Aggregating Local Preferences To Guide Policy[*]

Daniel J. Benjamin
Cornell University, USC, and NBER

Gabriel Carroll
Stanford University

Ori Heffetz
Cornell University and NBER

Miles S. Kimball
University of Michigan and NBER

First Draft: October 6, 2014
This Draft: October 9, 2014

Abstract

How could well-being data, for example those based on survey measures, be used for guiding policy? Exploring one direction, we analyze a mechanism that takes as inputs estimates of policy effects on different groups' utility proxies (constructed from the well-being data), and aggregates them into policy-change recommendations. We develop three justifications for the mechanism based on: an analogy with social welfare maximization, a formal equivalence to a voting procedure, and an axiomatic characterization. We show that iterated application of the mechanism has a stationary point that is Pareto efficient. Through analytic results and simulations, we assess potential limitations of the mechanism, such as its sensitivity to the units in which policies are measured.

JEL Classification: D69, H0, I38

Keywords: subjective well-being, well-being data, preferences, aggregation, policy, happiness

[*] We thank Greg Besharov, Steve Coate, Marc Fleurbaey, Nichole Szembrot, Maxim Troshkin, and E. Glen Weyl for valuable comments and suggestions. For excellent research assistance, we are grateful to Samantha Cunningham, Yuezhou (Celena) Huo, and Derek Lougee. For financial support we are grateful to Cornell's Samuel Curtis Johnson Graduate School of Management, USC's Center for Economic and Social Research, and NIH/NIA grant R01-AG040787 to the University of Michigan. E-mail: daniel.benjamin@gmail.com, gdc@stanford.edu, oh33@cornell.edu, mkimball@umich.edu.

Policymakers are expressing increasing interest in using data on individuals' well-being to guide policy. Much recent attention has focused on survey-based data. As part of a growing effort to collect such data, governmental organizations are including subjective well-being (SWB) questions in their national surveys. But how would such well-being data be used to make policy? In this paper we hope to contribute to answering this question by building on ideas briefly outlined in Benjamin, Heffetz, Kimball, and Szembrot (2013) and—as we recently found out—investigated more than thirty years earlier in unpublished work by Hylland and Zeckhauser (1980). We develop a mechanism that aggregates estimates of local effects of policy on individuals' utility proxies to yield recommendations of marginal policy changes.

The output of our mechanism is a proposed policy change relative to the current status quo policy vector. The main inputs into our mechanism are estimates, for each different group in the population, of the effect of small policy changes on a comprehensive well-being measure (formally, an ordinal utility representation of preferences). Each of the groups—also referred to below as "types"—consists of individuals with identical policy interests (formally, their ordinal utility is affected in the same way by policy). We do not address in this paper how to identify such groups, how to construct such a utility proxy, or how to estimate how it is affected by policy; we assume the existence of these inputs into our mechanism. This assumption is admittedly hopeful: while in principle the growing availability of individuals' well-being data may eventually enable government agencies to provide the inputs we assume, in practice many questions will first have to be addressed. These include: how to create a comprehensive measure of well-being (for example, one that is based on survey questions); what survey questions (or other indicators of well-being) it should be composed of; how its different components should be weighted relative to each other to create an overall measure of well-being; and how the effects of policy on this measure could be estimated. Recent evidence makes it increasingly accepted among researchers that a single SWB question is unlikely to be sufficiently comprehensive: it may not single-handedly capture all aspects of well-being that individuals care about (e.g., Stiglitz, Sen, and Fitoussi, 2009; for recent evidence from hypothetical and real choices, see Benjamin, Heffetz, Kimball, and Rees-Jones, 2012, 2013). Benjamin, Heffetz, Kimball, and Szembrot (2014) propose a methodology for identifying a set of SWB questions and their relative weights that may provide a better utility proxy, but much work still needs to be done, both on developing a reliable utility proxy for ordinal preferences and on estimating the effects of policy on that proxy.

The question that the present paper addresses is how to aggregate estimates of policy effects on different types into an overall policy recommendation. Since for most policy questions different individuals' interests do not fully coincide, aggregation across individuals has long been a central problem in economics. However, it has received relatively little attention in the SWB literature (for an excellent recent discussion of aggregation across individuals more generally, including in the SWB context, see Fleurbaey and Blanchet, 2013).

