should

[208] See section 3.2.1.1. above.
[209] Sher, *Beyond Neutrality*, 90.

count. In particular, desires that would not survive certain kinds of improvement should count for less.[210] To be sure, this intuition can partly be explained by cases where the satisfaction of one desire leads to the frustration of other desires. For example, if I desire not to go to the dentist because I have a phobia of the dentist, my not going to the dentist will satisfy that desire, but it will frustrate desires to have a healthy smile and not be in pain when eating. Yet, other cases do not seem to fall into this pattern. This is especially true when the pathway that I choose generates new desires that I would not want to fulfill from my present state. For example, I currently desire to achieve important things and make valuable contributions to society. I could imagine a set of actions that would make me desire things that are easier to obtain like watching TV and playing video games. This pathway might lead to a greater magnitude of satisfied desires but it conflicts with what I would want myself to want.

While there is some intuitive appeal behind the account, it also faces serious objections. In the next section I sketch out some of these objections.

3.4.2.4 Informed desires satisfaction accounts and problems with actual desire satisfaction accounts

One motivation for the informed desire account is that it is supposed to solve problems with the actual desire satisfaction account. As I have already argued, the most common problem that the account is designed to solve is not a problem for a sophisticated actual desire satisfaction view. I did, however, raise three objections to the actual desire satisfaction account that the informed desire account might, presumably, be expected to solve. In this section I will argue that the view

[210] The usual position for informed desire theorists is that desires that would not exist under conditions of full information are not valuable at all. However, I have framed the *desire improvement* intuition here to include the possibility that such desires do count for something.

does the following: 1) it solves the problem of pointless desires; 2) it is helpful in solving the problem of pointless lives and; 3) it does not solve the problem of remote desires.

For the actual desire satisfaction theory, the problem of pointless desires is that some possible actions will create desires that will then be immediately satisfied. And, this simultaneous creation and satisfaction of desires does not seem valuable. For example, if I create in you a desire to take an inert pill and you are able to take the pill, the actual desire satisfaction account says that this is valuable for you. But, this does not seem to be the case.

The informed desire account can solve this problem. The slogan for the informed desire account is that what is good for you is what an idealized version of yourself would want yourself to want. An idealized version of myself does not seem likely to want myself to want to take an inert pill every day, so on the informed desire account, this is not valuable for me.[211]

The informed desire account also provides some assistance with the problem of valuable lives. The general idea here is that in some cases the actual desire satisfaction account does not seem to fully capture our intuitions about what is valuable. It does not seem to capture that some kinds of lives (like the ones lived by teenage girls trading oral sex for crack) are deplorable in a way that does not seem to reduce to just the frustration of desires. Similarly, the actual desire satisfaction account does not seem to capture that some things like obtaining a PhD are valuable in a way that does not seem to reduce to just the satisfaction of desires.

The informed desire account is helpful on this problem because it specifies a privileged position on which to find the lives deplorable. It seems clear that girls addicted to crack would

[211] Of course one might argue that an idealized version of myself *would* want to take the pill every day. I find this implausible. However, because the details of what an informed person would want are unclear, this remains a life possibility. The difficulty of knowing what a fully informed person would do is a weakness of the account. I discuss this more fully later.

not want themselves to want to smoke crack and would not want themselves to want to live such lives. Thus, their lives are deplorable not just because they experience more pain than pleasure or more frustrated desires than satisfied desires, but because from some privileged epistemic position, this is not how they would want themselves to want to live. Whether this solves the problem depends on the intuitions that such cases generate. Sher seems to have an *externalist* intuition according to which what some things can be good or bad for a person regardless of what they (or an idealized version of them) want. So, the informed desire account is not quite in accordance with Sher's intuition, but it seems to get a better result than the actual desire satisfaction account.

However, the informed desire account does not solve the problem of remote desires. Recall that the problem of remote desire can be summed up with the following cases (adapted from Parfit):

> *Stranger*:* Suppose I meet a stranger who has what is believed to be a fatal disease. My sympathy is aroused, and I strongly want this stranger to be cured. We never meet again. Years late I am riding the same train and I happen to remember my meeting with the stranger. My sympathy is again aroused and I strongly want the stranger to be cured. Unknown to me, at that exact moment, the stranger is undergoes a therapy that cures the disease.

The actual desire satisfaction account holds that the stranger being cured is good for me. This seems hard to believe. Unfortunately, it seem likely that my informed self could form similar desires for me. Griffin explains:

> [T]he restriction to informed desires is no help here. I might meet a stranger on a train and, listening to his ambitions, form a strong, informed desire that he succeed, but never hear of him again ... any moderately decent person wants people living in the twenty-second century to be happy and prosperous.[212]

So, it seems that my informed self would want to stranger to succeed and so the remote desire remains. There are a few options at this point. The first is to bite the bullet and accept that the stranger being cured is good for me. The second is to find a way to restrict which satisfied desires increase a person's wellbeing such that the stranger being cured does not count. I will evaluate the plausibility of these options as part of the larger discussion about the costs and benefits of different accounts of wellbeing over the rest of this chapter. In the next section I discuss a problem that is unique to informed desire accounts.

3.4.2.5. Problems with the idealization process

A unique problem that arises for the informed desire account comes from difficulties in specifying the nature of the process that must be undertaken to become an ideal advisor. An impressive amount of discussion has been generated on this topic. Authors including Rosati,[213] Loeb,[214] Tiberius,[215] Sobel,[216] Enoch,[217] Carson,[218] Velleman,[219] Gibbard,[220] Anderson,[221] Daniels,

---

[212] Griffin, *Well-Being,* 17.
[213] Connie Rosati, "Persons, Perspectives and Full Information Accounts of the Good," *Ethics* 105, no. 2 (1995).
[214] Loeb, "Full-Information."
[215] Valerie Tiberius, "Full Information and Ideal Deliberation," *Journal of Value Inquiry* 31, no. 3 (1997).
[216] Sobel, "Full Information."
[217] David Enoch, "Why Idealize?," *Ethics* 115, no. 4 (2005).
[218] Thomas Carson, "Rationality and Full Information," in *Ethical Theory*, 2nd. ed. Russ Shafer-Landau: (Wiley-Blackwell, 2013).
[219] David Velleman, "Brandt's Definition of 'Good,'" *The Philosophical Review* 97, no. 3 (1988).
[220] Allan Gibbard, *Wise Choices, Apt Feelings* (Harvard University Press, 1990).
[221] Elizabeth Anderson, *Value in Ethics and Economics* (1993).

[222] and Harman[223] have found all or parts of the view problematic.[224] In this section I will discuss objections raised by Rosati[225] and Sobel[226] both of which deal heavily with questions concerning the process of becoming an idealized agent. I find that these objections pose serious problems for the informed desire satisfaction account. I then attempt to develop a model that attempts to avoid these criticisms.

3.4.2.6. Rosati's criticisms of the informed desire satisfaction account

One of the most influential criticisms of the informed desire account comes from Roasti.[227] She provides two arguments against the view. First, she argues that "the 'fully informed' person, though purportedly you, may not be someone whose judgment you would recognize as authoritative; thus, Ideal Advisor views lack normative force.[228]" Rosati then argues that "because of what it is like to be a person and to have a perspective, it appears that no person can be fully informed."[229] Both of these problems stem from the fact that the idealization process envisioned is "purely causal."[230] She argues that a consequence of this way of specifying the ideal advisor is that it cannot capture the normative force of our notion of a person's good which "does not await an answer to this empirical question."[231] However, she does argue that while the

[222] Norman Daniels, "Can Cognitive Psychotherapy Reconcile Reason and Desire?," *Ethics* 93 (1983).
[223] Gilbert Harman, "Critical Review," *Philosophical Studies* 42 (1982).
[224] Luke Muehlhauser and Chris Williamson, "Ideal Advisor Theories and Personal CEV," *Machine Intelligence Research Institute* (2013); Gerald Beaulieu, "The Normative Authority of Our Fully Informed Judgments," *The University of Manitoba* (1997).
[225] Rosati, "Persons."
[226] Sobel, "Full Information."
[227] Rosati, "Persons."
[228] Ibid., 299.
[229] Ibid.
[230] Ibid.,324.
[231] Ibid.

informed desire view cannot tell one what constitutes their good, it can be serviceable as a "regulative ideal that guides theoretical inquiry about a person's good."[232]

Rosati's first argument comes from considerations about the kinds of changes one would have to go in order to become fully informed. One shortcoming in the full information process comes from specifying exactly how one would gain full information. This matters because how events or experiences are ordered can have a profound effect on one's evaluations of them. As she puts it "poverty after wealth is experienced differently than poverty after near-poverty or wealth after poverty."[233] To counteract the problem of ordering, a person would need to experience all possible lives in all possible orders. But, such a person would have to have capacities in reason, memory and imagination that are far greater than one's actual abilities and one would need to be able to retain certain features of one's self after the information process has been completed. Thus, Rosati argues if I undergo the idealization process, we have not learned what I would want or even what someone like me would want. Instead, we have learned what a very strange person who happens to be continuous with me would want. Thus, the idea lacks normative force.

Recall the example from chapter 2 of Frank, who is uninformed, lacks self-control and is relatively irrational. Frank has a set of actual desires to eat unhealthy food and refrain from exercise, preferring instead to watch TV and play video games. If Frank were fully informed with more knowledge, rationality and self-control he would instead desire to be physically fit. On the informed desire account we ought to nudge Frank towards a healthy lifestyle. Imagine that we begin to work to make Frank more physically fit. We make it more difficult for Frank to get

[232] Ibid., 325.
[233] Ibid., 309.

to the McDonald's and easier to get to the gym. We make the salad menu more prominent and the hamburger menu harder to find. Let's imagine we're successful in this endeavor and as a result Frank grows healthier each day, but Frank does not develop any adaptive preferences and retains his original desires to be lazy and eat unhealthy food. We can imagine that Frank grows more miserable as his desire to sit around the house and eat Big Macs is consistently frustrated. No actual desire of Frank's is ever satisfied by living his new healthy lifestyle.

On the informed desire account what we have done for Frank is tremendously valuable because we have satisfied Frank's hypothetical desires. This, however, does not match our reflective judgments about value. It is not clear that it is good to nudge Frank in this way and it seems very implausible that what we have done for Frank is the tremendously valuable in the way that the informed desire account takes it to be. In addition, two considerations make the intuition stronger. First, we can run the situation in reverse. Imagine Frank's actual desire is to be fit, but Frank's hypothetical informed desire would be to sit around playing video games and eating Big Macs. In that situation it seems clear that nudging frank towards his hypothetical desire is not good. Second, even if one thinks that nudging Frank towards hypothetical desires at the expense of actual desires is good, it is at least the case that it counts against an action if it frustrates actual desires. That is, it would be better if Frank desired to be healthy and was nudged towards health. The informed desire view appears to be unable to take this into account.

Rosati's second argument is that "because of what it's like to be a person and to occupy a perspective, it appears that no person could be fully informed.[234]" The full information requirement entails that a person must take on all possible points of view in order to gain

[234] Ibid., 314.

information. Yet, the person must also "forget" those points of view when taking on a new possible point of view. But, because some traits experienced in some lives are mutually exclusive, the fully informed person cannot take both traits into account simultaneously, which renders her less than fully informed. As an example, imagine that the fully informed person lives one life with the trait of kindness and one life with the trait of meanness. Kindness and meanness change one's perspectives on the world. Kindness afford opportunities to help others and see the good in people; meanness affords opportunities to harm others and see the evil in people. The fully informed person cannot be both kind and mean because, as a conceptual matter, these traits cannot occur at once. It could be that through the experiences, the full informed person acquires certain traits in some order and then exits the idealization process with either meanness or kindness as the retained traits. Yet. this cannot be so as it violates the requirement that the ordering of information not have an influence on the outcome. So it seems that there is no way for the fully informed person to actually be fully informed.

Rosati's more general concern is that the full information process is causal in nature. That is, facts about what would happen to a person who underwent the full information process determine what is good for a person. Yet, it seems unusual that our basic normative view could depend on causal facts in this way.

3.4.2.7. Sobel's objection to informed desire satisfaction accounts

Sobel raises a number of objections, but I will consider the two objections that I take to be the strongest. The first objection is what I call the "too many voices" objection. The key idea here is that even if there is a privileged *epistemic* position for determining what is valuable for a person,

there does not seem to be a similarly privileged *temporal* position. Because agents change over time, and because what the fully informed advisor desires is intended to depend in some way on what the actual agent wants, it follows that what the fully informed advisor would desire would also change over time. So, there is not one informed desire, but many, each representing a different time slice. If it is possible that, for example, the fully informed version of me at age 60 wants the actual me to do different things than the fully informed version of me now, then we need some way of adjudicating between the competing voices. Yet, the informed desire account provides no principled way of adjudicating between these competing voices.

I see four interesting responses to this objection. The first would be to establish some privileged temporal position. For example, we could say that what is good for a person is what their *now* fully informed advisor would advise them. But, this seems arbitrary. It seems quite plausible that my 60-year-old fully informed advisor might be in a better epistemic position than my now advisor.[235]

A second response would be to aggregate the perspectives of the multiple temporal ideal advisors. However, this option removes the appeal of the account because it does not seem clear that an aggregative advisor is epistemically superior. This option also requires that one both determine what an ideal advisor would advise (which is challenging as noted earlier) and determine how the advice of multiple advisors would interact. Given the difficulty of imaging what an idealized advisor would advise, determining what an aggregation of idealized advisors would advise would seem to be a nearly insurmountable challenge.

---

[235] In addition, this view seems to require endorsing the A-theory of time which invites more metaphysical baggage than we might want to defend.

A third option is to deny that ideal advisors at different times would provide different advice. For example, it could be that part of the idealization process is to remove the time-bias that often influences decision making. However, this is a serious thesis that would require a robust defense and an explanation of the idealization process that allows for it. One challenge for such a defense is that if the ideal advisor does not make different recommendations at different *times*, it is unclear why the advisor would make different recommendations for different *people*. Such a view might be committed to thinking that what one should do does not depend on their circumstances. Finally, one might argue that the ideal advisor is somehow timeless. I'm not sure how such a view might operation, but it exists as a possibility.

Sobel's second objection is what I call the "amnesia" objection and it is similar to Rosati's second argument. The idea is that some lives can only be evaluated if they are experienced. Yet, experiencing some lives can leave one incapable of experiencing other lives in an unbiased way. To help deal with this issue, Sobel presents an amnesia model as an attempt at developing a plausible way for an idealized agent to gain the experience necessary to evaluate all the relevant lives. On this model, the idealization process involves experiencing each life sequentially but but then undergoing an amnesia experience before undergoing the next life. Once all the lives have been experienced the amnesia is removed.

