Penultimate Draft. Final Version Forthcoming in *Social Philosophy & Policy*

*Repudiated Preferences: Redefining the Knowledge Problem in Behavioral Welfare Economics*

Angela Barnes, Arizona State University

AngelaBarnes@asu.edu

3/9/24

**1. Introduction**

Nudging, to the dismay of many, has become increasingly popular. A tool of behavioral public policy, nudging is an intervention on choice architecture that is supposed to increase the chances that an agent chooses what is good for him without interfering with his liberty or autonomy. To not interfere with an agent's liberty or autonomy, nudges are supposed to be in service of the fulfillment of the agent's preferences - but not always the preferences he reveals through choice. If the agent's welfare maximizing preferences were expressed through his choices, there would be no need for a nudge - the agent has already maximized his welfare. However, due to findings in behavioral science and economics, we know that agents' choices do not always reveal preferences that maximize welfare. Instead policymakers, to pursue welfare maximization that is in line with an agent's preferences, need to know about some *other* set of preferences that the agent contains: what are the agent’s *true preferences?*

The literature trying to find *true preferences* to answer this call has found itself in a loop. One theorist - an economist, a philosopher, or an economic methodologist - offers a new picture of what true preferences are, and another shows why it is problematic to nudge people toward the fulfillment of those preferences. A new picture is offered, and the same objection is raised. I argue the reason this is happening is because we seek a single answer to two separate questions. When we separate the questions, it becomes clear why one answer won’t do. On one hand, we have a question about welfare: what is it and what preferences are related to it? On the other hand, we have an applied issue about policy: when is a nudge permissible? There is no prima facie reason to believe that the answer to these questions is the same.

This, I believe, has partially happened because of an implicit belief that we need a theory of true preferences before we can start work on the applied problem: behavioral public policy. Daniel Hausman, who argues against this strategy, thinks that we have to abandon the possibility of non-paternalistic nudges and instead be satisfied with using plausible theories of welfare to justify nudges even if they are not related to an agent's preferences.[[1]](#endnote-1) I argue that we do not need to jump to that conclusion. Our lack of ability to identify a complete set of true preferences should not be a reason to abandon what we *can* know about people’s preferences. We need not be able to identify the correct end of a true preference to understand some conditions for what makes a preference mistaken. I introduce *repudiated preferences* to fill this gap.

**2. Two Knowledge Problems**

What is the behavioral welfare economics “true preferences” knowledge problem? It is the tension that results from wanting to retain the foundational economic assumption that individuals are in the best position to decide what is good for them and the empirical evidence from behavioral science that they sometimes do not choose what is best for them.

Mario Rizzo and Glen Whitman, in their aptly titled “The unsolved Hayekian knowledge problem in behavioral economics,” define the knowledge problem as the “facts that must be discovered for a behavioral policymaker to have a reasonable chance of improving on the decisions that individuals would make for themselves.”[[2]](#endnote-2) They go on to set out criteria for success, starting with, not least of all, *knowledge of true preferences*. A policymaker cannot nudge an individual closer to his welfare without knowing what the optimal choices for that individual are.

Prompting Rizzo and Whitman’s article was an argument by Sunstein, who argues that nudging improves people’s lives *as judged by themselves[[3]](#endnote-3)*. He states,

“[B]ehaviorally informed policy ought to be based not on the preferences and values of social planners, but on learning from the choices of *informed and unbiased choosers*. It might well turn out to be possible to identify those choices. Once we have done so, we might be on the road toward identifying appropriate interventions, whether they involve nudges, taxes, subsidies, or mandates.”[[4]](#endnote-4)

Sunstein and Thaler’s ‘*as judged by themselves’* criterion has proved troubling to make sense of for the last two decades. Alexander Cartwright and Marc Hight argue that there is no coherent way to understand this criterion. They go on to argue that:

“Without clear welfare gains, libertarian paternalists must re-evaluate or re-specify their welfare criteria. If their program is to continue in its current form, Thaler and Sunstein should admit to being regular paternalists since this is what any workable interpretation of their welfare criterion reveals under scrutiny.”[[5]](#endnote-5)

And many, of course, have been trying to offer exactly that welfare criterion to allow us to keep the ‘libertarian’ in libertarian paternalism. Suggestions range from those claiming to be value neutral, such as Douglas Bernheim, who offers the *unified normative welfare framework*,[[6]](#endnote-6) to value-centered, such as C. Tyler DesRoches’ *value-based preferences*,[[7]](#endnote-7) and everything in between.

What these approaches have in common is that they are looking for “preferences that are sufficiently stable, reasonable, autonomous, and associated with wellbeing to ground the ‘welfare’ principle of libertarian paternalism,” or more broadly, behavioral welfare economics, which can then, in turn “be identified by choice architects and folded into policy design.”[[8]](#endnote-8)

The knowledge problem is both a question about when preferences are and are not tracking welfare and what we could possibly do about that from the position of policymaking. I propose that we pry these questions apart.

*The Welfare Knowledge Problem* (WKP): What set or subset of preferences, if fulfilled, would lead to one’s maximal welfare?

*The Policy Knowledge Problem* (PKP): What set(s) of subset(s) of preferences could non-objectionable nudging aim at fulfilling?

The literature has suffered from the blending of these two questions. Efforts to engage the PKP invariably think that they have to answer the WKP first. There is no prima facie reason to suppose that the answer to these questions is the same or that the WKP has to be answered before progress can be made on the PKP. It may be the case that it is objectionable for nudges to aim at the satisfaction preferences that would make constituents the most well-off. This, I believe, should not come as a surprise since these questions are aimed at the fulfillment of two different values. One aims at the value of welfare, and the other at values of governance and autonomy. While autonomy is surely instrumentally related to welfare, many recognize that it has its own intrinsic value.