Typically, when a researcher estimates the effect of policy on SWB, she regresses a SWB measure on policy and other variables using a pooled sample of individuals. Since such a regression aims to estimate the effect of the policy on the average SWB survey response, the researcher effectively assumes that SWB responses are interpersonally comparable.[1] This assumption is problematic if survey respondents whose interests are affected differently by the policy use the survey-response scale in systematically different ways. Replacing the single-question SWB measures that are currently widely used with a comprehensive well-being measure that captures everything individuals care about would not in itself solve this problem. Neither would this problem be entirely solved by random assignment of policy: if different types use the response scale in systematically different ways, some types will unjustifiably impact the estimated coefficients more than other types. What is needed is:

a) to use a comprehensive well-being measure—a utility proxy representing ordinal preferences—as the dependent variable;
b) to run regressions that are identified from exogenous variation in policy; and, most relevantly for this paper,
c) to conduct the regressions separately within single-type groups of individuals who have the same policy interests.

But one then faces the aggregation problem: one needs a method for aggregating the estimated effects across types.

[1] The precise nature of the implicit assumptions depend on which regression specification is run. For example, if the SWB question has a 5-point response scale coded 1–5 and ordinary least squares is used, then the assumption is that the 1–5 numbers have the same meaning across individuals. If ordered logit is used instead, then the assumption is that the latent variable underlying the response has the same meaning across individuals. If fixed effects are also included, then changes (but not levels) of the variable are assumed to be interpersonally comparable. See Ferrer-i-Carbonell and Frijters (2004) for related discussion.

There are two main traditions in economics regarding interpersonal (or inter-type) aggregation that we believe could be adapted to the context of well-being data. Both treat utility as ordinal and interpersonally non-comparable. One tradition is to measure the effect of a policy in terms of the amount of money lost or received that would make an individual indifferent; these "money-metric" effects of the policy are then aggregated. Much has been written on the advantages and disadvantages of such money-metric approaches. Fleurbaey and Blanchet (2013) provide a recent, comprehensive discussion and make a strong case for further development of such approaches.

The other tradition, most familiar from voting contexts, is to give each individual's ordinal preference equal weight in an aggregation process that determines policy. In this paper we develop and explore such a method, which we call the Normalized Gradient Addition (NGA) mechanism. It aggregates preferences that could in principle be estimated from regressions similar to typical SWB regressions, but that satisfy the conditions (a), (b), and (c) above.

The basic idea is straightforward. We assume that a government agency has estimated the effects of small changes in a variety of policies on a comprehensive (but not interpersonally comparable) well-being measure—a utility proxy—for each type in the population. In terms of each type's implied preferences over policy vectors, the agency has thus estimated each type's gradient (i.e., the vector of local-policy-change effects on the utility proxy for that type). The NGA mechanism proceeds in two steps. First, each type's gradient vector is normalized to have the same length. The gradients in their original units are not ordinal objects because the magnitude of the effect of policy on the utility proxy depends on the scale used by that type when answering well-being surveys. Normalizing the gradients ensures that every individual in the population contributes an equal input to the mechanism. The second step is to calculate the weighted sum of these normalized gradient vectors, where a type's weight is its fraction of the population. The mechanism then prescribes that the government agency marginally adjusts the vector of policy levels proportionally to the vector resulting from the normalized gradient addition.

In sections I, II, and III, we provide three different motivations for the NGA mechanism. In section I, we formally describe the mechanism and motivate it as analogous to locally climbing a social welfare function (SWF). Maximizing a SWF would require measuring utility in interpersonally comparable units—for example, a money metric—and moving policy (immediately) to the levels that globally maximize the SWF. In contrast, the NGA mechanism has four properties that collectively explain how the formula describing it differs from the SWF-maximization formula. First, the NGA mechanism treats each individual symmetrically: any individual's normalized gradient vector carries the same weight in the mechanism as any other individual's (in contrast, SWFs may weight individuals differently depending on their utility levels). Second, as already noted, it treats the utility proxy as ordinal and interpersonally non-comparable across types. Third, it is a local mechanism: its inputs are the slopes of indifference surfaces local to the status quo policy vector, and its output is a marginal adjustment of policy. While most of the attention in the social choice literature has focused on mechanisms that prescribe a particular policy endpoint, as opposed to an adjustment from the status quo, we view the localness of the mechanism as attractive from the perspective of practical usefulness because the effects of policy can be most credibly estimated locally. Fourth, the NGA mechanism has the property of being "first-order strategy-proof (FOSP)," meaning that the mechanism is strategy-proof when each type's indifference surface over policies is replaced by a linear approximation. Thus if the agency estimates local policy preferences from survey data, answering the survey honestly is approximately a dominant strategy. While we are not aware of any evidence of strategic misreporting on well-being surveys, we argue that FOSP is nonetheless an attractive feature of the mechanism, and that it is a property with low marginal cost once a mechanism already has the first three properties.