There are some important issues with this model. One issue is that the agent might experience vastly dissimilar lives that are evaluable from the perspective of each other. For example, the agent might experience a life where he lives in complete isolation with no experience of modern technology and one where he is an IT professional. These lives are completely dissimilar, so it seem possible that the agent might not be capable of rendering any

judgment about how to compare them. Relatedly, Sobel argues that the idealization process might simply drive a person insane. In fact, the process is so unusual and incoherent, that it might be a necessary outcome of idealization that it drives agents to insanity.

3.4.2.8. In defense of the informed desire account

As demonstrated above, there are genuine, serious concerns about the idealization process in the informed desire account as specified by Railton, Brandt and others. Yet, I do not think this fact implies that the informed desire account should be abandoned. Recall that, at its core, the informed desire account is committed to the two intuitions outlined in 3.4.2.3., namely, internalism and desire improvement. That is, an informed desire account should hold that what is good for me is related to what I desire and it should hold that not all desires are equivalently good for me. The problems raised with the idealization processes in the previous section are not problems for all theories that hold these two views. Instead, they are only problems with what I will call single-agent idealization processes. However, I think that multiple-agent idealization processes may also be possible which avoid these issues.

3.4.2.9. Multiple-agent idealization processes for informed desire satisfaction accounts of wellbeing

On the single-agent model of idealization, an individual's good consists in what he would want himself to want under conditions of full vividness and full information. Problems arise in this view because the goal is to create a single *idealized* advisor who has a privileged epistemic position. Yet, it seems that it may be impossible for a single person to be in such a position. For

example, a privileged epistemic position might require that the person have both the trait of kindness and the trait of meanness which seems to be in contradiction. I think there is a way around this dilemma but first allow me to introduce some terminology. As we have already discussed, an *idealized* agent, is a single agent which holds a privileged epistemic position. An idealized agent can have no epistemic gaps. There can be no obvious deficiencies in her epistemic view because otherwise she would not be idealized. Alternatively, we can imagine an agent who is merely *extrapolated* and is not necessarily idealized. An extrapolated version of an agent is what the agent would be like if they underwent certain changes and experiences. The extrapolated agent need not undergo all changes and experiences but merely undergo some of them such that the extrapolated agent have additional knowledge.

In most of the commonly cited cases supporting informed desire accounts, the intuitive force behind the case can be attained by only specifying that the agent is extrapolated in particular ways not that the agent is idealized. The problem, however, is specifying what extrapolation is required. To avoid this issue, I propose a multiple-agent model that makes use of many extrapolated agents instead of a single idealized agent.

Instead of the single-agent idealization process, imagine that the idealization process is conducted via a parliamentary model where all the members are different variants of myself who have experienced different possible lives. These parliamentary members start out as identical to me at my current moment of time and return as the version of me that would result from having a particular set of experiences. An example will reveal the intuitions.

Imagine that some person, call him Ted, has just graduated from college and is considering various career opportunities. Which career option is best for Ted? On this model, to

answer this question, we imagine that a version of Ted, call him Ted′, becomes the first of Ted's career options, an investment banker, and then reports back to parliament. Another version of Ted, call him Ted″ pursued another of Ted's options becoming a philosopher and reports back to the group. Another version of Ted, call him Ted‴ becomes a poet and then reports back to the group and so on for all possible options. Then, the parliament of Ted debates which option is best. We can imagine that the parliament of Ted has as much time for debate as they need, has unlimited patience and so on to promote reasonable debate. We can also imagine that each version of Ted is able to provide a maximally clear story about his experiences. After hearing all of the stories and understanding all of the experiences, they then vote for what would be best for the original Ted by rank-ordering all of the experience relayed by each other version of Ted according to which life seems best for him. The experience with the most votes is the one that is best for Ted.[236] Naturally, the experience that each version of Ted undergoes will shape their vote. Ted′ (the investment banker) may have grown to love money and so prefers lives that have a good deal of money. Ted‴ (the poet) many have come to love art and so prefers lives with art in them. This is accounted for in the vote.

While this example is slightly unusual, as an idealization process it solves some of the problems raised by the informed desire account. First, it may help to solve the issue that the informed advisor lacks normative force. Consider the case of Frank raised in 3.4.2.4. In that case, the informed desire view yielded the counterintuitive result that it is as good for Frank to be physically fit, but constantly have his desires to watch TV and eat McDonald's frustrated as it is for Frank to be physically fit and have an actual desire to be physically fit. The multiple-agent

[236] To be more specific, we can imagine that the delegates use any of a number of counting methods for ranked voting system like a Borda count.

idealization process avoids this problem. Because satisfied-desire-Frank dominates over unsatisfied-desire-Frank (in the sense it has all of the benefits of the other with no downsides), almost no voters in the parliament of Frank would rank unsatisfied-desire-Frank over the alternative. So, a frustrated Frank is not equivalent to a satisfied Frank.

Of course, Rosati's larger issue is that the idealization process does not tell me what is good for me, but it instead tells me what is good for someone like me. For the multiple-agent idealization process we might say that it does not tell me what is good for me, but instead tells me what a group of people like me would vote for. However, put this way, the objection seems to beg the question. For example, we might imagine someone complaining that hedonistic utilitarianism does not tell you what is good for people, it tells you what gets you the most pleasure. To this complaint a hedonistic utilitarian would respond "of course, that's what I mean by *good.*"

Rosati's original argument against the single-agent idealization process was not question begging because it was accompanied by a strong intuition about the strangeness of the idealized person. An idealized person would be very different from me, so it seems mysterious that the opinions of an idealized version of me should determine what is good for me. Yet, this argument seems to be weaker for the multiple-agent version of the informed desire account. It does not seem mysterious what Tom′ or Tom″ would want. Nor does it seem mysterious how the parliament of Tom will vote in certain cases. This is not to say that one does not pay an intuitive cost for holding an admittedly unusual and complex view, but it is unclear that the cost is higher enough to warrant rejection.

The multiple-agent version also solves the problem of no person being able to be fully informed. The multiple-agent view does not specify a privileged person who must be both kind and mean at the same time. Instead, the kind version of me and mean version of me will plead their case before all the other versions of me who will consider each option and rank them. Each version of me will have his own biases and traits, but the idea that a vote could be reached does not seem impossible.

3.4.2.10. Arguments against multiple-agent idealization processes for informed desire satisfaction accounts of wellbeing

Naturally, this alternative means of specifying the idealization process is not without its own objections. In this section I lay out some of those objections and show why I do not think the objections represent serious problems for the view.

One possible line of objection is that the parliamentary model is shaped too heavily by the starting place for the idealization process. For example, in the case of Ted, each of the members of parliament begins as Ted and adopt Ted's current attitudes and beliefs and then experience possible lives as deviations from that starting point. But, it seems possible that Ted could be a corrupt starting point. Ted could be addicted to drugs, misogynistic or morally compromised. Surely it is not good for Ted to be a merely extrapolated version of corrupted self.

For example, imagine Ted is a psychopath devoid of any significant empathy for others. Ted currently abuses his wife and we want to know if it is good for Ted to stop. To determine this, we might imagine that members of the parliament of Ted undergo experiences that would usually produce empathy towards women. For example, Ted′ might experience having his

mother abused or might experience lives where he is abused himself. However, if Ted is already psychopathic, this experience is likely to have no effect. Ted simply cannot "see" the empathy that the experience is intended to create. As a result, the parliamentary model might hold that beating his wife is not bad for Ted. This seems problematic.

However, I do not find this objection particularly troubling. First, remember than an account of wellbeing concerns what is good for Ted, not what is good all-things-considered. It seems possible that what is good for Ted might be very bad for others. I think such examples are simply the cost of internalism. If what is good for a person depends on their actual desires in any meaningful way, then occasionally we might find a desire that seems odd *to us*.

A second line of objection originates from the voting process itself. The parliamentary model is intended to reliably produce a single answer for what is best for a person. However, there are well known paradoxes in voting that might render this impossible. For example Arrow's Impossibility Theorem shows that when voters have three or more distinct options, no rank order voting system can convert the ranked preferences of individuals into a coherent ranking while also meeting the criteria of unrestricted domain, non-dictatorship, Pareto efficiency, and independence of irrelevant alternatives.[237] The voting system proposed here does not manage to avoid this issue and so falls victim to the impossibility theorem. A related problem is that there might be voting stalemates. For example it might be that the parliament of Ted casts an equal number of votes for A as not-A for a set of mutually exclusive actions. This creates an apparent paradox.

[237] Kenneth Arrow, "A Difficulty in the Concept of Social Welfare," *Journal of Political Economy* 58 (1950).

One way to resolve at least part of the paradox is to allow for ties. That is, if the parliament of Ted recommends A and not-A then we can conclude that either option produces the same amount of wellbeing for Ted. However, allowing ties does not solve the general Arrow's Theorem issue as Arrow's Theorem already allows for the possibility of ties. There are a number of methods of "getting out" of Arrow's paradox by rejecting some of its condition. A full discussion of the paradox and methods out of it is beyond the scope of this dissertation. However, what I can say is that I am willing to bite the bullet and accept that the multi-agent idealization account violates Arrow's theorem in the way any voting system violates the theorem. I do not find this fact to be particularly troubling. All accounts of wellbeing seem to bite the bullet somewhere and many seem to have trouble with certain extreme counterexample cases. It is logically possible for a paradoxical voting case to occur but I think the account is promising enough at resolving other serious issues with informed desire satisfaction accounts that this is a cost worth paying.

3.4.2.11. Conclusion for informed desire satisfaction accounts of wellbeing

In this section I discussed informed desire satisfaction accounts of wellbeing. I showed that while these accounts solve some of the issues with actual desire satisfaction accounts of wellbeing, they are not without their own unique set of issues. The most prominent kind of objection to these accounts concerns the idealization process itself. It seems that no single agent can have a privileged epistemic position and so the account is a non-starter. I argued for a multiple-agent idealization process that can avoid these problems while still providing agents

with an informed desire. I think this provides a promising alternative for the defender of the informed desire account.

3.5. Conclusion for chapter 3

In this chapter I considered the question of what we should nudge people towards. To answer this question I investigated the three most prominent accounts of what makes a person's life go better: hedonism, objective accounts and desire satisfaction accounts. I ultimately conclude that the best account of wellbeing is an informed desire account which utilizes a multiple-agent parliamentary model of the idealization process to get around objections raised to the single-agent model.

It is important to remember that this chapter only concerned accounts of *wellbeing* as opposed to considering accounts of *normative ethics*. From the fact that an action would increase wellbeing it does not follow that one ought to do it without some additional philosophical infrastructure. That infrastructure is called welfarism. Crisp explains it as follows:

> Well-being obviously plays a central role in any moral theory. A theory which said that it just does not matter would be given no credence at all. Indeed, it is very tempting to think that well-being, in some ultimate sense, is all that can matter morally. Consider, for example, Joseph Raz's 'humanistic principle': 'the explanation and justification of the goodness or badness of anything derives ultimately from its contribution, actual or possible, to human life and its quality' (Raz 1986, p. 194). If we expand this principle to

cover non-human well-being, it might be read as claiming that, ultimately speaking, the justificatory force of any moral reason rests on well-being. This view is *welfarism*.[238]

Arguing for welfarism is beyond the scope of this dissertation. However, I find it to be an extremely plausible view. In any case, I think it is reasonable to assume a close connection between wellbeing and moral decision making and thus, I think choice architects should, in general, attempt to maximize the wellbeing of choosers.

[238] Crisp, "Well-Being."

Chapter 4: When is a Nudge Morally Preferable?

4.1 Introduction

In chapter 2 of this dissertation I considered the question of when a nudge is morally acceptable. In that chapter I considered the possibility that nudges are never morally acceptable and I considered some factors that might determine whether a particular nudge is acceptable. In the third chapter I considered the question of what we ought to nudge people towards. Or, put another way, I considered the question of what value theory ought to be followed by choice architects.

In this chapter I will consider the question of when a nudge is the best available tool. That is, imagine a case where it seems acceptable to either nudge a person or to attempt to rationally persuade them. When should we opt for the nudge? When should we opt for rational persuasion? In this section I will compare the nudging option to its three main competitors: hard paternalism, libertarianism and rational persuasion. As I will define the terms, hard paternalism involves removing the ability to do otherwise through sanctions, penalties or removing options; libertarianism involves non-interference; and rational persuasion involves providing reasons and arguments in favor of a particular behavior. To distinguish between the four, imagine a government wants to increase the number of organ donors in the country so that the organs can be used to save lives. Below is an example of how the government might react based on each approach.

Hard Paternalism – the government forces all citizens to register for the organ donor list or face consequences like fines or jail time.

Libertarianism – the government does nothing and lets people donate their organs if they so choose or, alternatively, the government set up a market to allow people to sell organs if they desire.

Rational Persuasion – the government provides information pamphlets that outline the benefits of organ donation and argue for why signing up to be an organ donor is a good thing to do.

Nudging – the government automatically opts all citizens into the organ donation program. Citizens can easily remove themselves from being an organ donor by filling out some simple paperwork.

In what follows I will define the key alternatives to nudges more clearly and I will outline the circumstances under which each alternative is preferable.

4.2 Nudges and hard paternalism

4.2.1. The difference between nudges and hard paternalism

In this section I will consider when a choice architect should opt for a nudge and when a choice architect should opt for hard paternalism. First, however, let me distinguish between nudges and hard paternalism more clearly. Recall that my definition of a nudge as outlined in chapter 1 is as follows:

> Nudge: A nudges B when A intentionally makes it more likely that B will φ, primarily triggered by B's shallow cognitive processes, while A's influence preserves B's choice-set.

The primary difference between a nudge and hard paternalism is that hard paternalism does not preserve B's choice set while nudges do preserve the choice set. For example, here is how Thaler and Sunstein define one condition on whether an action is a nudge.

> The Choice-Set Preservation Condition. A preserves B's choice-set when the choice-set is unaltered or expanded compared to a baseline representing B's situation prior to A's influence attempt.[239]

The Choice-Set Preservation Condition does help us identify certain cases as hard paternalism. For example, imagine I lock you in a jail cell. You are now physically incapable of engaging in a large number of activities that you might have otherwise engaged in. So, in that sense, I have not preserved your choice set. However, this does not seem to be the usual case we are concerned with. That is, some cases that seem like hard paternalism do not involve physically straining another.

In fact, while this definition tells us something it does not answer the most important questions at hand. Namely, it does not tell us what precisely a choice set is or when something has been removed from a chooser's choice set. An example may help reveal the intuitions. In the 1980's many states wanted to increase the use of seatbelts in motor vehicles, so they enacted laws that define penalties if a driver is caught driving without using a seatbelt. Such a law seems like a clear case of hard paternalism. It forces the driver to wear a seatbelt because it is best for the driver. However, it is not so clear that such a law violates the Choice-Set Preservation Condition. A driver is still entirely able to choose not to wear a seatbelt, so, in some absolute sense all of the same choices are available to the driver.