It becomes clear that these problems are distinct when we recognize that they have different success conditions. When we separate these two knowledge problems, we can critique answers to them according to their success conditions (and not the success conditions of the other knowledge problem).

***2.1 What would it take to offer a satisfactory answer to the* Welfare Knowledge Problem*?***

*The Welfare Knowledge Problem*: What set or subset of preferences, if fulfilled, would lead to one’s maximal welfare?

Hausman argues that economists can answer this question in one of two ways, either by claiming that preference satisfaction *constitutes* welfare or, as he prefers, claiming that the satisfaction of some preferences is *good evidence* for welfare.[[9]](#endnote-9) It is surely true that there is an answer to the WKP, regardless of whether or not preference-satisfaction is the correct theory of welfare. There is some set of preferences, the fulfillment of which will make one better off than the fulfillment of another set of preferences. Those might be preferences for internal states, outcomes, character traits, or lifestyle narrative arcs. They might be the person's actual real set of expressed preferences. However, either a constitutive or an evidentialist account of preference-satisfaction as welfare will have to measure up to the well-known challenges of preference-satisfaction theories and the new challenges posed by behavioral science.

A satisfactory answer to the WKP has to respond to the well-known challenges to preference-satisfaction (or desire-satisfaction) theories of well-being. Of course, the most pressing objection to welfare being realized by our preference satisfaction is that some of our preferences have manifestly bad outcomes when satisfied.[[10]](#endnote-10) This is because our actual (revealed through choice) sets of preferences include things like *uninformed preferences, maladaptive preferences,* and *hedonic adaptive preferences.*

Uninformed preferences, when satisfied, can lead to manifestly bad outcomes: we can imagine the preference for a blueberry pancake over a buttermilk pancake when one doesn’t know they are deathly allergic to blueberries. The satisfaction of some kinds of adaptive preferences (such as maladaptive and hedonic adaptive) also do not appear to be welfare-conducive. Maladaptive preferences are preferences for bad situations that one has grown accustomed to, never been without, or had to adjust to for survival. This is why we find women who seem to ‘prefer’ (to continue to seek out relationships with) men who abuse them. Hedonic adaptive preferences, on the other hand, are preferences that iterate once fulfilled. A preference for *one more pair of shoes*, *a new car*, or *a bigger house* are preferences that, when satisfied, tend just to produce another identical, unsatisfied preference.

Aside from well-known objections to preference-satisfaction theories of well-being, behavioral science has given us new reasons to be concerned that our revealed preferences don’t track our welfare. Instead of acting purely based on our preferences, we are pushed and pulled by our heuristics and biases, responding to non-relevant information to make our choices. Moreover, many of us turn out to have preferences that are inconsistent and, at times, contradictory - we do and don’t prefer to save our money. We do and don’t prefer to quit smoking. When we have these inconsistent preferences, the impulse is to label one of them as *mistaken*, but there are good reasons to be skeptical about our ability to decide *which* preference is mistaken. The policymaker is not in a privileged position to the agent in determining when an error is being made.[[11]](#endnote-11)

The central question then is, if well-being is not tied to the satisfaction of *all* of one’s preferences, then it must be tied to some subset (or non-actual set) of preferences. If it is some subset, then we need a non-circular reason to suppose that the satisfaction of some preferences leads to welfare and others don’t. For example, if someone prefers (*smoking* to *not smoking*) and (*not having lung cancer* to *having lung cancer*), we can’t merely say that the second preference constitutes well-being, whereas the first doesn’t *because it leads to increased welfare.*

An answer to *The Welfare Knowledge Problem* has to offer us a non-question-begging rationale to think that satisfaction of some preferences leads to increased welfare, while the satisfaction of other preferences does not. Or, if one prefers the evidentialist account, a reason to think that the satisfaction of some preferences is good evidence for welfare, while the satisfaction of other preferences is not. Otherwise, it has to offer some reason to suppose that people are better off given the satisfaction of their uninformed, maladaptive, hedonic-adaptive, and inconsistent preferences.

Authors continue to offer creative answers to the WKP. Bernheim suggests that only our “direct judgments” (our ends as opposed to means) are relevant to our welfare.[[12]](#endnote-12) DesRoches suggests that it is the fulfillment of our value-based preferences.[[13]](#endnote-13) Fabian and Dold argue that we should be thinking about self-actualized preferences.[[14]](#endnote-14) Sunstein is still trying to save laundered preferences - counterfactual preferences fostered under ideal conditions - long abandoned by Hausman.[[15]](#endnote-15)

These accounts, as they offer answers to the WKP, ought to be judged according to the success criteria laid out above. However, success according to these criteria does not mean they are good candidates for answers to the PKP, where we will now turn our attention.

***2.2 What would it take to offer a satisfactory answer to the* Policy Knowledge Problem*?***

*The Policy Knowledge Problem*: What set(s) of subset(s) of preferences could non-objectionable nudging aim at fulfilling?

What counts as a satisfactory answer to this question will turn on what “non-objectionable” means. However, here I will be referring to the primary objections lobbied against nudges: that they are problematically paternalistic and that they undermine the autonomy of the constituents. Therefore, for our purposes, non-objectionable nudges are those which are not problematically paternalistic and do not undermine autonomy.[[16]](#endnote-16)

What preferences does a policy have to aim at to be non-paternalistic? This was the purpose of the original “as judged by themselves” clause from Thaler and Sunstein’s *Libertarian Paternalism.*[[17]](#endnote-17)Since nudges are geared to make people better off by their own standards, then to that extent, they are not problematically paternalistic. A policy is problematically paternalistic if it substitutes the policymaker’s idea of welfare for the constituents. To be non-problematically-paternalistic (hereon, ‘non-paternalistic’ for brevity), then, a policy has to appeal to the constituent’s idea of their welfare instead of merely the policymaker's idea of their constituent’s welfare. The preferences that nudgers take as their object must be, in a robust sense, preferences *of* the individual instead of normatively imposed preferences. If choice architects appeal to what an individual *should* want, as decided by the policymaker, then the resulting policy is clearly objectionably paternalistic according to libertarian paternalism. Non-paternalistic nudging, then, aims at preferences that people *do* have instead of preferences that people *should* have.