In section II, we show that beyond being in the same tradition as preference aggregation via voting, the NGA mechanism is in fact equivalent to a particular voting procedure. Specifically, each individual allocates across the set of policies a fixed budget of votes in favor of increasing or decreasing each policy. The budget constraint is quadratic: the sum across policies of the *squared* votes cannot exceed the fixed budget. For each policy, the change prescribed by the mechanism is the sum of the votes across all individuals. To the extent that voting is viewed as an attractive way to aggregate preferences, the equivalence of the NGA mechanism to this voting procedure may be viewed as a justification for the mechanism. However, we emphasize that the mechanism can be implemented by a government agency that collects and analyzes well-being data—and thereby simulates how an individual *would* vote if well informed about the consequences of the policies—without any individual in the population actually having to vote.

In section III, we provide an axiomatic characterization of the NGA mechanism. In particular, we show what assumptions are needed, in addition to the four properties listed above, for the combined set of assumptions to be necessary and sufficient to pin down the NGA mechanism.

We envision that if the NGA mechanism were used in practice, it would be implemented iteratively, with the government agency regularly collecting well-being data (e.g., conducting well-being surveys), re-estimating policy effects at the current status quo, and implementing the next policy change. For this reason, in Section IV we study the properties of the iterated mechanism. Among other results, we prove that a stationary point always exists; provide some conditions for there to be a unique stationary point; and show that any stationary point is Pareto efficient. In addition to proving several analytic results, we conduct a wide range of simulations of the iterated mechanism to assess potential limitations. We find few multiple stationary points in our simulations and—in spite of trying hard to generate them—we find no limit cycles.

One of the biggest limitations of the NGA mechanism is that it is not invariant to the units in which policy is measured, and thus the choice of units is potentially subject to manipulation by the government. Indeed, in section IV we prove that in some cases, complete manipulation is possible: units could be chosen such that one of the types in the population achieves its bliss point, even when that type is a relatively small fraction of the population and has extreme preferences in all but one policy dimension. In some of our simulations, we assess the scope for manipulation by studying the sensitivity of the stationary point of the iterated mechanism to the choice of units. In our view, this limitation of the mechanism is not necessarily crippling, for two reasons. First, no alternative aggregation mechanism can fully escape sensitivity to some parameters that would be subject to potential political manipulation. For example, voting is always potentially manipulable via agenda-setting, and, similarly, the output of social welfare maximization may depend on the functional form of the social welfare function. Second, we believe that there are reasonable criteria for choosing the units, and (in section I) we mention some possible procedures for choosing units that satisfy these criteria.

In section V, we conclude by outlining how we envision that the NGA mechanism might someday be applied, as well as by discussing other applications of the mechanism, additional limitations, and possible extensions that could overcome those limitations. We also discuss the challenge of identifying types in the population, a challenge that can probably only be addressed imperfectly. Appendix A contains all proofs, and Appendix B contains some examples of results from our random simulations of the iterated mechanism.

The papers most closely related to ours are Benjamin, Heffetz, Kimball, and Szembot (BHKS, 2013) and Hylland and Zeckhauser (HZ, 1980)—both mentioned above—and a recent working paper by Chung and Duggan (2014). BHKS propose the NGA mechanism but focus exclusively on the one-shot mechanism. Section I of the present paper is closely based on BHKS (including reproducing BHKS's numerical example and figures). Many years earlier, HZ had essentially proposed the NGA mechanism (BHKS do not cite HZ because they were unaware of it). We say "essentially" because HZ focused on what we refer to as stationary points of the iterated mechanism. HZ described the NGA mechanism as a voting procedure, and section II of the present paper closely matches their set-up. Unlike HZ, in section II we view voting as a *motivation* of the mechanism that we envision would be implemented without the actual need for frequent voting—an *as-if voting* mechanism in which the government uses estimates of policy effects on well-being to construct individuals' would-be votes. Like HZ, Chung and Duggan (2014) view the mechanism as a voting procedure and focus on stationary points but (beyond the overlap with us and HZ) focus on analytic questions distinct from ours (particularly the relationship of stationary points to the core, generic local uniqueness, non-cooperative stability, and other stability issues). Our axiomatic characterization of the mechanism in section III as yet another motivation of the mechanism is entirely new, as is the simulation-based analysis of the iterated mechanism in section IV. Some of our results on properties of the stationary points in section IV previously appeared in HZ; we cite HZ when stating those results.

I. The Normalized Gradient Addition (NGA) Mechanism

In this section, we describe the NGA mechanism and motivate it by analogy with maximizing a social welfare function. The development of the mechanism in this section closely parallels BHKS.