---

[239] Sunstein and Thaler, "Nudge: Improving Decisions," 15.

One potential way to understand why the seatbelt case does not preserve the driver's choice set is to look at some of the work done on coercion since coercion is typically thought to be a paradigmatic case of removing options for a chooser. In an influential essay, Nozick argues that P coerces Q if and only if:

1. *P* aims to keep *Q* from choosing to perform action *A*;
2. *P* communicates a claim to *Q*;
3. *P*'s claim indicates that if *Q* performs *A*, then *P* will bring about some consequence that would make *Q*'s *A*-ing less desirable to *Q* than *Q*'s not *A*-ing;
4. *P*'s claim is credible to *Q*;
5. *Q* does not do *A*;
6. Part of *Q*'s reason for not doing *A* is to lessen the likelihood that *P* will bring about the consequence announced in (3).[240]

However, even on this analysis of coercion, it still seems that *Q* still has the option to do *A*. One potential way of explaining why coercion removes options is to say that in the coercion example above, *Q* no longer has the option to both do *A* and avoid the consequence announced in (3). Therefore, an option was removed for *Q*. The problem with this line of reasoning is that it includes cases that do not seem to remove options. For example, imagine I am happily eating a steak when someone explains to me that animals have suffering and that by buying my steak I am complicit in a system that causes animals undue suffering. This seems like a paradigmatic case of rational persuasion. Yet, it removes the option for me to both enjoy my steak and not have to think about animal suffering.

[240] Robert Nozick, "Coercion," in *Philosophy, Science and Method: Essay in Honor of Ernest*, eds. Ernest Nagel, Sidney Morgenbesser, Patrick Suppes, and Morton White (New York: St. Martin's Press, 1969) 441-445.

A different way to think about the distinction between hard paternalism and nudging is that nudges generally do not change the relevant facts of the situation whereas hard paternalism does. In the seatbelt case, for example, the introduction of seatbelt laws actually changes the driver's reasons for wearing a seatbelt in the sense that now the driver has the reason of "avoid a ticket" for wearing a seatbelt. Nudges do not change the chooser's reasons in this way. For example, a nudge to increase seatbelt use would be to use a dashboard indicator light when one's seatbelt is not on. This option does not change the driver's reasons for wearing a seatbelt.[241]

Similarly, this idea of changing a chooser's reasons distinguishes hard paternalism from rational persuasion. In rational persuasion, one does not change the chooser's reasons, but instead merely calls them to attention. In the case of arguing about animal suffering over steak, the fact of animal suffering has not been changed, instead, the fact of animal suffering has merely been called to attention.

In conclusion, two related but separate cases constitute hard paternalism as I am using the term. The first is a case where a chooser is physically prevented from being able to make some choice. The second is a case where the chooser's reasons are changed by the intervention of the choice architect. Examples of this second type would include penalties and sanction for performing or not performing certain actions. Also worth noting is that for my purposes I will only consider cases where the choice set available to the chooser is made worse as a whole through either physical restriction of freedom or by making some choice more costly in terms of time, trouble, social sanction and so forth. It is also possible to encourage some behavior by

[241] They might give the driver a reason of "get rid of annoying sound" for wearing a seatbelt, but I am assuming these noises disappear after a short period of time.

making a specific choice more appealing through incentives. I will briefly show how my analysis of hard paternalism applies to this case, but that will not be the focus of this section.

4.2.2. Moral considerations relevant to the use of hard paternalism

In this section I will consider some of the moral considerations that are relevant to the use of hard paternalism. In particular, I will argue that while both nudging and hard paternalism are *pro tanto* morally wrong, the *pro tanto* wrongness of hard paternalism is greater than the *pro tanto* wrongness of nudging. Thus, in a case where a choice architect wants to create some desirable outcome, the choice architect should only opt for hard paternalism instead of a nudge if there is a high degree of certainty that the hard paternalist intervention will produce the desired outcomes and if there is good reason to believe that other options like rational persuasion or nudging are unlikely to be effective.

4.2.2.1 The *pro tanto* wrongness of hard paternalism

In chapter 2 of this dissertation, I argue that nudges are manipulative and that they get a chooser to act in a way that does not necessarily track the chooser's reasons. As a result, they are *pro tanto* morally wrong. However, I also argue that this wrongness can be outweighed relatively easily by other considerations like the beneficial effects of particular nudges on choosers. In this section I will also argue that hard paternalism is *pro tanto* morally wrong, but I will not argue that this wrongness can be so easily outweighed. As a result, the degree of *pro tanto* wrongness is much greater in the case of hard paternalism and so, choice architects need stronger reasons to engage in hard paternalism than they do to engage in a nudge.

As I argued earlier, hard paternalism involves two related cases. In the first, a chooser's options are restricted in the sense that the chooser is physically prevented from choosing some option. For example, imagine I want to prevent some person from abusing recreational drugs. I might isolate them in a jail cell in order to make it impossible for them to use the drugs. There are a few ways of specifying why this is *pro tanto* wrong. The first is that such a restriction of options often causes pain to the chooser. This occurs in two ways. First, the mere fact that one's options are restricted in an unusual way can cause alarm or anger. For example, if I am told that I must eat hotdogs for lunch every Wednesday, I might find this fact troubling even if I happen to like eating hotdogs on Wednesdays simply because it restricts my options. Second, restricting options prevents individual from pursuing options that might more effectively bring them pleasure or lessen their pain. For example, if I am forced to eat hotdogs on Wednesdays, I might find that on Wednesday that I would get more pleasure by eating a burrito. But, since I am forced into a particular course of action, I cannot make the choice that might maximize my pleasure. In addition, restricting options is *pro tanto* wrong because it takes away something that most people find valuable. This is independent of any pain that the restriction causes.

In the second kind of hard paternalism case, an option is made more costly in terms of time, trouble, social sanction and so forth. We can divide our analysis of this case into two possibilities: if the sanction is imposed because the choosers opts to do something other than what the choice architect intends, and cases where the sanction is not imposed and the chooser does as the choice architect intends. If the sanction is imposed, the wrongness is obvious. Namely, it is *pro tanto* wrong to cost someone time, trouble or the like. A more interesting case is one in which the sanction is never actually imposed. For example, imagine I run a company

and I want my employees to work harder. I might threaten to dock the employee's pay if they are caught leaving work before 5pm. Imagine that none of my employees leave before 5pm and, accordingly, no one's pay is docked. Have I done something *pro tanto* wrong in this case?

One way that this might create a *pro tanto* wrong is that is might cause an employee to forego choosing an option that would have been best absent my penalty. For example, imagine my employee has an illness and has a doctor's appointment that requires her to leave work early. It might be better for my employee that she go to the doctor than stay at work. However, given the sanction, it might be better than she stay at work than go to the doctor and face a penalty. So, in this case, the option that is best for the employee – going to the doctor with no penalty – is removed. So, imposing sanctions may cause the choosers to select from among options that are not optimal for the particular situation.

Yet, we can imagine that in some situations the sanction never takes effect and the chooser does not forego options that would be best were it not for the sanction. In such a case, I think the wrongness of hard paternalism is similar to the wrongness of nudging. Namely, it bypasses the ability of the chooser to exercise her autonomy and it subjects the chooser to the will of the choice architect.

4.2.2.2 Considerations relevant to the use of hard paternalism

In the previous section I argue that both nudges and hard paternalism are *pro tanto* wrong, but that the *pro tanto* wrongness of hard paternalism is, in general, greater than that of nudges. This fact places a special burden on the use of hard paternalism that is greater than the burden placed on nudging. Namely, because the costs are greater in hard paternalism than in nudging, it must

also be the case that the benefits are greater in order to justify the use of hard paternalism. In general, I think hard paternalism can be justified if a few conditions are true. I identify those conditions below.

4.2.2.3 Hard paternalism to create universally-desirable actions

First, in cases of hard paternalism, the desired behavior should be universally or nearly-universally desirable. This is because otherwise hard paternalism runs the risk of eliminating an option that would be best for the chooser absent the intervention or of forcing the chooser to continue to make the same choices as they would prior to the intervention, but with an added burden. Consider 401(k) savings as an example. We might imagine that a choice architect imposes a $500 penalty on those employees who fail to save for retirement through the company 401(k) system. While saving for retirement is usually best for an employee, we can imagine many plausible exceptions. For example, imagine an employee has a large amount of credit card debt and would gain more financial security and peace of mind by paying the debt off instead of saving for retirement. Or, imagine the employee has no emergency savings such that a routine car repair would represent a financial emergency. In either of these cases, a penalty simply makes the situation much worse for the employee. Neither of these cases seem particularly far-fetched, so the benefit of retirement savings probably does not meet the universality necessary for a hard paternalist intervention. On the other hand, seatbelts might meet this universality requirement. Seatbelts place such a small imposition on the driver that it seems far more difficult to imagine a situation where one is better off not wearing one. As a result, seatbelts might be a better case for using hard paternalism.

4.2.2.4 Hard Paternalism when other options are unavailable

A second condition is that hard paternalism should only be used when other options for behavioral change will not accomplish the intended goals. For example, if one can achieve the same rate of seatbelt wearing through rational persuasion (e.g. by providing drivers with a pro-seatbelt leaflet), then choice architects should opt for rational persuasion instead. This argument is similar to the argument I made in chapter 2 and that I will make later in this chapter about the relationship between rational persuasion and nudging. I argue that nudging is a small *pro tanto* moral wrong because it is manipulative. Yet, the wrongness of the nudge can be easily outweighed by other considerations. However, rational persuasion is not coercive and so, all other things being equal, a choice architect should opt for rational persuasion. Similarly, hard paternalism is a *pro tanto* moral wrong, but in general, the wrongness of hard paternalism is more difficult to outweigh. This is because hard paternalism tends to be more coercive, tends to decrease chooser's ability to act otherwise and tends to impose costly sanctions. Therefore, if the outcomes of an intervention can be achieved through a nudge or rational persuasion, the choice architects should not opt for hard paternalism.

One way that nudging or rational persuasion might not work is a case where the decision making calculus does not work out in favor of the desired action. The standard discussion of nudges usually focuses on cases where we presume that some action is in the chooser's best interest and we want to ensure that the chooser selects this action. However, we might imagine cases where an action is desirable but it is not currently in the chooser's best interest. An example will reveal the intuition.

Imagine a large automaker is debating what to do about a defective part on one of their cars. The executives presume that the goal of the company is to make money for their shareholders, so they calculate how much money a recall of the part would cost versus how much they expect to lose in lawsuits should they let the defective part stay on the road and cause injuries. They see that it is cheaper to face the lawsuits than recall the part, so the parts stay on the road and people are injured. In this case, the socially desirable response is that the automaker recalls the parts. Yet, it does not appear that either rational persuasion or nudging is going to get the automaker to choose the recall because it is not in the best interest of the company.[242] So, one might need to change the variables that go into the calculation to ensure that the company makes the socially desirable decision. In this case, altering the calculation can involve financial sanction or criminal prosecution by the government to make the costs of failing to recall the parts higher than it would have otherwise been and thus make the rational decision to recall the parts.

4.3 Nudges and libertarianism

In this section I will discuss the circumstances under which a choice architect should opt for libertarianism instead of a nudge. In this section I will use the term libertarian in a slightly unusual sense. I will not use libertarian in the way that it is most often used in political philosophy. Instead, by "libertarian" I will mostly be referring to cases where the choice architect does nothing in the sense of not favoring any particular choice from among those available. The next section briefly explains the distinction between nudges and libertarianism.

[242] Here I'm just granting that the only goal of a company is to make money to simply the situation. It may be the case that companies ought to focus on other goals as well.

4.3.1 The difference between nudges and libertarianism

As an example of the difference between nudges and libertarianism, imagine a choice architect is interested in retirement savings. We might imagine a case where a choice architect might nudge employees to save for retirement through automatically opting in their employees to the company 401(k) plan. But, the choice architect might remain neutral on which investment options employees ought to pick. We might imagine that they provide employees with a neutrally-designed pamphlet that contains a list of options. Thus, the choice architect is nudging them towards savings, but is libertarian about the particular savings options. This distinction exists because the choice architect has the intention of altering behavior in the first case, but not intention of altering behavior in the second case.

Astute readers of Thaler and Sunstein might argue that libertarianism is impossible in the sense that there is no neutral framing. Going back to the 401(k) example, one might argue that no matter how you present the investment options, you are going to create some effect on which options the choosers select. So, there is no such thing as a neutral option and therefore, no libertarian approaches to these issues. To avoid this issue, when I say that an intervention is libertarian, I am mostly interested in the choice architect's intent. If the choice architect has no preference for which option the chooser should be more likely to choose and makes no attempt to frame the choice such that a particular option is chosen more frequently, then I would say that the choice architect is behaving in a libertarian way.

Of course, the line between libertarianism and nudges may not always be so clear. For example, we can imagine a situation where a choice architect, call her Jill, thinks that it would be best if employees saved their retirement income through so-called Target Date Retirement Funds

which automatically create a reasonable asset allocation based on an employee's intended date of retirement. To encourage employees to select these options, Jill places them at the beginning of the investment pamphlet where employees will see them first. On the other hand, Sam is not concerned with which retirement options employees choose. Sam places the retirement option in the pamphlet at random and by sheer luck she creates a pamphlet identical to Jill's. On my interpretation of the difference between nudging and libertarianism, Sam was libertarian whereas Jill nudged even though they ended up with precisely the same outcomes. While this is an interesting result, I do not think it is an objection to libertarianism as I am conceiving of it. When a choice architect engages in libertarianism she has no reason to suspect that the framing will impact the chooser in any particular way. It might be the case that by sheer luck she frames the choice in a way that creates some outcome, but this is merely a coincidence.

Cases of required active choosing raise similar concerns. Required active choosing occurs when the choice architect forces the chooser to select an option instead of allowing for a default option. For example, in the case of 401(k) enrollment, there are at least three options. One might automatically opt employees out of the 401(k) plan and require them to complete some paperwork to participate. You might automatically opt employees into the 401(k) plan and require that they complete some paperwork to cease participation. Or, you might require active choosing such that all employees are required to choose between opting in and opting out. The required active choosing could either be a nudge or libertarianism depending on the intentions of the particular choice architect. It could be a nudge because a choice architect might decide to mandate required active choosing because they know that participate rates in the 401(k) plan under that condition are higher than those under the more common opt-out condition. It could be

libertarian if the choice architect does not have a preference for what employees choose and instead requires a choice because it seems like the most neutral option. Below I outline when this is the best option for the choice architect.