But that may, unfortunately, be the easiest part of this question to answer. Perhaps harder to answer is this: what preferences would a nudge have to aim at to keep from undermining autonomy?

There are two ways to conceive of what it might mean to respect autonomy, or “the control an individual has over his or her own evaluations and choices.”[[18]](#endnote-18) One way to understand this demand is that the individual, *at the choice point*, needs to be able to reject the nudge, i.e., choose for himself. *Which* choices we are nudging people towards need not weigh into that question. It is a question of which *tactics* we can use when trying to nudge people. Uncontroversially, authors answer: only the ones that they could reasonably resist. This conception of protecting autonomy is unrelated to the question at hand.

However, there is another way to understand the demand that nudges respect autonomy. An individual's ability to choose for himself extends beyond the choice point toward the shape of his life. To respect his autonomy, a nudge must respect the individual's ability to decide for himself, make mistakes, and engage in life planning.[[19]](#endnote-19) The intervention must not *create* a preference within the individual that interferes with his ability to make and remake himself. “This contravenes the ‘libertarian’ in libertarian paternalism, which requires not just the preservation of choice but also that the nudge helps the agent to realise their own preferences.”[[20]](#endnote-20)

Imagine a high school student who enjoys tinkering with electronics who, in a nudge-free world, would end up working under an apprenticeship towards an electrician’s license after graduation. An autonomy-undermining nudging agenda would be one that uniformly offers robustly effective college nudges to all secondary students. When he was nudged into attending university, this came at the cost of allowing the boy to choose freely how to plan his life, even if he ended up endorsing the preference for university over apprenticeships after the fact.[[21]](#endnote-21)

We know that choice points are often the catalyst for creating preferences, as our given preference sets are not complete. As such, choice architects create choice points that have the opportunity to create post-hoc endorsed preferences, such as the boy’s endorsed preference to attend college while he was there.[[22]](#endnote-22) While the individual may not regret the choice, it did undermine his ability to plan for himself. Nudging that targets preferences that the individual does not have either way then has the potential to be autonomy undermining. When we set choice architecture to elicit a certain response, instead of nudging people towards (or away from) a pre-existing preference, we may be constructing and instilling one in them.[[23]](#endnote-23)

A satisfactory answer to the PKPwould identify a set of preferences that belongs in an important way to the individual instead of one that the policymaker thinks should (or wishes would) belong to the individual. Moreover, it should not aim at *creating* preferences that undermine an individual's ability to plan his life. Otherwise, it has to give us a non-paternalistic reason to think that this is not especially threatening to one’s conception of their own well-being or a special kind of imposition on the way to planning one’s life. That is, an answer can not just appeal to giving the individual the *right* desires or *right* life plan (from the point of view of the policy maker), they have to give an answer that is justifiable to the constituent.

And perhaps obviously, if the individual is already succeeding at the fulfillment of this set of preferences, then there is no role for choice architects to intervene. An answer to the PKP needs to identify a set of preferences *of* the agent *that* the agent fails to satisfy; after all, behavioral welfare economics is premised on our making *mistakes*. Only then does it make sense to introduce nudging strategies. If the only preference set we can identify that meets the non-objectionable criteria offered above is one that the agent is already fulfilling, then there are no behavioral interventions to be suggested.

Of course, there are also concerns about effectiveness and unforeseen consequences when it comes to nudges[[24]](#endnote-24) and those concerns need to be addressed. However, they are not concerns about the object of nudges but rather about the implementation. As such, I leave those concerns out of this discussion.

***2.3 Detangling***

Authors addressing ‘the’ knowledge problem in behavioral welfare economics have (on the whole) thus far not been clear about which of these knowledge problems they are addressing or on which grounds they are objecting to other theories. As I hinted at in the beginning, I believe we can observe a pattern in this literature, where theories are *offered* according to the success conditions of the WKP, and *assessed* according to the success conditions of the PKP.

Let’s use laundered preferences as an example. Laundered preferences - those counterfactually formed under ideal conditions - are attempts to explain why some preferences, and not others, are welfare-enhancing. Claims about why laundered preferences, informed preferences, or preferences formed under ideal epistemic conditions are welfare tracking are going to have to answer to the success conditions of the WKP.

However, laundered preferences are critiqued, not just under the success conditions of the WKP but also for not answering to the success conditions of the PKP.[[25]](#endnote-25) This is, of course, because we have a mistaken understanding of the knowledge problem, and we feel that an answer to the WKP *should* give us an answer to the PKP. The critiques are not misplaced as the authors present them as though they do give us answers about what kinds of policy we are justified in creating, without acknowledging for themselves the separate conditions that need to be satisfied for an account to be successful on this ground.[[26]](#endnote-26)

Clarity about which problem we are addressing can blunt unnecessary criticisms. Mark Fabian and Malte Dold make the same mistake: “We have established how self-actualisation can lead to the emergence of relatively stable preferences that are sufficiently related to welfare to form the basis for nudging.”[[27]](#endnote-27) Their *agentic preferences* offer us an answer to the WKP: those preferences that take us towards self-actualization - towards autonomy, competency, and belonging - are those related to our welfare. But again, these preferences don’t actually exist (at least not in most of us). Self-actualized preferences are an ideal, they are not the preferences *of* the agent, and thus don’t measure up to the success criteria for the PKP. If they professed to be answering the WKP instead of the PKP, they could save themselves from unnecessary objections.

**3. Prioritizing the PKP**

As I’ve said above, I hypothesize that we find ourselves running in this loop because of an implicit belief that an answer to the WKP will get us an answer to the PKP, or at the very least, we can’t have an answer to the PKP without it. Those who have argued that we should abandon the WKP project[[28]](#endnote-28) seem to think that we need to also abandon the idea of a non-objectionably-paternalistic nudge. I hope to have convinced you that these questions and the conditions needed to satisfy them are separate and distinct. As such, I will suggest a different path forward for the PKP, one that doesn’t require an answer to the WKP.

I start with a simple observation: If there is any non-objectionable nudge, it is the nudge that the individual deploys on himself. A self-nudge retains autonomy in the important (life planning) way and is definitionally non-paternalistic.[[29]](#endnote-29) To identify the preferences that a non-paternalistic nudge could aim at, we should think about the preferences that self-nudges aim at: which are the objects of a self-nudge?

What is the object of a self-nudge? What does it reveal? A self-nudge is an attempt to *change* behavior to increase the likelihood that some outcome does (or, more often, does *not*) occur. Sometimes, self-nudges are pre-commitment devices. However, self-nudges are not always an attempt to follow through on a long-term or future-payoff preference. They are often attempts to change preferences themselves. This extends beyond mere pre-commitment devices.

Behavioral economists have been talking about pre-commitment devices for decades.[[30]](#endnote-30) Pre-commitment devices are used to follow through on long-term preferences when one knows that they will have short-term preferences (and suffer from present bias) that undermine that goal in the meantime. Christmas clubs are a classic example: you prefer to have money to buy Christmas presents for your kids, but you know that you won’t prefer that every time you go to the store between now and Christmas, so you use a pre-commitment device to squirrel away some money each month that you can’t spend before the holiday. A pre-commitment device is a nudge that people deploy on themselves so that they can carry out a long-term goal when they have inconsistent preferences in the short term.

However, there are other kinds of self-nudges. Sometimes, people self-nudge when they are trying to go through a process of preference transformation. In these cases, people often don’t know what they want, but they know they don’t want what they have. An attempt to change behavior (or preferences) does not ensure that the person prefers the new outcome it only shows that the person has a preference to move away from the old outcome.

***3.1 Repudiated Preferences: what we can learn from self-nudges***

Self-nudges come in all shapes and sizes: Throwing away just-purchased packs of cigarettes, hanging around people that don’t smoke tobacco, keeping backup e-cigarettes (vapes) on hand. Here's an all too common tale: someone engages in all the aforementioned self-nudges and starts vaping because they want to stop smoking tobacco: the preference was to *quit* smoking, and vaping seemed like it could help to accomplish that goal. However, many people who have made this switch *regret it:* they have ended up with a higher nicotine consumption and a harder habit to kick. And yet, they still prefer not to smoke cigarettes.

What can we learn from this self-nudge? We can’t link self-nudging actions to *true* preferences because it is possible that the person does not prefer the new outcome (vaping) and may not think their preferences, after having gone through this process of preference reformation, are conducive to their welfare. However, we learn something about what the person *does not* prefer (smoking cigarettes). Self-nudges teach us about preferences that the individual has that he is rejecting as being conducive to his welfare. For this reason, I introduce *repudiated preferences*:

**Repudiated preferences**: A repudiated preference is one that the person has but also understands as not being a preference whose satisfaction will lead to their own wellbeing.

What is distinctive about a repudiated preference is that it is one that *the individual understands for themselves* *that it is mistaken*. A preference can be repudiated regardless of whether other people around the individual think that the preference is mistaken. I take no theoretical position on whether that is because the repudiated preference is uninformed, short-sighted, maladaptive, or otherwise faulty. Any of these things could count as repudiated preferences. Repudiation is not time-sensitive: repudiation of a preference may occur before, during, or after one acts upon it. To have repudiated preferences is not uncommon. They can exist whether or not one acts on them. I have not had a cigarette in over 18 months, and yet, sometimes, I would prefer to be smoking a cigarette. My preference to have a cigarette is a repudiated preference: I can identify and understand for myself that it is not a preference whose satisfaction will lead to *my* increased wellbeing. This takes no stance on whether or not the preference for a cigarette is a repudiated preference for anyone else. Repudiated preferences are necessarily subjective: whether a preference is repudiated is contingent on the specific individual who has it.

Examples of reputed preferences can vary quite wildly. When I sign on to Facebook.com to escape my loneliness, knowing that it will not help me escape my loneliness, I am acting on a repudiated preference. When I buy a pack of cigarettes after having just thrown away a pack of cigarettes, I am acting on a repudiated preference. My preference to eat potato chips while writing this essay (hoping that it will make me a better writer since I can’t turn to cigarettes) is a repudiated preference. When I voluntarily work 60 hours a week after I affirm to myself that time with my family is more important than my wages, I act on a repudiated preference.

These are easy examples because we have access to the ends that are being frustrated (or at least not satisfied) by the means: to feel a sense of connectedness, to not be a smoker, to be a better writer, to spend time with my family. However, we don’t always have access to these ends. Sometimes, we merely have access to the frustration. We work too much and realize that we are not living - we don’t know how we would like to be living, but we know that working too much isn’t it. We feel bad after we go on Facebook, and we don’t know why, but we know we don’t want it.

This clarification helps to distinguish repudiated preferences from what Frankfurt coined “second order desires,” or we might say, second order preferences.[[31]](#endnote-31) One way to interpret an example above is to say that I have a first order preference to work too much, another first order preference to spend time with my family, and a second order preference to *‘prefer the preference to spend time with my family over working*.’ Talking about self-nudges as revealing second order preferences is misleading for two reasons, i) we often know that a preference is not conducive to welfare before we know which preferences *would be* conducive to welfare, and ii) we are often (as in the vaping case) wrong about which preferences would be (second order) preferable to the preferences that we have currently. In other words, second order preferences, again, commit us to the existence of “true preferences.” For these reasons, I offer a picture that lets us identify a mistake for ourselves (a non-welfare-conducive preference) without having to know what our own “true preferences” would be. This I take as being non-controversially descriptive of how we actually live our lives and update our preferences.