4.3.2 Libertarianism as the default option

In this section I will argue for a general framework according to which libertarianism can be thought of as a kind of default option. That is, when a choice architect has insufficient evidence in favor of any choice being preferable or does not have any evidence that a particular intervention will change the choices of choosers, then the choice architect should opt for libertarianism. In this section I will argue for this view briefly. In the next section I will also argue for some particular conditions under which choice architects should opt for libertarianism.

It is tempting to think of libertarianism as an option with no moral costs and no moral benefits. It has no moral costs because, unlike nudging or hard paternalism, it is not *pro tanto* morally wrong to opt for libertarianism and unlike hard paternalism, it does not impose any obvious costs on choosers. It has no benefit because it does not attempt to make choosers better off. Accordingly, we can think of libertarianism as a kind of default option. On this view, therefore, libertarianism is the default approach which is then altered depending on the evidence from the particular situation at hand.

There are two lines of objections to this view. The first is a Thaler and Sunstein style argument that there is no such thing as a neutral framing. No matter how a choice is set up, it will always have some effect on the choices made by the chooser. The second, related argument, is that all option must be good or bad compared to *some other option*. So, it may not make sense to

claim that libertarianism has no moral cost or benefit if one must compare it to some other option like a nudge. I think we can deal with both of these objections with a single argument. Even if we grant the claim that there is no neutral framing, I do not think this serves as an objection to the use of libertarianism as a default position. The particular framing of choices in libertarianism will have some effect, but that effect will be essentially random. Since the question in choice architecture is 'how ought choice architects to set up choices for choosers,' we might think that random occurrences of this type are not open to moral scrutiny. That is, if a choice architect's actions create an unintended and unforeseen negative consequence, this does not impact the moral worth of the action. So, even if it is the case that all framings have consequences, it might be the case that intending and foreseeing some particular consequence is morally relevant whereas neither intending nor foreseeing the consequence is morally irrelevant. Therefore, it is plausible that libertarianism can be seen as a kind of default option.

4.3.3. Libertarianism under conditions of uncertainty

In the previous section I argued that libertarianism can be thought of as a kind of default option when compared to other options like persuasion, nudging and hard paternalism. In this section I will argue for a particular implication of this view, namely, that libertarianism is preferable under conditions of uncertainty.

If libertarianism is the default option, then it seems to follow that libertarianism is preferable under conditions of uncertainty. In the case of choice architecture, there are two kinds of uncertainty. First, there is uncertainty about which option is better for choosers. This kind of uncertainty can be further divided into two types. The first type is uncertainty about the

circumstances that choosers actually face. Take the 401(k) enrollment case as an example. As I argued above in section 4.2.2.3, there are some circumstances under which it might not be in the best interest of chooser to enroll in a company 401(k) plan even given a generous employer match. For example, it might be a more responsible idea for employees to pay off high-interest credit card debt or create an emergency fund. If a choice architect is uncertain about the actual circumstances faced by most employees – namely, one does not know what percentage of employees would be better off not participating in the 401(k) – then, the choice architect might opt for libertarianism. The second kind of uncertainty about which option is better for choosers is uncertainty about value. That is, if a choice deals with incommensurable values or value judgments where many people disagree, then it might be preferable to be libertarian. As an example, it might be good to be libertarian about how employees spend their leisure time given that there is substantial individual differences in what people find valuable.

The second type of uncertainty is uncertainty about whether a particular intervention will be successful, even given that some particular choice is preferable. For example, we might know that Target Date Retirement Funds are better for employees, but we might not know what interventions will be best at getting employees to select that option. Under such conditions we can rule out hard paternalism because it often imposes a cost (e.g. sanctions) and such a cost should not be levied without a reason to suspect that the intervention will improve outcomes. We can also rule out nudging because nudging too imposes a moral cost (albeit a small one). So, we are left with either rational persuasion or libertarianism. As I will argue later, rational persuasion is morally costless (e.g. it does not constrain chooser autonomy), so it may be worth attempting to rationally persuade choosers even in absence of evidence that rational persuasion will be

effective. Whether this is a viable option depends in large part on the specifics of the behavioral intervention in question. If the choice architect is a government, for example, disseminating a rationally persuasive message is not costless as it requires designing such a message, marketing it and so on. However, if I am trying to change the behavior of a friend or acquaintance the only cost might be minor social discomfort, so rational persuasion might be worth attempting. In cases where rational persuasion is not costless, then choice architects ought to prefer libertarianism instead.

4.4 Nudges and rational persuasion

In this section I will consider when a choice architect should opt for rational persuasion instead of a nudge. I will argue that rational persuasion is morally costless in the sense that it is not coercive and does not violate autonomy, yet it can also achieve positive outcomes. As a result, rational persuasion is preferable to nudges all else being equal. Nudges should only be preferred where there is evidence that a nudges would be much more effective than rational persuasion at reaching a positive outcome.

4.4.1 The difference between nudges and rational persuasion

In this section I will explain the difference between nudges and rational persuasion. Recall that my definition of a nudge is as follows:

> Nudge: A nudges B when A intentionally makes it more likely that B will φ, primarily triggered by B's shallow cognitive processes, while A's influence preserves B's choice-set.

The key distinguishing characteristic from the definition is the term "shallow cognitive processes." Nudges rely on the automatic/heuristic reasoning systems (i.e. system 1 in Kahneman's parlance) to make it more likely that B will φ. Rational persuasion on the other hand relies on deep cognitive processing as it requires choosers to consider and accept reasons. A related way to think about the difference between nudging and rational persuasion is that nudging bypasses or subverts an agent's rational capacities, where "rational capacities" are defined as follows:

> Rational Capacities: those capacities that enable agents to assess and revise their beliefs in accordance with the basic canons of logic; to evaluate their epistemic and practical options against criteria generated by their beliefs, values and preference sets; to make adjustments to these beliefs, values, and preference sets in light of new information; and to act in accordance with their judgments about what they have most reason to do.[243]

Those activities that engage these capacities are rational persuasion whereas those that do not are probably a form of manipulation (e.g. nudging).

However, the situation is more complicated than it might first appear. Many (perhaps all) cases of rational persuasion also involve some degree of appeal to one's shallow cognitive processing. In fact, it seems possible for choosers to be both rationally persuaded and nudged at the same time. For example, imagine an older male patient has been diagnosed with prostate cancer. His doctor wants to persuade him not to undergo surgery because it is very unlikely that the prostate cancer will spread and cause any long-term health issues. Yet, in the context of a conversation about the medical outcomes of surgery, how the doctor describes the alternative to

[243] Gorin, "The Nature," 30.

surgery could be a nudge. For example, the doctor might describe the options as either surgery or doing nothing, or she might describe the options as surgery or "watchful waiting" or "active surveillance." While the substance of the conversation might be about medical facts, framing the choice as one between two *treatment protocols* and not as a choice between *doing something and doing nothing* can have a powerful, non-rational influence on the patient.

In this sense, anything a choice architect does, even in the context of rational persuasion could also become a nudge because no matter how the rational persuasion is delivered there is an opportunity to nudge choosers as well. In my view, the key difference comes from the intent in each case. Typically in rational persuasion the persuader is motivated by more than just achieving the desired ends (e.g. that the chooser has some particular view). Instead, the rational persuaded is also motivated by the independent value of the chooser arriving at the ends in a particular way. That is, the rational persuader wants others to accept the view for the reasons the rational persuader has for accepting the view. Whereas the manipulator only has the goal of getting others to accept the view even if they do it for the wrong reasons or for no reasons at all. Therefore, the intent in rational persuasion and cases of manipulation like nudging is very different and this difference in intent distinguishes the two kinds of cases.

4.4.2 Rational persuasion as a morally costless action

I take it as a common view (especially among philosophers) that rational persuasion is an acceptable, even preferred form of behavioral change. In this section I will sketch out why one might hold this view.

One way to see why rational persuasion is preferred is to reflect on what gives an action moral worth. The distinction between the moral worth and the moral rightness of an action was famously made by Kant. For example, Kant argued that a merchant who does not overcharge new customers is morally right, but that does not mean that his actions have moral worth. If, for example, the merchant's only reason for not overcharging is that he does not want to develop a bad reputation, then his action would lack moral worth. Kant extends this argument to say that only those actions performed from the motive of duty have moral worth whereas those action motivated by the desire to make others happy or better off do not.

Markovits calls Kant's view that an action has moral worth if and only if it is performed because it is right the *Motive of Duty Thesis*.[244] This view has the counterintuitive implication that acting on a selfless desire to help others lacks moral worth. To help salvage the view Markovits proposes the *Coincident Reasons Thesis* which holds the following:

> *My action is morally worthy if and only if my motivating reasons for acting coincide with the reasons morally justifying the action*—that is, if and only if I perform the action I morally ought to perform, for the (normative) reasons why it morally ought to be performed.[245]

On this view, moral worth is about a relationship between the reasons one ought to do something and the reasons one does it. Gorin[246] extends this account of moral worth to a more general account of normative worth, which applies in any case where it is appropriate to explain or

[244] Julia Markovits, "Acting for the Right Reasons," *Philosophical Review* 119, no. 2 (2010).
[245] Gorin, "The Nature," 205.
[246] Ibid., 115-116.

justify a behavior. He argues that an action has normative worth "*if and only if the motivating reasons that explain the behavior coincide with the reasons that justify the behavior*."[247]

Whatever one thinks of Markovits and Gorin's view on moral and normative worth, their work does seem to point to an important intuition, namely, that it is better to do something for the right reasons than to do the same action but for reasons unrelated to the reasons one ought to do it. Part of the intuition surely stems from the fact that if I act correctly, but for the wrong reasons, I am much more likely to act incorrectly next time than if I act correctly for the right reasons. But, the intuition does not seem to be limited to this fact. It also seems to be the case that, all else equal, acting for the right reasons gives an action more worth than acting for the wrong reasons.

Extending this example to the case of rational persuasion and nudging, if I nudge a chooser towards some action, I cause the chooser to undertake the action for reasons other than the reasons the chooser ought to do the action. For example, if I nudge employees to save for retirement via an opt-out 401(k) plan, then the employees who save, might do so *because it was easier* which is not the reason one ought to save. On the other hand, if I engage in rational persuasion, and if choosers save, they do so because of the reasons that one ought to have for saving. On this line of reasoning, all else being equal, it is better to rationally persuade someone than to nudge them.

This view is complicated by the difficult metaphysics of what it means for a reason to be the "motivating reason." On this view, rational persuasion is preferable to nudging because nudging causes choosers to act for reasons other than the ones that should govern their behavior.

---

[247] Ibid., 116.

The fact that choosers are not acting for the right reasons is inferred from the fact that the behavior of choosers changes under various nudges Imagine Tom is not participating in the company 401(k) plan. I institute a nudge that automatically opts Tom into the plan unless he completes some paperwork to get out of the plan. Tom is aware of the change, he does not complete the paperwork and so he participates in the 401(k) plan. It is tempting to say that the "motivating reason" that Tom is in the 401(k) plan is because of my nudge. But, this does not quite follow. We can see this because it seems very plausible that if we altered the details of the 401(k) plan Tom would not participate despite the fact that I nudged him. For example, imagine the 401(k) plan was particularly poor. Imagine it offers no employee match, no return on investment, and has high fees. If we think it is plausible that Tom would not participate in such a plan even if he was nudged, then the desire to save money is part of Tom's motivating reason. The upshot of this line of reasoning is that nudging a chooser to φ does not necessarily mean that the motivating reason that the chooser φ's is the nudge. However, a chooser who φ's because of rational persuasion always has the motivating reasons that ought to govern her action.

4.4.3. When should a choice architect not opt for rational persuasion?

In this section I will consider the question of when a choice architect should opt for rational persuasion as opposed to the other available options. My general framework for this discussion is that, all else being equal, one should opt for rational persuasion. In this section I will define some ways in which all else might not be equal and, accordingly, one might not opt for rational persuasion.

4.4.3.1 Barriers to rational persuasion

One circumstance under which a choice architect might opt for a nudge or hard paternalism instead of rational persuasion is a case where a clear barrier to rational persuasion is present. Put another way, if, absent the intervention of the choice architect, the chooser is going to φ for reasons other than those that she ought to φ, then rational persuasion may not be the best option.

A few types of cases come to mind. One type of case is one where a clear cognitive bias is present in the decision. For example, a study by Danziger et al.[248] found that the percentage of favorable rulings in va Jewish-Israeli parole hearing "drops gradually from ≈65% to nearly zero within each decision session and returns abruptly to ≈65% after a break," suggesting that how long it has been since the judge ate constitutes an extremely important variable in legal decisions. This represents a fairly obvious cognitive bias because no one can reasonably argue that when the judge last ate ought to be an important variable in the decision. In such a case, it might be appropriate to nudge the judges to counteract this bias e.g. by providing food within reach of the judges as they make their decisions.

Another type of case is one where the chooser is not reasons-responsive. For example, imagine a jilted lover who contemplates suicide. Rational persuasion is unlikely to change his mind, so more drastic action (in this case, hard paternalism to prevent the suicide) is called for. However, this kind of case need not be restricted to only drastic examples. In some cases, the reasons in favor of some action are sufficiently complex, that most agents would not be able to respond sufficiently to the reasons. One interesting examples of this came in 2001 when the FAA

[248] Shai Danziger, Jonathan Levav, Liora Avnaim-Pesso, "Extraneous Factors in Judicial Decisions," *Proceedings of the National Academy of Sciences of the United States of America* 108, no. 17 (2010).

was considering making it mandatory that children under 2 be in car seats on airplanes. The American Academy of Pediatrics endorsed the potential move, saying in a statement:

> Occupant protection policies for children younger than 2 years on aircraft are inconsistent with all other national policies on safe transportation. Children younger than 2 years are not required to be restrained or secured on aircraft during takeoff, landing, and conditions of turbulence. They are permitted to be held on the lap of an adult. Preventable injuries and deaths have occurred in children younger than 2 years who were unrestrained in aircraft during survivable crashes and conditions of turbulence.[249]

However, Newman et al., calculated that requiring parents to pay for an extra seat for children would cause more parents to switch to car travel instead of air travel to reach destinations.[250] Since car travel is far more dangerous than air travel, in effect, the policy would cause more children to die. Therefore, airlines should not require carseats on airplanes for children. However, this reason is unlikely to comfort a parent who wants to know how to keep their child safe on an airplane. So, while this might be the actual reason to not require or provide child seats on airplanes, providing this reason to a parent is unlikely to lead to positive outcomes and so it may be acceptable to attempt to cause the intended behavior through techniques other than rational persuasion.

4.4.3.2 Cognitive costs of rational persuasion

[249] American Academy of Pediatrics, "Restraint Use on Aircraft," *Pediatrics* 108, no. 5 (2001): 1218.
[250] Thomas Newman, Brian Johnston, David Grossman, "Effects and Costs of Requiring Child-Restraint Systems for Young Children Traveling on Commercial Airplanes," *Archives of Pediatrics and Adolescent Medicine* 157, no. 10 (2003).

While I have argued that rational persuasion is morally costless, it can be costly both in terms of time and cognitive resources. Not all choosers are in a position to consciously consider all the choices they make, and they might find it unnecessarily taxing to do so. In such cases, nudging so that the socially desirable option is easiest may be preferable to rational persuasion. This is especially useful in cases where each individual choice is relatively unimportant to the chooser, but where the aggregative of all decisions by all choosers can be meaningful. For example, it might not matter much to each individual shopper at a grocery store whether they bag their groceries in paper or in plastic. If a choice architect wanted more shoppers to select paper, rational persuasion would be unlikely to be a good strategy because for most shoppers the decision is not important enough to justify carefully weighing the costs and benefits of each option. So, nudging shoppers may be the best choice.

This kind of situation is complicated by the fact that attempts to rationally persuade choosers for trivial decisions might operate as nudges instead of rational persuasion. For example, hotels often try to encourage guests to reuse their towels to save resources. We might imagine that the hotel places a standard sign in the room that asks guests to "help save the environment" by reusing their towels. If guests reuse the towels and their motivating reason is that they want to help save the environment, then the hotel has succeeded in rationally persuading guests. But, it is not clear that the *argument* was the motivating reason. In an experiment, the hotels also tried a different message, namely that "75 percent of the guests who stayed in this room (room 313)" had reused their towels.[251] The results indicated an increase of twelve percent in reusing towels. This shows that perhaps the motivating reason in the "rational

[251] Alex Mindlin, "Dos and Don'ts of Gentle Prodding," *The New York Times*, March 17, 2008, http://www.nytimes.com/2008/03/17/business/17drill.html?_r=2&ex=1363406400&en=96da4&oref=slogin&.

persuasion" case is not saving the environment, but it instead complying with a social norm which is set by the existence of the sign. In sum, if choosers are not going to devote much cognitive resources to the decision, rational persuasion may not be the best option.

4.5 Conclusion to chapter 4

In this chapter I considered four options for behavioral change – namely, hard paternalism, libertarianism, rational persuasion, and nudging, and considered the circumstances under which each out to be preferred. We can divide these four options along two dimensions – moral cost and likely moral benefit. In terms of cost, both libertarianism and rational persuasion have no moral cost whereas nudging has a small moral cost and hard paternalism has a large moral cost. In terms of likely benefit, libertarianism generally has no benefit, nudging usually has a small benefit, hard paternalism usually has a large benefit and rational persuasion can have either a small or large benefit depending on the circumstances. Choice architects should choose the option that provides them with the most favorable combination of costs and benefits for the particular case they are considering.

Dissertation chapter 5

5.1. Introduction

Over the past four chapters of this dissertation I have been developing the machinery necessary to help choice architects determine if they should engage in particular kinds of nudges. In chapter 1, I develop a picture of what exactly a nudge is and what it is not. In chapter 2, I develop the position that nudges are *pro tanto* morally wrong but that this wrongness can be exacerbated or mitigated by a wide range of considerations. I develop a three-factor framework for evaluating the permissibility of nudges which considers the nudge type/mechanism, agent-relative considerations and ends-based considerations to develop a full picture of what factors choice architects should evaluate in determining if they ought to perform a nudge. In chapter 3, I take up the question of what it is that we should be nudging choosers towards. I develop a revised form of the informed desire account as my answer to this question and I conclude that what makes a person's life go better is the satisfaction of the desires they would have if fully informed and rational. In chapter 4, I considered the question of when a nudge is morally preferable. That is, when should we opt for a nudge as opposed to other tools like rational persuasion, hard paternalism, or libertarianism? In this section of the dissertation I apply this machinery to a few real world cases to show how it might be used in practice.

5.2. Nudging for the greater good: Charity donation nudges in the workplace

One of the most widely-discussed types of nudges is the 401(k) nudge. On this nudge, employees are automatically enrolled in their company 401(k) plan unless the employee completes paperwork to opt out of the program. This nudge has a large impact on the savings rate of

employees and this is thought to be beneficial for them. Many find this nudge to be highly beneficial and relatively unproblematic. In this section, I consider a more difficult case of whether choice architects ought to engage in nudges to get employees to donate money to highly effective charities like those recommended by GiveWell.[252] There are a number of possible variations on how a nudge of this type could be constructed. I will discuss how alternative formulation impact the assessment of the nudge, but I will take the following to be a kind of base case:

> *Charity donation nudge:* A company wants to encourage more of their employees to give to charity. To do so, they automatically enroll all employees in a program that donates 5% of their income to charity. The default charity is chosen from GiveWell's list of Top Recommended Charities which means that the default charity will be evidence-backed, thoroughly vetted, and underfunded. Employees are given clear advanced notice of the change and are provided with clear instructions on how to opt out of the new program.

5.2.1. Motivating the case

In chapter 2, I argued that nudges are *pro tanto* morally wrong. To motivate the case, I will first provide reason to think that the *pro tanto* wrongness of the nudge might be outweighed by its positive impact.

To do this, we can add a few details to the basic case. The company in question has 10 employees who each have an average income of $50,000. Without the charity donation program none of the employees would donate 5% of their income; with the program half of the employees

[252] http://www.givewell.org/

will donate. This means that the program will result in an additional $12,500 per year donated to a GiveWell recommended charity.[253]

The easiest way to get an intuitive sense of how much impact an additional $12,500 per year in donations might have, is to use GiveWell's top recommended recommended charity (as of late 2015), Against Malaria Foundation (AMF). AMF donates long-lasting insecticide treated bednets to areas in Africa at a cost of $5.80 per net. These bednets are then used by inhabitants to help prevent the spread of malaria by reducing the number of mosquito bites they get at night. Malaria causes symptoms including fever, fatigue, vomiting, headaches, yellow skin, seizures, coma or even death.[254] Children and the elderly are especially vulnerable to more severe cases of the disease. GiveWell estimates that for every $2,838 donated to AMF, a child under 5 that would have died from malaria will live.[255] This means that the $12,500 dollars donated through the nudge will save the lives of 4.4 children every single year and will prevent many others from getting the disease.

Therefore, it seems extremely plausible that the benefits caused by this nudge might outweigh the general *pro tanto* wrong of nudging. However, this is only intended to motivate the case. We need to investigate the nudge more thoroughly to determine if, all-things-considered, a choice architect ought to use it and if there might be alternative ways to achieve this end that are preferable. In the next section I discuss how this nudge relates to the question of what we ought to nudge choosers towards.

---

[253] I specify that the charity be GiveWell recommended because, to the best of my knowledge, GiveWell provides the best available charity recommendations for individual donors.
[254] Hector Caraballo, "Emergency Department Management Of Mosquito-Borne Illness: Malaria, Dengue, And West Nile Virus," *Emergency Medicine Practice* 16, no. 5 (2014).
[255] "Against Malaria Foundation," *Givewell*, December 2015, http://www.givewell.org/international/top-charities/AMF.

5.2.2. Wellbeing and improving the life of others

In chapter 3 of this dissertation, I developed an account of wellbeing according to which what makes a person's life go better is the satisfaction of the desires they would have if fully informed and rational. I also noted that answering this question does not necessarily tell us what we ought to do especially in cases where our wellbeing comes into conflict with the wellbeing of others. For example, one might be a hedonist about wellbeing but think that there are deontological side constraints such that certain methods of maximizing the total amount of pleasure in the world are never morally permissible. I intend to show how the interplay between the personal wellbeing and the wellbeing of others might play out in this case and how this interplay may practically impact the decision to nudge.

First, let us take it as a premise that the amount of wellbeing in the world would be increased by having employees in a developed country like the US donate to prevent people in the developing world from dying from malaria. I also think it is safe to specify that if the employee were to make the donation absent a nudge, the situation would be entirely unobjectionable. What might be objectionable in this case is the choice architect using a nudge to get someone else to donate to charity.

With the stipulations above, I think there are two relevant questions. First, does the nudge decrease the wellbeing of the chooser and second, if it does decrease their wellbeing, does that mean that the nudge is impermissible? I address each of these questions in turn.

5.2.2.1. Does the charity nudge decrease the chooser's wellbeing?

The charity nudge decreases the amount of disposable income that the chooser has by $2,500 dollars per year. It seems plausible that money improves a person's wellbeing so we may assume that the decrease in income for the chooser decreases their wellbeing. However, I do not think this is necessarily so. I will briefly show how accounts of wellbeing that I discarded might evaluate the matter and then I will discuss what the informed desire account might have to say.

According to hedonism what makes a person's life go better is an increase in the balance of pleasure over pain. Because money can be turned into a wide variety of pleasurable experiences, one might assume that donating to charity is bad for hedonic wellbeing. This is an empirical question and the best available evidence on this point is mixed. One noted effect is the Easterlin Paradox according to which high earnings do correlated with happiness, but increased income does not correlate with increased happiness.[256] Additionally, a study by Daniel Kahneman concluded that an increase in income had an effect on evaluation of one's life but not on emotional wellbeing.[257] Other studies have found strong correlations between reported happiness and charity donations and some evidence for a causal connection between the variables. There are plenty of reasons to be skeptical of the psychological literature on this question including whether happiness in the sense measured in the laboratory is the same as please in the sense meant by hedonism. However, the point is to indicate that it is not clear that the charity donation nudge decreases the hedonic wellbeing of choosers

According to objectivist accounts of wellbeing, what makes a person's life go better is the attainment of any of a number of a list of intrinsically good states of affairs. Let's take

---

[256] Richard Easterlin, "Does Economic Growth Improve the Human Lot? Some Empirical Evidence," in *Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz*, eds. Paul A. David and Melvin W. Reder (New York: Academic Press, Inc., 1974).

[257] Daniel Kahneman and Angus Deaton, "High Income Improves Evaluation of Life But Not Emotional Well-Being," *PNAS* (2010).

perfectionism as an example. On the one hand, donating to charity may qualify as an example of becoming morally better which is included on perfectionist lists like Sher's. Yet, it is unclear how this trades off against other items on the list. That is, it is unclear what the typical person might spend the money on instead and it is unclear how typical uses of money contribute to the attainment of the perfections. If, for example, the loss of $2,500 made it difficult for the person to meet basic needs (like housing or food), then it would have an enormous impact on perfection. But, if it traded off with frivolous status-seeking purchases, then perhaps the harm is not so great.

The most serious reason to be concerned about charity nudging on a perfectionist account of wellbeing is that being nudged to donate to charity might not develop the relevant perfection at all. That is, perhaps one must execute agency or rational choice to develop a perfection. Further, donating to charity via this program might trade off with charity donations that *do* involved the relevant agency. This concern can be mitigated if nudging someone to donate to charity is a plausible stepping stone on the path to endorsing the charity donation via rational choice or exercising agency. But, it is unclear if this would be a plausible result.

Finally we come to informed desire satisfaction accounts of wellbeing. On the multiple-agent version of the informed desire satisfaction account of wellbeing that I have developed, the relevant question is what a parliament of versions of the employees that each experience a different version of the possible lives that result from donating money would want the actual employee to do. The answer to this question is unclear. It probably depends a good deal on the individual's starting point. For example, if the person starts with some altruistic desires that the donation fulfills, then I can imagine that many of the representatives of the parliament would retain some variant of these desires and would be persuaded to follows a path

that satisfies those desires. If the person does not start with any altruistic desires, then the parliament is significantly less likely to value these outcomes.

One reason that actual people are often not motivated to donate to charity is because the impact of their donation is not sufficiently emotionally salient. This explains why we would ruin a $2,500 suit to save a child drowning in a pond, but are less certain about donating $2,500 to save a child from malaria. The outcome of the parliamentary debate may also depends on some details about how the parliamentary delegates transmit information. Imagine that in one of the possible lives, the parliamentary delegate goes and sees the children that would be saved by donating. This would be an extremely powerful emotional experience that would make it very likely that this delegate would favor options that including donating. But, in communicating their experience, do we suppose that the other delegates gain a similar level of emotional salience?

To avoid the problem of needing the delegates in the parliament to become full idealized (and thus reintroducing the problems the model was attempting to solve), we probably need to specify that the delegates communicate their story using normal means of communication. If so, then while many delegates will be persuaded by communicating the experience, not all of them will. It is ultimately unclear whether, in general, donating to charity satisfies a person's informed desire without knowing a good deal more about their current desires. It is also unclear whether most people would have an informed desire for charity donation (although the optimist in me hopes that they do).

In sum, it seems most plausible that the charity donation nudge increases the chooser's wellbeing on the informed desire account of wellbeing. It is unlikely to increase wellbeing on the objectivist account and somewhat likely to increase wellbeing on the hedonist account.

5.2.2.2. Tradeoffs in wellbeing and welfarism

Given the large positive benefit to others that the charity donation nudge has, if the charity donation nudge improves the wellbeing of the chooser, then it is likely to be unobjectionable. In this section I consider the harder question of whether the nudge is objectionable if it does not benefit the chooser.

The answer to this question depends on how one thinks about the connection between wellbeing and morally correct action. One plausible view might be called *welfarism*. Welfarism is the view that all the justificatory force of any moral reason rests entirely on wellbeing. Put another way, those actions which increase aggregate wellbeing are good and those which decrease it are bad. If welfarism is true, then we only need to plug in the relevant account of wellbeing and then determine if the charity nudge increases or decreases aggregate wellbeing. In our charity donation nudge, it seems extremely plausible that aggregate wellbeing is increased on any plausible account of wellbeing. The pleasure/informed desire satisfaction/perfection opportunities that the child will gain from not dying are almost certainly far greater than those that will be lost by having an individual employee lose $2,500. Let us then give a tougher case for the welfarist approach:

> *Charity donation paternalism:* A company wants to encourage more of their employees to give to charity. To do so, they force all employees to donate 5% of their income to charity. The default charity is chosen from GiveWell's list of Top Recommended Charities which means that the default charity will be evidence-backed, thoroughly vetted

and underfunded. Employees are given notice of the change and are fired if they do not wish to comply.

In this case, the 10 employees now donate $25,000 per year and save nearly 9 children per year from death due to malaria. Therefore, a welfarist might presumably say that this is twice as good as the nudge approach. Those who want to resist this implication might point to some mitigating factors, but I think the basic picture will remain correct. We can imagine still more draconian policies as well. Why stop at 5% of income? Why not 50%? We can imagine that the actor is not a company but the IRS who automatically donates all tax returns to worthy charities and so on. Some may take this as a *reductio* to welfarism and perhaps it is. The point for our purposes is that welfarism would at least justify the charity donation nudge and perhaps much more.