If we only had two options for how to live our lives, perhaps self-nudging would reveal a stable set of preferences that we could call “true preferences.” To smoke or not to smoke? If I have a repudiated preference for smoking, then I must have a “true preference” for being a non-smoker. But being a ‘non-smoker’ has little content. It might mean vaping, using nicotine patches, or gum. It might mean eating a tin of Altoids a day, coupled with a near-intravenous supply of coffee, trying to stay awake and alert in meetings. It might mean putting on 50lbs due to my sudden increased sugar intake. It might mean yelling at my kids more than I would like while the withdrawals wear off. It could turn out that all the things that I do in my attempts to be a non-smoker are *also things that I would prefer not to be doing*.

However, we live in a world with many, many options, and we often don’t have access to what the outcomes of those options will be - our expected utility has little to do with the experienced utility. When we choose, we find out that we are often wrong about what we thought we wanted. Given these two things together, that attempts to change behavior reveal a preference away from some current situation and that they *don’t* reveal a preference for the option that they are getting as they attempt to change, what can we say about the revealed object of self-nudging? A self-nudge is often evidence of a preference that the individual understands to be a mistake - to be non-welfare-conducive. A self-nudge to quit smoking is a self-identified mistaken preference for smoking, it is not, however, necessarily a preference for some specific alternative, including the chosen alternative.

Recall that the paternalism critique in the nudging literature is that the policy maker is not in a better position than the agent to identify mistakes when contradictory preferences are present.[[32]](#endnote-32) A non-paternalistic choice architect does not decide for the agent which preferences he *should* have or which ones he *should* fulfill. Instead of trying to base behavioral welfare policies on policy-maker-identified mistakes, we could instead try to base them on self-identified mistakes: repudiated preferences.

Repudiated preferences retain the normative “should” that we are so drawn to when trying to answer the WKP. However, they avoid the problematic paternalism insofar as it is a ‘should’ that the agent is imposing on herself instead of one imposed by the policymaker. As Fabian and Dold note, our preferences should change over time; it is a sign of growth. “As it matures, individuals will detach from values and behaviours associated with negative affect and deepen their engagement with those associated with positive feedback.”[[33]](#endnote-33) It is unlikely that an individual possesses completely self-actualized preferences. It is unlikely that we, as policymakers, will know which of their preferences are self-actualized. It is unlikely that the individual knows which of his own preferences are self-actualized. If we are waiting for a complete normative welfare thesis to act, to locate the one true set of preferences, there will be no room for action. However, without knowing which of his own preferences are self-actualized, the individual can identify which of his preferences are surely mistakes.

Instead, if we want an answer to the PKP, we should pay attention to individuals identifying mistakes for themselves. Individuals themselves live with theoretical uncertainty. Policymakers should not be in the business of imposing theoretical certainty on those lives. Few of us, if any, move through our days with a robust idea that we ought to *maximize utility* from our choice sets. We try to serve those things that are good and try to discard those things that are no longer serving us. We find out that some parts of our lives are not what we would prefer them to be, and we move away from those things towards uncertain futures. Agents go through this iterative process, as Fabian and Dold describe, and in time maybe (hopefully) find themselves actualized with *agentic preferences*, but first they learn a lot of things about what they *don’t* want.

When an individual is trying to change a pattern, break out of a cycle, or put distance between himself and some behavior, we are not justified in making claims about what they are after. Trying to infer from mistakes back to true preferences can send us into a normative tailspin back to the WKP. If someone prefers to quit smoking, he has not revealed what he prefers to smoking. Charming white teeth? Lack of a smoker's cough? General health? Happy wife? If we want to make progress on the PKP, we are better off paying explicit attention to what people are telling us are their *mistakes* than trying to infer from their mistakes what they want their self-actualized life to look like.

We do not need to try to infer when an individual *would* affirm that they have made a mistake in an opaque, counterfactual, laundered way.[[34]](#endnote-34) We merely need to trust that individuals are making progress when they themselves identify mistakes. When they engage in *small group de-biasing* by willfully discussing their choices with their communities and *self-regulation* strategies, both including pre-commitment devices and robust self-nudging,[[35]](#endnote-35) they are attempting to correct for something *they themselves see as a mistake.*

Against Rizzo and Whitman’s criteria for how to identify mistakes,[[36]](#endnote-36) it can seem like we will never be able to identify *any* mistake from the outside. However, this is again letting the perfect be the enemy of the good. People sometimes tell us that they are making mistakes. Sometimes, they tell us through their actions and we should put more time and effort into studying those actions, those inconsistencies and self-nudges that articulate a personal repudiation: *this preference is not for me, and I would like to shed it.*

***3.2 Application, Difficulties, and Further Work***

If we want to make progress on creating non-objectionable nudging policies, we should begin with trying to satisfy the success conditions of the PKP instead of the WKP. I believe that repudiated preferences are a step in the right direction. Here I will offer a brief sketch of what further action in this direction might look like. There are clear challenges to identifying and creating policies based on repudiated preferences, and I discuss those and reasons to think those challenges could be overcome.[[37]](#endnote-37)

Recall the claim from earlier: a policymaker cannot nudge an individual closer to his welfare without knowing what the optimal choices for that individual are.[[38]](#endnote-38) It is exactly here that we turned off course. When we know what the *suboptimal* choices for an individual are, then progress can be made. This is, in fact, how an individual makes progress for himself. Not by identifying the optimal choices but by identifying the suboptimal choices and moving away from them. We don’t need to know how far off the mark someone is in their mistake. Their goal is a direction, not an endpoint.[[39]](#endnote-39) Instead of being bewildered that we can not non-paternalistically identify and respond to *all* mistakes, we should be encouraged that we may be able to respond to some. We have the opportunity to do this when we listen to people reporting their repudiated preferences.