Another view that one might take is that maximizing aggregate wellbeing is not all that matters. There are constraints on how one ought to behave that are not directly related to wellbeing considerations. A framework for reviewing these considerations was developed in chapter two of this dissertation. I turn now to applying elements of that framework to the present case.

5.2.3. The three-factor framework and charity nudges

In chapter two of this dissertation I develop a three-factor framework for evaluating nudges. The framework included: nudge type/mechanism considerations, agent-relative considerations, and ends-based considerations. I will now see how each of these considerations might apply to the charity donation nudge case.

5.2.3.1. The charity donation nudge and the nudge type/mechanism

One way that a nudge can have a problematic mechanism is if the mechanism makes the nudge difficult to notice or call to attention. In chapter 1 I discussed the case of Asparagus-Lovers in which a false childhood memory about loving asparagus is implanted into a choose to get them to eat more asparagus. This nudge is problematic because it is extremely difficult for the chooser to resist which infringes on her autonomy. There are some possible instantiations of the charity donation nudge that could make it more difficult to resist. For example, we could implant a false memory about wanting to sign up to donate to charity. More plausibly, the choice architect could simply fail to tell the employees that 5% of their income is going to charity. Presumably many employees would notice a 5% decrease in their paycheck, but those who do not might have difficulty resisting the nudge. In the case as specified however, it does not appear that any choosers will have difficulty resisting the nudge. They are told about the nudge beforehand and will be constantly reminded about it when they see the size of each subsequent paycheck. Therefore, the nudge is easily resistible.

A second way that a nudge can be problematic is if the chooser would not endorse the nudge if he became aware of it. For example, imagine the HR person at the company is an extremely conservative Republican, yet all of the other employees are very liberal Democrats. It would be problematic if the HR person instituted an automatic charity donation to the Republican party because the others in the company would not endorse this decision if they became aware of it. Again, this does not seem to apply in our case. The concern is likely moot because this nudge has a high degree of resistibility but also most people would endorse saving the lives of children if they became aware that they were doing so.

5.2.3.2. Agent-relative considerations in the charity donation nudge

Another way that the charity donation nudge might be problematic is that the nature of the relationship between the choice architect and the chooser might make the nudge problematic. In chapter 2 I proposed that nudges that take place in contexts of high trust may be problematic because this context makes the nudge harder to resist. However, I do not think there's any reason to suspect that a high-trust relationship exists between employers and employees, so this is probably not a realistic concern.

A different worry might be that there is a natural power disparity between employers and employees and that this nudge is problematic because of this power disparity. For example, it could be that employees have an implicit fear of reprisal for failing to do something that the company clearly wants them to do. This might make the donation decision feel to the employee like a subtle form of coercion which may make the donation decision not entirely voluntary.

This concern depends largely on the specifics of how the program is unveiled and executed. Consider two alternatives:

> *Opt-out mechanism one:* A company wants to encourage more of their employees to give to charity. To do so, they automatically enroll all employees in a program that donates 5% of their income to charity. The default charity is chosen from GiveWell's list of Top Recommended Charities which means that the default charity will be evidence-backed, thoroughly vetted and underfunded. To opt out of the program employees must submit their opt-out form to the company leadership at the weekly staff meeting and must explain why they do not wish to participate.

Opt-out mechanism two: The program is the same as before, but the opt-out mechanism is different. Employees may out out of the program online with relatively anonymity. Only the company's accountant will know if an employee has opted out and the accountant has made it clear that the decision will be confidential.

It seems plausible that opt-out mechanism one uses the nature of the relationship between employer and employee in a problematic way whereas opt-out mechanism two does not. This may be true even if employees suffer no actual reprisal from their decision. Because of the power disparity between the two parties, the perceived threat of reprisal makes the decision more coercive. Therefore, all else being equal, opt-out mechanism two is preferable.

5.2.3.3. Ends-based considerations in the charity donation nudge

In chapter two I considered two ends-based considerations that are relevant to various kinds of nudges. The consideration that is relevant to the present case is whether there is something problematic about instituting a nudge that is done not for the benefit of the chooser but for the benefit of someone else. I can imagine three possible positions on this question. The first would be that it is never OK to institute a nudge that is not for the benefit of the chooser. The second is that it is always OK to institute such a nudge provided that the total amount of wellbeing is increased. The final option is that it is sometimes OK to institute such nudges but that the decision does not depend on entirely on whether the nudge increases wellbeing.

I think we can rule out the first option. This view would entail extremely counterintuitive conclusions. For example, it would imply that it would be impermissible to nudge doctors to wash their hands (which costs them time) if it would save patient lives. Or, it would entail that it

is impermissible to nudge drives to slow down in school zones (which costs them time) in order to prevent children from being hit by cars. To exacerbate the implausibility of the claim, remember that nudges necessary do not force choosers to pick any particular option. So, some of the usual concerns about coercion and violations of autonomy do not readily apply. As such I think we can discard this view.

This leaves two claims: that a nudge is acceptable so long as it increases total wellbeing and the alternative claim that factors other than wellbeing play some role in determining whether a nudge is morally acceptable. It should be noted that taken a certain way, distinguishing between these views is tantamount to resolving the fundamental ethical debate between utilitarianism and deontology. For example, Sen argues that the following two claims, when combined, add up to utilitarianism:

> *Welfarism*: The judgment of the relative goodness of alternative states of affairs must be based exclusively on, and taken as an increasing function of, the respective collections of individual utilities in these states.
>
> *Sum-ranking*: One collection of individual utilities is at least as good as another if and only if it has at least as large a sum total.[258]

Taking a stance on the debate between utilitarians and deontologists is beyond the scope of this dissertation. Instead, I will confine myself to noting how one might evaluate the charity donation nudge if one accepts welfarism and how one might evaluate the nudge if one rejects welfarism.

As noted above, if one accepts welfarism, it seems extremely difficult to find the charity donation nudge problematic. On any reasonable theory of wellbeing, saving a child from dying

[258] Amartya Sen, "Utilitarianism and Welfarism," *The Journal of Philosophy* 76, no. 9 (1979): 468.

from malaria is likely to produce a greater amount of wellbeing than giving a well-off person some additional money. As noted previously. the most difficult issue for welfarism is going to be the question of why nudge at all. That is, many nudges are attractive for reasons outside of wellbeing. Nudges allow choice architects to improve the wellbeing of choosers without violating their autonomy or otherwise restricting their options. But, why should a welfarist be concerned about violations of autonomy? It is possible that such violations have a negative effect on wellbeing, but they also often have a profound impact on producing the desired behavior. For example, while nudging can produce a large increase in 401(k) participation rates, forcing employees to participate in the 401(k) program would presumably produce an even larger effect. But, this justifies forcing choosers to undertake a large number of actions which seems to be problematic. I leave it to welfarists to grapple with these questions.

Those who reject welfarism will hold that a state of affairs may contain an increase in wellbeing but not produce a corresponding increase in goodness. For example, one who rejects welfarism may hold that if two states of affairs are identical in wellbeing, but one was produced through coercion and the other was not, the state of affairs produced through coercion is worse. Indeed, many of the considerations raised in chapter two of this dissertation presume some kind of rejection of welfarism. So, on this view, those considerations become relevant. I have already noted how various constraints on action might come into play in the charity donation nudge case, so I will not repeat those here. However, I will note that, on balance, the charity donation nudge seems to be justifiable.

5.2.4. Alternatives to the charity donation nudge

In the previous section I noted that nudging employees to donate money to charity through a default opt-in mechanism is likely to be justified on most reasonable conceptions of the relationship between wellbeing and goodness. In this section I consider whether nudging is the proper mechanism for producing the desired outcomes.In particular, I use the distinction developed in chapter 4 between nudging, paternalism, libertarianism and rational persuasion to analyze whether alternatives to nudging would be preferable.

5.2.4.1 Libertarianism and the charity donation nudge

In the charity donation case, the libertarian option would be to take no action to encourage charity donations at all. The specification in this case that employees would not donate absent the program combined with the discussion in the previous section entails that nudging would be preferable to libertarianism in this case. The relevant question then from a libertarian standpoint is whether one is obligated to do those things that are preferable. This boils down to the question of supererogation, that is, it boils down to whether all action that are morally better are also morally required. This is a question that I do not intend to take a stand on here. Instead I note that given the preceding discussion, one would need to hold some version of the supererogation thesis in order to justify libertarianism. Yet, even those who hold this thesis might still prefer to nudge choosers because doing so would be preferable.

5.2.4.2. Paternalism and the charity donation nudge

Another option would be paternalism. In this case, employees would be forced to donate to charity by their employer. Above I specified such a case as follows:

> *Charity donation paternalism:* A company wants to encourage more of their employees to give to charity. To do so, they force all employees to donate 5% of their income to charity. The default charity is chosen from GiveWell's list of Top Recommended Charities which means that the default charity will be evidence-backed, thoroughly vetted and underfunded. Employees are given notice of the change and are fired if they do not wish to comply.

The paternalism option would result in more donations to charity. Assuming the participation rate increases to 100%, this would result in an addition $12,500 per year and would save the lives of an additional 4.4 children per year. Yet, I suspect that many would feel uncomfortable with the paternalistic option. In fact, it seems plausible that employees would be outraged by the instantiation of such a policy.

Part of this reaction is framing. In particular, the example as written plays on loss aversion, the cognitive bias to avoid losses more strongly than to seek gains. Imagine that instead of forcing employees to donate 5% of their income, employees are are all paid 5% less than the standard for their industry and a $2,500 donation is made in the name of each employee to charity every year. Framed like this, employees may actually feel good about the donation instead of feeling like they have been wronged. Practically speaking however, paternalistic options that do not attempt to force donations from employees make little sense. If a company wishes to donate money to charity it can simply donate its profits to a charitable cause. It can also negotiate a lower salary for the employee (which most companies do anyway) and donate the excess money to the charitable cause. So, the only way that the company can gain access to additional money for charitable donations that would not have been available to it otherwise is to

attempt to take money out of an employee's paycheck and donate that money to charity. Presumably, many would find this intuitively problematic. Is there something that can be said in defense of this intuition?

One obvious first pass at a defense is that forcing employees to donate to charity is a violation of their autonomy. Employees should be free to spend their money however they see fit and one's employer should not have a say on how an employee spends their money. Unfortunately, this response is probably too quick. Autonomy involves "shaping one's own life in ways that one finds valuable or important, as opposed to going through life mindlessly or based on other people's agenda.[259]" The problem with thinking that the employer nudge removes the ability for employees to shape their own lives is that employees can simply choose to work elsewhere. Indeed, we generally do not hold that companies are doing something morally wrong in a wide range of cases where something is done against the employee's wishes. For example, an employer is not violating an employee's autonomy in giving them a pay cut because the relationship between the employee and the employer is voluntary.[260] In fact, provided the the employer announced the new policy and gave employees sufficient time to decide if they wished to leave the company, forcing them to donate to charity is really no different than announcing a pay cut or some other reduction in employee benefits.

Perhaps the paternalistic employer is not violating autonomy in some absolute sense, but is instead doing it relative to the available options. That is, the employer would be better enabling employees to shape their lives as they see fit by allowing them to control how they

[259] Dworkin, *The Theory*.
[260] Of course not all relationships between employees and employers are voluntary in this way. We can imagine cases where an employer is the only available option for employees or where employers have imposed high economic costs on switching jobs. I am excluding theses cases from consideration.

spend their money. This seems plausible, but from the fact that the employer is harming autonomy is a relative sense, it does not follow that the employer is doing something wrong. For example, the government is paternalistic in the sense that it takes taxes from me and then spends them for me in an attempt to improve society. We might not want to say that taxation is impermissible, so perhaps a paternalistic charity nudge is not impermissible either.

It appears then, that it is difficult to find an argument that would forbid a paternalistic charity donation program. It may be that such a program is unwise due to practical considerations. It seems likely that employees would be unhappy with such a program and this unhappiness may manifest itself in decrease productivity, loss of employees and so on. If so, this may provide a strong practical reason not to engage in a paternalistic program of this sort even if it is not impermissible.

5.2.4.3. Rational persuasion and the charity donation nudge

The final alternative is rational persuasion. On this option, the employer would provide information to employees that explains the reasons they should donate to charity and then asks them to make a donation. I have argued elsewhere that all else being equal, rational persuasion is preferable to nudging. So, if an employer could get the same rate of charity donations through rational persuasion, then that is what ought to be done. Whether rational persuasion is likely to be as effective as nudging is an empirical question. My guess is that this is unlikely. Given the other consideration developed in this section and the large amount of potential good created by the charity donation nudge, one ought to prefer nudging to rational persuasion if nudging is more effective.

5.3. End of life care and nudging

In this section I consider whether it would be appropriate to use nudges to influence the decisions of patients with regards to end of life healthcare decisions. This is an important question because end of life care represents an estimated $125 billion in medicare spending and for 40% of household out-of-pocket medicare bills accrued in end of life care exceed their assets.[261] In addition, the total cost of hospice care for the last year of life is $8,700 less than for nonhospice patients, a savings which would amount of billions of dollars across the US healthcare system.[262] Yet, patient desires are not served by the system. The majority of patients die in the ICU in pain and discomfort when they would prefer to die in familiar situations like at home. In addition, most terminally ill patients would prefer only comfort care whereas acute care is the norm in hospitals and is the ultimate end for most patients. Yet, the prospect of nudging patients at the end of life is complicated by the fact that end of life decisions are irreversible, often made by patients without full use of their mental capacities (or by their families), and are often emotionally and politically charged.

In this section I explore some nudges that might be used in end of life decisions and the ethical considerations that are relevant in these decisions.

5.3.1. Nudge mechanisms in end of life care

[261] Penelope Wang, "Cutting the High Cost of End-of-Life Care," *Time,* June 13, 2013, http://time.com/money/2793643/cutting-the-high-cost-of-end-of-life-care.
[262] Paula Span, "An Easier Death, and Less Costly, Too," *The New York Times*, November 20, 2014, http://newoldage.blogs.nytimes.com/2014/11/20/an-easier-death-and-less-costly-too/?src=recg&_r=0.

A few potential nudge mechanisms are available in the case of end of life care. One is the use of defaults in the creation of advanced directives. On this nudge, the advance directive that patients complete might have options for comfort care already selected but with an option to select some other plan if they wish. In one experiment that used this nudge, the comfort care default resulted in 77% of patients keeping that choice whereas only 61% selected comfort care in the case where no option was selected and only 43% selected comfort care when life-extending options were selected as the default. So, defaults can serve as a powerful options for choice architects.

The question of defaults can be taken even further. Currently many patients receive life-extending care because that is the default intervention option for doctors and surgeons. If the patient's wishes are not known, it is assumed that the patient wants to be kept alive by any means necessary. So, life-extending care is the default. We could imagine changing this default so that patients who will have very poor quality of life are given comfort care instead of life-extending care as a default. We could further imagine that comfort care is presented as the default to family members or other proxy decision makers for the patient and they must override that default in order to get life-extending care in these cases. The question of how doctors are to determine which patients should get life-extending care and which should get comfort care is a difficult issue that could pose serious problems for this attempted use of defaults.

Another option is required active choosing. In required active choosing, you simply require that terminally-ill patients make a choice as to their end of life care plan without attempting to persuade them to choose any particular option.[263] If the research suggesting that most patients would prefer only comfort care is correct, then required active choosing would

[263] It is not clear that required active choosing is in fact a nudge because it may not rely primarily on shallow cognitive processes. However, since it is one option a choice architect might choose, I include it here.

increase the number of people who have their end of life desire fulfilled without forcing choosers to select any particular option. Note that required active choosing can be combined with other nudge options. For example, one could require that people choose an end of life plan while defaulting them to comfort care only. This makes required active choosing a very powerful mechanism. However, one challenge for this nudge is determining the mechanism by which the choosing can truly be required. It may be that, practically speaking, we cannot require that patients in a wide range of physical and mental conditions make a choice on this matter.

A third nudge option would be to utilize framing in the presentation of options to the relevant decision makers. Bellicose expressions are common in discussions of health care alternatives. Patients are urged to "fight the disease" and when they die, they are often said to have "lost the battle." The fight metaphor assumes an opponent, namely the disease, and it presumes that the goal is to defeat one's opponent. If choice architects utilize fighting metaphors, it frames the discussion such that patients are more likely to choose the option that allows them to defeat their opponent -- like life-extending care -- and it makes them less likely to choose options that feel like defeat -- like comfort care. One nudge would be to reframe the discussion. This could be done in two ways. First, comfort care could be reframed as the fighting option. Instead of thinking of comfort care as giving up, comfort care could be presented as the option that helps the patient fight the best. Comfort care allows the patient to have more inner calm and less pain which makes it easier for them to fight the disease whereas life-extending care often causes them more pain and discomfort which makes it harder for the patient's body to do the fighting. A second option would be to use a different metaphor entirely. For example, British

doctors more frequently use journey metaphors[264] which may allow patients to see comfort care as the end of a journey instead of as losing the fight.

5.3.1.1. Selecting end of life nudges

While there are many options for end of life nudges, not all nudges are appropriate in all situations. In this section I discuss some of the factors that one might consider in determining which nudge to select.

One important factor is the decision making capacity of the chooser -- be it the patient or the patient's family. If the end of life is coming suddenly or traumatically, the chooser may lack the mental capacity to make well-reasoned decisions. The patient may be scared, in pain, confused or groggy from medicine, and family members may be distraught, confused and worried. This situation can present a particularly tricky challenge for choice architects. On the one hand, nudging can be especially helpful in these cases. Choice architects may be in a better position to evaluate the situation and may have the benefit of significant past experiences that the chooser may not have access to. A nudge in the right direction may save people from significant pain, financial hardship and emotional trauma. Yet, on the other hand, a nudge can easily become a prod (to borrow Saghai's phrase) in these cases. A prod is a nudge that cannot be easily resisted by the chooser. A nudge can be easily resisted if:

1. B has the capacity to become aware of A's pressure to get her to φ (attention-bringing capacities); and
2. B has the capacity to inhibit her triggered propensity to φ (inhibitory capacities); and

[264] Paula Span, "Fighting Words are Rarer Among British Doctors," *The New York Times*, April 22, 2014, http://newoldage.blogs.nytimes.com/2014/04/22/fighting-words-are-rare-among-british-doctors.

3. B is not subject to an influence, or put in circumstances that would significantly undermine the relatively effortless exercise of attention-bringing and inhibitory capacities.[265]

Nudges in the case of end of life care may not be easily resistible in a number of ways. Patients may not have the capacity to become aware of the doctors pressure to get them to φ because they may be emotionally or mentally compromised. In such a case, they may not be able to resist the propensity to φ and they may not be able to call the fact that they are being nudged to attention. Nudging someone who is in tremendous pain, may amount to the doctor imposing her will on the patient instead of the doctor attempting to steer the patient in particular ways.

The nudge may also be difficult to resist because patients may trust the doctor and assume that the doctor will not nudge them. For example, many patients assume that if the doctor advocates for a particular course of action, it is probably correct. They may not realize that there are tradeoffs between various options and that reasonable people can disagree about which course of action they would prefer. It may be that if a doctor presents comfort care as the default option, the patient will not be able to call to attention the fact that there are alternative options.

To avoid this issue, it may be best to simultaneously nudge and rationally persuade. For example, the patient's advanced directive could have the comfort care option preselected, but it could also contain detailed information on what each option entails and arguments for why many people might be better off choosing comfort care. Here, the rational persuasion makes it more likely that the chooser brings the issue to conscious attention while the nudge makes it more likely that the chooser selects the option that will improve their situation. This is as compared to

[265] Saghai, "Salvaging," 3.

a pure nudge, like providing an advanced directive with one option selected and no sufficient information about what each option entails. This nudge would make it very difficult to bring the nudge to conscious attention.

Similarly, doctors should frame the situation in such a way that it is promoted to conscious attention. For example, if a doctor discussed the option to pursue comfort care or life-extending care with the patient while framing the discussion such that the patient is more likely to choose comfort care, this ensures that the issue is brought to conscious attention while still realizing the benefits of utilizing the nudge. Nudges of this type are probably ideal in this case and may be necessary to acquire informed consent.

5.3.2. Death and the informed desire satisfaction account

In chapter 3 I argued that we ought to nudge people towards the actions that would be chosen under conditions of superior information. This was developed through a parliamentary model according to which what is best for a person is what one's future selves would be in favor of if each of them experienced a different possible future and then voted on the future that they found most appealing. On this account, a nudge makes a person's life go better if it makes it more likely that they will choose the option that the hypothetical parliament would have chosen.

Death, however, poses unique challenges for this model. One obvious problem is that the parliament members may need to die in order to evaluate whether a path would make a person's life go better. That is, to know whether death via life-extending care is better or worth than death via comfort care, the parliament members may need to experience these deaths. But, it may be nonsensical to wonder about the votes of hypothetical dead people.

There are two possible responses to this concern. The first is that the parliamentary model is intended to be a thought experiment and is not intended to be taken literally. Just as there is no real Original Position for Rawls, there is no real parliament for the informed desire account and so, objections to the details of the parliament might be seen to miss the point. I think this response is successful as long as the situation is not logically incoherent. In the version of the informed desire account discussed in the literature, a hypothetical person would need to contain mutually exclusive attributes and have mutually exclusive experiences. This is logically impossible and so, it is meaningless to discuss what actions such a person would choose. If the parliamentary models leads to a similar logical incoherence then it too might fail. Fortunately, I do not think the parliamentary model is logically incoherent in this way. It is not incoherent to imagine that a member of the hypothetical parliament has all of the experiences of dying but is not dead.

A second response is that all of the relevant information might be obtained before the person is dead. Imagine for example that all of the information about the person's life is preserved just before death and then the parliament votes based on that information. Since no information is gained from actually being dead (because there is no person to experience death), this should be sufficient for the parliament to debate the question. So, if we rephrase the parliament model slightly we get a sensible view that avoids these issues.

5.3.3. When does death make a person's life go better?

Given the discussion above, we may not be in a position to determine when nudging patients towards comfort care might be preferable to care that prolongs their life. The short answer, of

course, is that nudging patients towards comfort care makes their life go better if that is what would be chosen given full information. Yet, it would be helpful to say more. In this section I consider some specific kinds of cases where a chooser's reported actual desires are particularly likely to diverge from their idealized desires.

One example is cases where patients are required to reason about experiences that they have never had and for which they do not have good reference experiences. Decisions about life prolonging care versus comfort care often involve the patient attempting to reason about what their quality of life might be while undergoing life prolonging care and subsequent to life prolonging care. Yet, most patients simply have no reference experiences about which to make these judgments. On the informed desire satisfaction model we ought to help patients choose what they would choose if they had the benefit of the experience. So, doctors who have seen many other patients go through the dying process might be in an especially good position to nudge patient by giving them the benefit of their experience.

A second case is a decision where the key information is statistical in nature since it is hard for some individuals to reason statistically. One well known example is prostate cancer. When men learn that they have prostate cancer, what is most salient to them is that they have **CANCER** and, accordingly, they assume that aggressive treatment to remove the cancer is the sensible choice. Yet, statistically speaking, monitoring the development of the cancer and opting for surgery later if necessary does not increase the chances of cancer-related mortality[266] but intervention increases the rate of impotence and other issues. So, the informed desire of most men would be to opt for watchful waiting if they understood the statistics. Since choice architects

[266] Christopher Warlick, Bruce Trock, Patricia Landis, Jonathan Epstein, Ballentine Carter, "Delayed Versus Immediate Surgical Interventions and Prostate Cancer Outcome," *Journal of the National Cancer Institute* 98, no. 5(2006).

might be better equipped to make sense of the relevant information, a nudge might be appropriate.

A final example is cases where patients are making use of misguided heuristics. This can occur because the complexity of medical decision making is often very high and so heuristics can be a useful way to make sense of the information. But, sometimes these heuristics can lead patient's astray. Earlier in this chapter I discussed the "fight" heuristic where patients see the disease as an opponent that must be defeated. While this provides a useful way of understanding one's experience it can also cause patients to make decisions that they might not have made with full information. A second heuristic is that drugs are bad. Here the idea is that taking too many drugs is bad for you and you should only take them if you really need them. This heuristic does not distinguish between different kinds of drugs or evaluate their side effects, so choice architects who understand the nuances better can be in a good position to nudge the patient.

5.4. Conclusion

In this chapter of the dissertation I have taken two specific topics and shown how the considerations of the preceding four chapters of this dissertation might apply. First I discussed a specific hypothetical charity donation nudge and showed that the considerations outlined elsewhere in the dissertation can probably be used to justify such a nudge. Second, I discussed nudging in the case of end of life care and showed what kinds of nudges might be appropriate and what considerations come to bear on selecting such a nudge.

What is interesting about the nudge space is that it is vast. Nudges have already been used to design the 401(k) pension scheme,[267] suggest a tax refund system,[268] improve compliance in tax reporting[269] and lower youth alcohol consumption[270] among many other things and the field is still young. My hope is that this dissertation provides tools that policymakers can use to ensure that nudges do not becomes shoves and that the nudge mechanism delivers on its promise to improve the lives of millions of choosers.

[267] Andrews, "Obama Outlines."
[268] Surowiecki, "A Smarter Stimulus."
[269] Cabinet Office and Behavioral Insights Team, "Applying Behavioral Insights."
[270] Ibid.

BIBLIOGRAPHY

American Academy of Pediatrics, "Restraint Use on Aircraft," *Pediatrics, 108, no. 5* (2001).

Andrews, Edmund, "Obama Outlines Retirement Initiatives," New York Times, (2009).

Anderson, Elizabeth, *Value in Ethics and Economics*, (Harvard University Press, 1993).

Aristotle, *Nicomachean Ethics*, ed. Gregory Crane (The Perseus Digital Library: 4th Century BCE).

Arneson, Richard, "Human Flourishing versus Desire Satisfaction," *Social Philosophy and Policy 16, no. 1* (1999).

Arrow, Kenneth, "A Difficulty in the Concept of Social Welfare," *Journal of Political Economy, 58,* (1950).

Arras, John, "Theory and Bioethics," *The Stanford Encyclopedia of Philosophy*, (2010).

Baber, Harriet "The Experience Machine Deconstructed," *Philosophy in the Contemporary World 15, no. 1* (2008).

Baggini, Julian & Fosl, Peter S. *The Ethics Toolkit: A Compendium of Ethical Concepts and Methods* (Wiley-Blackwell, 2007).

Balz, John. "A nudge on a hot button issue: Abortion," *The Nudge Blog*, (2008).

Beaulieu, Gerald, "The Normative Authority of Our Fully Informed Judgments," *The University of Manitoba* (1997).

Bentham, Jeremy, *An Introduction to the Principles of Morals and Legislation*, eds. J. Burns and H.L. A. Hart (Oxford: Clarendon Press, 1789).

Blumenthal-Barby, Jennifer, "Between Reason and Coercion: Ethically Permissible Influence in Health Care and Health Policy Contexts," *Kennedy Institute of Ethics Journal 22, no.4* (2012).

Blumenthal-Barby, Jennifer, "Choice Architecture: A Mechanism for Improving Decisions While Preserving Liberty?" in *Paternalism: Theory and Practice*, eds. C. Coons and M. Weber (Cambridge University Press, 2013).

Blumenthal-Barby, Jennifer & Burroughs, Hadley "Seeking Better Health Care Outcomes: The Ethics of Using the 'Nudge,'" *The American Journal of Bioethics 12, no. 2* (2012).

Bovens, Luc, "The Ethics of Nudge," in *Preference Change: Approaches from Philosophy, Economics and Psychology*, eds. Till Grüne-Yanoff and Sven Ove Hansson, vol. 42, (Berlin and New York: Springer, Theory and Decision Library A, 2008).

Bradford, Gwen, "The Value of Achievement," *Pacific Philosophical Quarterly 94, no. 2* (2013).

Bradford, Gwen, "Problems for Perfectionism," Unpublished Manuscript, (2014).

Bradley, Ben, "Objective Accounts of Well-Being" quoted in Ben Eggleston and Dale Miller, eds., *The Cambridge Companion to Utilitarianism* (Cambridge: Cambridge University Press, 2014).

Bradley, Ben; Feldman, Fred & Johannson, Jens eds., *The Oxford Handbook of Philosophy of Death* (2013).

Brandt, Richard, *A Theory of the Good and the Right* (Prometheus Books, 1979).

Brentano, Franz, *Psychology From An Empirical Standpoint*, ed. Linda McAlister (London: Routledge and Kegan Paul, 1973).

Broad, Charlie Dunbar, *Five Types of Ethical Theory* (London: Routledge and Kegan Paul, 1930).

Buss, Sarah, "Valuing Autonomy and Respecting Persons: Manipulation, Seduction, and the Basis of Moral Constraints," *Ethics 115* (2005).

Buss, Sarah, "Personal Autonomy," *Stanford Encyclopedia of Philosophy* (2013).

Butler, Joseph, *Fifteen Sermons Preached at the Rolls Chapel* (London: James and John Knapton, 1729).

Cabinet Office and Behavioral Insights Team, "Applying Behavioral Insights to Health," (2011).