Cartwright and Hight seem to suggest that there is no way to conceive of an individual identifying his own actions or preferences as welfare-conducive or not welfare-conducive - that there is no way to make sense of an individual's ability to do this.[[40]](#endnote-40) However, this move, too, has an uncomfortably paternalistic flavor. To tell an individual that what he is identifying as a mistake is *not* a mistake is just as unjustified, or as it seems to me - *more* unjustified, than telling an individual that he *is* making a mistake when he thinks he is not. I would rather allow the possibility that people are mistaken in identifying their repudiated preferences and give them the opportunity to learn that through trial and error than not trust an individual's ability to do this at all.

We can envision a role for nudging, where nudges are designed to push people away from their repudiated preferences. This, of course, does not lend itself to neatly packaged policies. There are at least two main challenges facing this vision: First, we have to have a mechanism for gaining knowledge of someone's repudiated preferences. One challenge here lies in the generalizability of repudiated preferences. Since repudiated preferences are inherently subjective, they are tied to individuals and can be different for different people. There will be no ‘one size fits all’ policy that we are able to justify on the grounds of repudiated preferences. Second, we have to find ways to nudge people *away* from those things without also simultaneously nudging them *towards* something else. The policies that we should expect to be justified under the framework of repudiated preferences are not going to look like nudge policies that we are familiar with. Nudges and other behavioral interventions that are non-paternalistic and respectful of autonomy will almost definitely be more expensive, more resource-heavy, and less effective than Sunstein and Thaler’s low-cost, high-reward *Libertarian Paternalism*.

The further, socially, one gets away from an individual, the less able one will be to know what that person's repudiated preferences are. To infer knowledge of repudiated preferences from individuals that one is not personally familiar with, we have to use a lot of imperfect information. There are different ways we could consider doing this. One way would be through trying to use data, economic or otherwise, to try to pick out behaviors that are likely to be self-nudges and observe what the individual is trying to nudge themselves away from.

There has been some discussion about nudge personalization in the literature using access to big data, among other things.[[41]](#endnote-41) Much of that discussion has had the goal of improving the *effectiveness* of nudges. However, in other instances, it is introduced to overcome problems of generalizability - when policymakers determine that some individuals need different nudges than others (setting different defaults for people dependent on whether they are determined to be ‘undersavers’ or ‘oversavers’). What I am suggesting is a little different. Instead of using data to register when we think people are making mistakes (under-saving or over-saving), we could use data to look for signs that individuals are acting on repudiated preferences, signs that they are nudging themselves.

This might take the form of looking for signs that people are trying to nudge themselves out of particularly vicious habits. For example, if someone is looking up rehabilitation facilities with available beds, information on smoking cessation, or tips on building long-term wealth, these seem like safe signs that have repudiated preferences. This information could then be used to design personalized nudges.

An alternative to trying to infer repudiated preferences from economic or behavioral data would be to have more options for opportunities for individuals to explicitly register their repudiated preferences or to ‘*opt-in’* to certain nudges. This might sound far-fetched, but we already have some precedent for this kind of policy. For example, in most states, the state government offers services for individuals to self-exclude from gambling. This is a voluntary sign-up process that individuals can choose to go through when they come to understand their preference to gamble as a repudiated preference. When you file self-exclusion paperwork with the state, you are banned from physical casinos and from making accounts on gambling websites. This is one place where we have set up a system for people to register their repudiated preferences and ask for help from the government. Of course, outright bans such as this are not typically the policy that behavioral welfare economists tend to be interested in.[[42]](#endnote-42) However, one might imagine a similar system being available where we can register our repudiated preferences and ask for nudges. *[Select to receive long-term wealth-building nudges] [Select to receive health nudges] [Select to receive green nudges]*

Another virtue of opt-in nudges is that they would respect individuals who, even if they have repudiated preferences, would prefer not to have help in augmenting their behavior. For individuals who would prefer to struggle to thrive on their own, without the help of nudges, there might be no general nudging policy proposal that could respect their preferences and autonomy. Personalized nudges could, in theory, be both non-paternalistic and respectful of autonomy, and they may be the only nudges that are. However, personalized nudges come with their own challenges, including high costs and the potential for injustice in treating citizens differently, but these are new and different challenges than the classic objections.

The second challenge is that we must still be careful to avoid *creating* preferences in the individual in a way that problematically interferes with their ability to plan their lives for themselves. It also requires us to think of creative new ways that we might nudge. Many nudges are designed to *speed up* the decision-making process so that the ‘right’ decision is more apparent to the chooser than the ‘wrong’ decision(s). We instead could think of nudges that are designed to *slow down* the decision making process. For the aforementioned individual who was looking up methods for smoking cessation, they might get a pop-up nudge on their Google Maps when they search for “smoke shops near me:”

*Are you sure? [yes, continue] [no, exit]*

The “are you sure?” nudge is designed not to recommend any course of action to the individual but to remind them that this is a preference that they previously treated as a repudiated preference. This kind of nudge has many challenges. For one, this kind of personalization, in design and in delivery, is significantly resource-heavy. This is costly. Secondly, this nudge is invasive. To be clear, it is not more invasive than the data collection that we are already subject to whenever we use the internet. However, some might object to the government having this kind of data access to its constituents. Again, these are new and different challenges than the classic objections.

This kind of nudge, however, is also not unprecedented. I recently experienced a similar nudge offered by the state of Arizona. Arizona’s Medicaid system (AHCCCS) keeps track of people who sign up for Medicaid and what kind of referrals they ask for. If one fails to use their health insurance for some period of time, AHCCCS will call and ask the individual if they can help them to sign up for health services they did not follow through on. This nudge respects the individual's preferences in that the individual had to enroll in AHCCCS and request services before this kind of nudge was triggered. It does not create preferences in the individual. It merely lowers barriers to following through on preferences the individual has already expressed.

This is clearly not an exhaustive list of the challenges facing us as we try to create nudge policies based on repudiated preferences, nor are my replies comprehensive. However, I hope that this sketch of what a research agenda along these lines might look like is reason for optimism that we need not abandon the goal of being policymakers who are non-paternalistic and respect autonomy.

As well as pursuing research agendas along these lines, we might consider how we might build it into our systems to let people make and profess mistakes. Instead of assuming that an individual is a consistent utility function who can’t, won’t, or shouldn’t change his mind, perhaps behavioral welfare economics can focus on building into systems opportunities for nonjudgmental change. Given the costs associated with changing behavior - that it requires effort, time, and often uncomfortable emotions - we tend not to attempt to change behavior when we can avoid it. Perhaps we should try to build systems that, instead of nudging someone one way or another, lower the cost of changing their minds.

**4. Conclusion**

The conceptual landscape that makes up the Hayekian Knowledge Problem in behavioral welfare economics has been muddled by a confusion of the theoretical and practical problems related to true preferences. Here, I have separated out these two different problems, with two different success conditions, to show where we can start to find more clarity. I argue that an impulse to try to answer the practical (policy) knowledge problem by first answering the theoretical (welfare) knowledge problem is misguided. We don’t have to wait until we have found *true preferences* before we can start to think about what non-objectionable nudges might look like. I offer some thoughts about how we can reconceptualize this landscape using *repudiated preferences* to dispel some of the confusion around the practical issue. I have tried to present us with the clearest case: a non-objectionable nudge will be that which is aimed at keeping an agent from fulfilling a preference that *he professes* is a mistake. Without generalizability, large-scale implementation would likely require personalized nudges. Whether those nudges would be effective at changing behavior or cost-effective to implement are separate questions. I sketched what a few nudges taking these insights into account might look like, but by no means offer a full policy proposal. Further research along these lines is needed. Furthermore, I suggest that a productive line of research within normative economics might be to try to design ways that we can identify individuals engaging in self-nudging behaviors and build more chances into the system for people to register repudiated preferences or to ‘opt-in’ to nudges. [[43]](#endnote-43)