Cabinet Office and Behavioral Insights Team, "Applying Behavioral Insights to Reduce Fraud, Error and Debt," (2011).

Caraballo, Hector, "Emergency Department Management Of Mosquito-Borne Illness: Malaria, Dengue, And West Nile Virus," *Emergency Medicine Practice, 16, no.* 5, (2014).

Carson, Thomas, "Rationality and Full Information," in *Ethical Theory*, 2nd. ed. Russ Shafer-Landau: (Wiley-Blackwell, 2013).

Chisholm, Roderick, *Brentano and Intrinsic Value* (Cambridge: Cambridge University Press, 1986).

Crisp, Roger, *How should one live?: Essays on the Virtues,* (New York: Oxford University Press, 1998).

Crisp, Roger, "Well-Being," *Stanford Encyclopedia of Philosophy*, (2013).

Crisp, Roger & Hooker, Brad, *Well-Being and Morality: Essays in Honour of James Griffin*, (New York: Oxford University Press, 2000).

Daniels, Norman, "Can Cognitive Psychotherapy Reconcile Reason and Desire?," *Ethics, 93,* (1983).

Danziger, Shai; Levav, Jonathan & Avnaim-Pesso, Liora, "Extraneous Factors in Judicial Decisions," *Proceedings of the National Academy of Sciences of the United States of America, 108, no. 17* (2010).

Darwall, Stephen, *Impartial Reason*, (Cornell University Press, 1983).

David, Paul A. & Reder, Melvin W., *Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz*, (New York: Academic Press, Inc., 1974).

Dolan, Paul; Hallsworth, Michael; Halpern, David; King, Dominic & Vlaev, Ivo "Mindspace: Influencing Behavior through Public Policy," *Institute for Government,* (2010).

Dorsey, Dale, "Three Arguments for Perfectionism," *Nous 44, no. 1* (2010).

Dworkin, Gerald, *The Theory and Practice of Autonomy* (Cambridge University Press, 1988)

Easterlin, Richard, "Does Economic Growth Improve the Human Lot? Some Empirical Evidence," in *Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz*, eds. Paul A. David and Melvin W. Reder (New York: Academic Press, Inc., 1974).

Eggleston, Ben & Miller, Dale, *The Cambridge Companion to Utilitarianism* (Cambridge: Cambridge University Press, 2014).

Enoch, David, "Why Idealize?," *Ethics 115, no. 4* (2005).

Epicurus, *Principal Doctrines*, trans. Robert Drew Hicks (1925).

Flegal, Katherine; Graubard, Barry; Williamson, David & Gail, Mitchell, "Cause-Specific Excess Deaths Associated with Underweight, Overweight, and Obesity," J*AMA: Journal of the American Medical Association*, (2007).

Frankena, William K., *Ethics*, 2nd ed. (New Jersey: Prentice-Hall, 1973).

Furrow, Dwight, *Ethics*, (New York and London: Continuum, 2005).

Gauthier, David, *Morals by Agreement*, (Oxford University Press, 1986).

Gibbard, Allan, *Wise Choices, Apt Feelings* (Harvard University Press, 1990).

GiveWell, "Against Malaria Foundation (AMF)" (2015).

Gorin, Moti, "The Nature and Ethical Significance of Manipulation," (PhD diss., Rice University, 2013).

Griffin, James, *Well-Being: Its Meaning, Measurement and Moral Importance* (Oxford: Clarendon Press, 1986).

Griffin, James, *Value Judgment: Improving our Ethical Beliefs* (New York: Oxford University Press, 1996).

Griffin, James, "Replies," in *Well-Being and Morality: Essays in Honour of James Griffin*, eds. Roger Crisp and Brad Hooker (New York: Oxford University Press, 2000).

Grüne –Yanoff, Till, "Old Wine in New Casks: Libertarian Paternalism Still Violates Liberal Principles," in *Social Choice & Welfare*, eds. J. Duggan, B. Dutta, M. Fleurbaey and C. Puppe, vol. 38, (Springer, 2012).

Hansen, Guldborg & Jespersen, Andreas. "Nudge and the Manipulation of Choice: A Framework for the Responsible Use of the Nudge Approach to Behavioral Change," *European Journal of Risk Regulation* (2013).

Hare, Richard Mervyn, *Moral Thinking*, (Oxford: Oxford University Press, 1981).

Harman, Gilbert, "Critical Review," *Philosophical Studies, 42*, (1982).

Harsanyi, John, "Morality and the Theory of Rational Behavior," in *Utilitarianism and Beyond*, ed. Amartya Sen and Bernard Williams (1982).

Hausman, Daniel & Welch, Brynn "Debate: To Nudge or Not to Nudge," *The Journal of Political Philosophy*, 18, no. 1 (2010): 125.

Haybron, Daniel, *The Pursuit of Unhappiness: The Elusive Psychology of Well-Being* (Oxford University Press, 2010).

Heathwood, Chris, "The Problem of Defective Desires," *Australasian Journal of Philosophy 83, no. 4* (2005).

Hooker, Brad, "Does Moral Virtue Constitute a Benefit to the Agent?," in *How should one live?: Essays on the Virtues*, ed. Roger Crisp (New York: Oxford University Press, 1998).

Hume, David, *An Enquiry Concerning the Principles of Morals* (London: A. Millar, 1751).

Hurka, Thomas, *Perfectionism* (Oxford: Oxford University Press, 1993).

Jackevicius, Cynthia; Mamdani, Muhammad & Tu, Jack, "Adherence With Statin Therapy in Elderly Patients With and Without Acute Coronary Syndromes," *JAMA: Journal of the American Medical Association* (2002).

Johnson, Robert, "Kant's moral philosophy" *Stanford Encyclopedia of Philosophy*, (2008).

Kagan, Shelly, *Normative Ethics* (Boulder and Oxford: Westview Press, 1998).

Kagan, Shelly, "The Limits of Well-Being," in *The Good Life and the Human Good*, eds. Paul & Miller (1993).

Kahneman, Daniel & Deaton, Daniel, "High Income Improves Evaluation of Life But Not Emotional Well-Being," *PNAS* (2010).

Kahneman, Daniel & Krueger, Alan, "Developments in the Measurement of Subjective Well-Being," *Journal of Economic Perspectives* (2006).

Kahneman, Daniel & Tversky, Amos*, Choices, Values, and Frames*, Cambridge University Press (2000).

Kauppinen, Antti, "Working Hard and Kicking Back: The Case for Diachronic Perfectionism," *Journal of Ethics and Social Philosophy*, (2007).

Kraut, Richard, *What is Good and Why* (Cambridge: Harvard University Press, 2007).

Kwak, James "Improving Retirement Savings Options for Employees," *John M. Olin Center for Law, Economics and Business Fellows' Discussion Paper Series* (2014).

Lau, Joe & Deutsch, Max, "Externalism About Mental Content", *The Stanford Encyclopedia of Philosophy* (2014).

Lewis, David, "Dispositional Theories of Values," *Proceedings of the Aristotelian Society, 63,* (1989).

Loeb, Dan, "Full-Information Theories of Individual Good," *Social Theory and Practice 21, no. 1* (1995).

Loewenstein, George & Haisley, Emily, "The Economist as Therapist: Methodological Ramification of 'Light' Paternalism," in *Foundations of Positive and Normative Economics*, eds. Andrew Caplin and Andrew Schotter (2008).

Lott, Maxim, "Gov't Knows Best? White House Creates 'Nudge Squad' to Shape Behavior," *Fox News*, (2013).

Luper, Steven, "Death," *The Stanford Encyclopedia of Philosophy* (2014).

Luper, Steven, "Exhausting Life," *The Journal of Ethics* (2010).

Luper, Steven, "Retroactive Harms and Wrongs," in Ben Bradley, Fred Feldman and Jens Johannson, eds., *The Oxford Handbook of Philosophy of Death* (2013).

MacAskill, William, "Normative Uncertainty" (Diss., Oxford University, 2014).

Markovits, Julia, "Acting for the Right Reasons," *Philosophical Review 119, no. 2*, (2010).

Mill, John Stuart, *Utilitarianism* (London: Parker, Son and Bourn, 1863).

Mindlin, Alex, "Dos and Don'ts of Gentle Prodding," *The New York Times* (2008).

Mitchell, Gregory, "Libertarian Paternalism is an Oxymoron." *Northwestern University Law Review*, (2005).

Moore, Andrew, "Objective Human Goods" in *Well-Being*, eds. Crisp and Hooker (2000).

Moore, G. E., *Principia Ethica* (Cambridge: Cambridge University Press, 1903).

Morgenbesser, Sidney; Suppes, Patrick & White, Morton, *Philosophy, Science and Method: Essays in Honor of Ernest Nagel*, (New York: St. Martin's Press, 1969).

Muehlhauser, Luke & Williamson, Chris, "Ideal Advisor Theories and Personal CEV," *Machine Intelligence Research Institute,* (2013).

Murphy, Mark, "The Simple Desire-Fulfilment Theory," *Nous 33, no. 2* (1999).

Newman, Thomas; Johnston, Brian & Grossman, David, "Effects and Costs of Requiring Child-Restraint Systems for Young Children Traveling on Commercial Airplanes," *Archives of Pediatrics and Adolescent Medicine, 157, no. 10,* (2003).

Nietzsche, Fredrich, "Twilight of the Idols," in *The Portable Nietzsche*, trans. Walter Kaufmann (New York: Viking Press, 1968).

Nozick, Robert, "Coercion," in *Philosophy, Science and Method: Essays in Honor of Ernest Nagel*, eds. Sidney Morgenbesser, Patrick Suppes, and Morton White (New York: St. Martin's Press, 1969).

Nozick, Robert, *Anarchy, State, and Utopia* (Oxford: Basil Blackwell, 1974).

Nussbaum, Martha, *Women and Human Development: The Capabilities Approach* (Cambridge: Cambridge University Press, 2000).

Nussbaum, Martha & Sen, Amartya, *The Quality of Life,* (Oxford University Press, 1993).

O'Neill, Brendan, "A Message to the Illiberal Nudge Industry: Push Off," *Spiked* (2010).

Paul, Ellen & Miller, Fred, *The Good Life and the Human Good*, (Cambridge University Press, 1993).

Parfit, Derek, *Reasons and Persons* (Oxford: Oxford University Press, 1984).

Plato, *Philebus*, ed. Gregory Crane (The Perseus Digital Library: 4th Century BCE).

Plato, *Protagoras*, ed. Gregory Crane (The Perseus Digital Library: 4th Century BCE).

Rachels, James, *The Elements of Moral Philosophy*, International ed. (McGraw-Hill, 2005).

Railton, Peter, "Facts and Values," *Philosophical Topics, 14, no. 2* (1986).

Railton, Peter, "Moral Realism," *Philosophical Review 95* (1986).

Rice, Christopher, "Defending the Objective List Theory of Well-Being," *Ratio 26, no. 2* (2013).

Rosati, Connie, "Persons, Perspectives and Full Information Accounts of the Good," *Ethics 105, no. 2*, (1995).

Rosen, Bernard, *Ethical Theory: Strategies and Concepts* (Mayfield Publishing Company, 1993).

Ross, W. David, *Foundations of Ethics* (Oxford: Clarendon Press, 1939).

Ryle, Gilbert, *Dilemmas* (Cambridge: Cambridge University Press, 1954).

Saghai, Yashar, "Salvaging the Concept of Nudge," *J Med Ethics* (2013).

Sahadi, Jeanne, "Average tax return tops $2,800" *CNN Money* (2015).

Saunders, Jason, ed., *Greek and Roman Philosophy after Aristotle* (New York: Free Press, 1997).

Scanlon, T. M., "The Status of Well-Being," *The Tanner Lectures on Human Values*, (1996).

Scanlon, T. M., *What We Owe to Each Other* (Cambridge: Belknap Press of Harvard University, 1998).

Scanlon, T. M., "Value, Desire and Quality of Life," in *The Quality of Life,* eds. Martha Nussbaum & Amartya Sen (1993).

Sen, Amartya, "Utilitarianism and Welfarism," *The Journal of Philosophy, 76, no. 9* (1979)..

Sen, Amartya, *Commodities and Capabilities* (Amsterdam: North-Holland, 1985).

Sen, Amartya & Williams, Bernard, *Utilitarianism and Beyond*, (Cambridge University Press, 1982).

Shafer-Landau, Russ, *Ethical Theory*, 2nd. ed., (Wiley-Blackwell, 2013).

Sher, George, *Beyond Neutrality: Perfectionism and Politics,* (Cambridge University Press, 1997).

Sidgwick, Henry, *The Methods of Ethics*, 7th ed., ed. John Rawls (Indianapolis: Hackett, 1982).

Sobel, David, "Full Information Accounts of Well-Being," *Ethics 104, no. 4* (1994).

Span, Paula, "An Easier Death, and Less Costly, Too," *The New York Times*, (2014).

Span, Paula, "Fighting Words are Rarer Among British Doctors," *The New York Times*, (2014).

Sumner, L. W., *Welfare, Happiness and Ethics* (Oxford: Oxford University Press, 1996).

Sunstein, Cass, *Simpler: The Future of Government*, (Simon & Schuster, 2013).

Sunstein, Cass & Thaler, Richard, "Libertarian Paternalism is Not an Oxymoron," *Social Science Research Network,* (2003).

Sunstein, Cass & Thaler, Richard, *Nudge: Improving Decisions About Health, Wealth, and Happiness*, Penguin Books, (2009).

Surowiecki, James, "A Smarter Stimulus," New Yorker, (2009).

Tiberius, Valerie, "Full Information and Ideal Deliberation," *Journal of Value Inquiry 31, no. 3* (1997).

Turton, Daniel Michael, "Reviving Hedonism about Well-Being: Refuting the Argument from False Pleasures and Restricting the Relevance of Intuitive 'Evidence,'" (MA thesis, Victoria University of Wellington, 2008).

Velleman, David, "Brandt's Definition of 'Good,'" *The Philosophical Review 97, no. 3* (1988).

Wall, Steven, "Neutralism for Perfectionists: The Case of Restricted State Neutrality," *Ethics*, (2010).

Wall, Steven, "Perfectionism in Moral and Political Philosophy," *Stanford Encyclopedia of Philosophy* (2012).

Wang, Penelope, "Cutting the High Cost of End-of-Life Care," *Time*, (2013).

Warlick, Christopher; Trock, Bruce; Landis, Patricia; Epstein, Jonathan & Carter, Ballentine, "Delayed Versus Immediate Surgical Interventions and Prostate Cancer Outcome," *Journal of the National Cancer Institute, 98, no. 5,* (2006).

Wenar, Leif, "Rights," *Stanford Encyclopedia of Philosophy* (2015).