1. Daniel M. Hausman, “Banishing the inner Econ and justifying paternalistic nudges,” *Behavioural Public Policy* (2022): 1–12. https://doi-org.ezproxy1.lib.asu.edu/10.1017/bpp.2022.19. [↑](#endnote-ref-1)
2. Mario J. Rizzo and Glen Whitman, “The unsolved Hayekian knowledge problem in behavioral economics.” *Behavioural Public Policy* 7, no. 1 (2023): 199-211. [↑](#endnote-ref-2)
3. Cass R. Sunstein and Richard H. Thaler, “Libertarian paternalism is not an oxymoron,” *The University of Chicago Law Review* 70, no. 4 (2003): 1159-202. Also, Richard H. Thaler and Cass R. Sunstein, *Nudge: Improving decisions about health, wealth, and* *happiness.* (New Haven: Yale University Press, 2008). [↑](#endnote-ref-3)
4. Cass R. Sunstein, “Hayekian behavioral economics,” *Behavioural Public Policy* 7, no.1 (2023): 185. (emphasis added) [↑](#endnote-ref-4)
5. Alexander C. Cartwright and Marc A. Hight, “‘Better off as judged by themselves’: a critical analysis of the conceptual foundations of nudging,” *Cambridge Journal of Economics* 44, no. 1 (2020): 52. [↑](#endnote-ref-5)
6. B. Douglas Bernheim, “The Good, the Bad, and the Ugly: A Unified Approach to Behavioral Welfare Economics.” *Journal of Benefit-Cost Analysis* 7, no. 1 (2016): 12–68. [↑](#endnote-ref-6)
7. C. Tyler DesRoches, “​​​Value Commitment, Resolute Choice, and the Normative Foundations of Behavioural Welfare Economics,” *Journal of Applied Philosophy* 37, no. 4 (2020): 562-577. [↑](#endnote-ref-7)
8. Mark Fabian and Malte Dold, “Agentic Preferences: A Foundation for Nudging When Preferences Are Endogenous.” *Behavioural Public Policy* (2022): 4. [↑](#endnote-ref-8)
9. Daniel M. Hausman, “Why satisfy preferences?” *Papers on Economics and Evolution*, 1124 (2011). Also, Daniel M. Hausman, *Preference, Value, Choice, and Welfare,* (Cambridge, UK: Cambridge University Press, 2012). [↑](#endnote-ref-9)
10. Chris Heathwood, “Which desires are relevant to well-being?” *Nous* 53, no. 3 (2019): 644–688. [↑](#endnote-ref-10)
11. Bernheim, “The Good, the Bad, and the Ugly.” [↑](#endnote-ref-11)
12. Bernheim, “The Good, the Bad, and the Ugly.” To be clear, I don’t think this is the project Bernheim takes himself to be interested in doing, however, he spends a good deal of time making sure that he is meeting these success conditions. [↑](#endnote-ref-12)
13. DesRoches, “Value Commitments.” [↑](#endnote-ref-13)
14. Fabian and Dold, “Agentic Preferences.” [↑](#endnote-ref-14)
15. Sunstein, “Hayekian behavioral economics,” and Daniel M. Hausman “On the Econ within,” *Journal of Economic Methodology* 23, no. 1 (2016) 26-32. [↑](#endnote-ref-15)
16. While some policy makers, and supporters of BPP might not find these to be particularly pressing concerns, they tend to not be the people who care particularly much about the knowledge problem. If one is fine with acting paternalistically, they won’t need to discover what “true preferences” are to make their policy recommendations. Moreover, it might turn out to be the case that nudging is never ethically permissible for reasons beyond these (e.g. because some agents might be categorically opposed to having this kind of government assistance in their struggle to thrive). [↑](#endnote-ref-16)
17. Sunstein and Thaler, “Libertarian paternalism is not an oxymoron,” and Thaler and Sunstein, *Nudge.*  [↑](#endnote-ref-17)
18. Daniel M. Hausman and Brynn Welch. “Debate: To Nudge or Not to Nudge.” *Journal of Political Philosophy* 18 (2010): 128. [↑](#endnote-ref-18)
19. Malte Dold, “Back to Buchanan? Explorations of welfare and subjectivism in behavioral economics,” *Journal of Economic Methodology* 25, no. 2(2018): 160-178. [↑](#endnote-ref-19)
20. Fabian and Dold, “Agentic Preferences,” p3. [↑](#endnote-ref-20)
21. This, of course, is the kind of influence that we allow our families and friends to have in our lives, the kind that we practice with our children, but there are good reasons to suppose that our being treated like this by our governments is much more problematic than that. [↑](#endnote-ref-21)
22. Despite seeming problematic, this process is endorsed by: L. A. Paul and Cass R. Sunstein, “'As Judged By Themselves': Transformative Experiences and Endogenous Preferences.” (2019) Available at SSRN 3455421. [↑](#endnote-ref-22)
23. See discussion of this in, among others, Bernheim, “The Good, the Bad, and the Ugly,”; Fabian and Dold, “Agentic Preferences,”; and Robert Sugden, *The Community of Advantage: A Behavioural Economist's Defense of the Market.* (Oxford: Oxford University Press, 2018). [↑](#endnote-ref-23)
24. Rizzo and Whitman, “The unsolved Hayekian knowledge problem.” [↑](#endnote-ref-24)
25. Rizzo and Whitman, “The unsolved Hayekian knowledge problem.” and Gerardo Infante, Guilhem Lecouteux, and Robert Sugden, “Preference purification and the inner rational agent: a critique of the conventional wisdom of behavioural welfare economics”, *Journal of Economic Methodology* 23, no. 1 (2016): 1-25. [↑](#endnote-ref-25)
26. Sunstein, “Hayekian behavioral economics.” [↑](#endnote-ref-26)
27. Fabian and Dold, “Agentic Preferences,” 12. [↑](#endnote-ref-27)
28. See Sugden, *The Community of Advantage*; Hausman, “Banishing the inner Econ,”; Rizzo and Whitman, “The unsolved Hayekian knowledge problem” [↑](#endnote-ref-28)
29. Samuli Reijula and Ralph Hertwig. “Self-nudging and the citizen choice architect.” *Behavioural Public Policy* 6, no. 1(2022): 119-149. [↑](#endnote-ref-29)
30. Richard Thaler, “Toward a positive theory of consumer choice*,” Journal of Economic Behavior & Organization* 1, no. 1 (1980): 39-60. [↑](#endnote-ref-30)
31. Harry Frankfurt, “Freedom of the Will and the Concept of a Person,” *The Journal of Philosophy* 68, no. 1 (1971): 5-20. [↑](#endnote-ref-31)
32. Bernheim, “The Good, the Bad, and the Ugly,” and Rizzo and Whitman, “The unsolved Hayekian knowledge problem.” [↑](#endnote-ref-32)
33. Fabian and Dold, “Agentic Preferences,” 10. [↑](#endnote-ref-33)
34. As discussed in Bernheim, “The Good, the Bad, and the Ugly.” [↑](#endnote-ref-34)
35. For discussion of these phenomenon, see Rizzo and Whitman, “The unsolved Hayekian knowledge problem,”; and Reijula and Hertwig, “Self-nudging and the citizen choice architect”; and Mario J. Rizzo and Glen Whitman, “Deficient Foundations for Behavioral Policymaking,” in *Escaping Paternalism: Rationality, Behavioral Economics, and Public Policy* (Cambridge: Cambridge University Press pp), 190-234. [↑](#endnote-ref-35)
36. Rizzo and Whitman, “The unsolved Hayekian knowledge problem,” [↑](#endnote-ref-36)
37. Thanks to an anonymous review for the suggestion to discuss challenges facing this research agenda, of which much of this section is a response to. [↑](#endnote-ref-37)
38. Rizzo and Whitman, “The unsolved Hayekian knowledge problem,” p4. [↑](#endnote-ref-38)
39. This should sound rather familiar to Aristotelians, when we find ourselves in excess or deficiency, we aim to move *away* from that excess or deficiency. The virtue, the endpoint, is only found by overshooting and circling back, something each of us must do for ourselves. [↑](#endnote-ref-39)
40. Cartwright and Hight, “Better off as judged by themselves.” [↑](#endnote-ref-40)
41. For a recent overview of this, see Stuart Mills, “Personalized Nudging.” *Behavioural Public Policy* 6, no. 1 (2022): 150–59. [↑](#endnote-ref-41)
42. Thaler and Sunstein do talk about this in *Nudge* (235), but it is in principle much different from the classic nudges that the debate is about. [↑](#endnote-ref-42)
43. I have special thanks to offer David Schmitz and his editorial team for making this a smooth and painless process, for the other authors in this issue for their discussion of my manuscript, for an audience at INEM in Venice who helped me start to think through these issues in a new way, for Tyler DesRoches for his endless encouragement and willingness to provide comments at the latest possible moment, and for Aubrial Harrington, a friend and colleague who is responsible for thinking of the term “repudiated preferences.” [↑](#endnote-ref-43)