I.A. Model Set-Up

We imagine that there is a government agency that has control over $P \geq 1$ policies. The level of each policy is a real number that can be adjusted independently of other policies. We denote the policy levels by the vector $\boldsymbol{p} \equiv (p_1, \ldots, p_P)$ and the status quo policy vector by $\boldsymbol{p}_0$.

The natural units of the *P* policies are in general not comparable (e.g., tax rate vs. limits on particulate concentrations in the air). For ease of exposition, we will postpone addressing this issue by assuming that each policy is measured in comparable "policy units"; we return in section I.D. below to discuss the important issue of how these policy units should be chosen. However, we note here that "1 policy unit" will be chosen to be a small enough change in policy that individuals' policy preferences are well approximated as locally linear over this range. As explained below, the mechanism will limit the policy change vector $\Delta\boldsymbol{p}$ to have a Euclidean length of 1 policy unit.

We partition the population of individuals into Θ > 1 types, where each individual of a given type θ has identical preferences over policies represented by utility function $u_\theta(\boldsymbol{p})$ (a reduced-form representation based on policy effects on the arguments of utility). Each $u_\theta(\cdot)$ is continuously differentiable, strictly concave, and achieves a maximum at a bliss point $\boldsymbol{p}_\theta^{\text{bliss}}$. In this section and the next, we assume that the status quo $\boldsymbol{p}_0$ is at a substantial distance from every type's bliss point; in section IV, we discuss how to deal with the more complicated case of being near or at the bliss point of one of the types. For each type θ, we denote its fraction of the population by ϕ_θ.

We denote the true gradient, or utility gradient, of type θ at $\boldsymbol{p}_0$ by $\boldsymbol{\nabla u}_\theta(\boldsymbol{p}_0) \equiv \left(\frac{\partial u_\theta(\boldsymbol{p}_0)}{\partial p_1}, \dots, \frac{\partial u_\theta(\boldsymbol{p}_0)}{\partial p_P}\right)$ or simply $\boldsymbol{\nabla u}_\theta$ when there is no ambiguity. The utility gradients (and the utilities themselves) are not observable. Instead, the government agency tracks an ordinal utility proxy $r_\theta(\boldsymbol{p})$ constructed from responses to a well-being survey, estimates the effects of each policy *p* on each type θ's utility proxy r_θ, and constructs the reported gradient at $\boldsymbol{p}_0$, $\boldsymbol{\nabla r}_\theta(\boldsymbol{p}_0) \equiv \left(\frac{\partial r_\theta(\boldsymbol{p}_0)}{\partial p_1}, \dots, \frac{\partial r_\theta(\boldsymbol{p}_0)}{\partial p_P}\right)$, denoted $\boldsymbol{\nabla r}_\theta$ when there is no ambiguity. While the utility proxy may be constructed from *any* well-being data, for concreteness we focus on the case of *reported* survey responses and, below, on the possibility of misreporting them; the possibility of misreporting can be generalized to any attempt by individuals to distort their well-being data (for example, attempts to distort non-survey-based indicators such as an individual's medical or economic data). We assume that each $r_\theta(\cdot)$ is continuously differentiable and has no inflection points (i.e., $\boldsymbol{\nabla r}_\theta(\boldsymbol{p}) = 0$ if and only if $\boldsymbol{p}$ is an extremum). The agency's aggregation mechanism is its method for choosing a policy adjustment $\Delta\boldsymbol{p}$ as a function of the observed $\boldsymbol{\nabla r}_{\boldsymbol{\theta}}$'s.

Since $r_\theta(\boldsymbol{p})$ is an ordinal utility proxy, the direction of the gradient $\boldsymbol{\nabla r_\theta}$ is invariant to monotonic transformations of $r_\theta(\cdot)$, but its length is arbitrary. In particular, if individuals of a given type tend to use the extreme ends of the well-being survey-response scale, then the policy effect estimates will tend to be larger in magnitude, and therefore $\boldsymbol{\nabla r_\theta}$ will tend to be a longer vector. In contrast, if individuals of a given type use little of the scale, then $\boldsymbol{\nabla r_\theta}$ will be a shorter vector. To eliminate these differences, we define the normalized gradient to be $\widetilde{\boldsymbol{\nabla r_\theta}} \equiv \frac{\boldsymbol{\nabla r_\theta}}{\|\boldsymbol{\nabla r_\theta}\|}$ if $\|\boldsymbol{\nabla r_\theta}\| > 0$ and $\widetilde{\boldsymbol{\nabla r_\theta}} \equiv \boldsymbol{0}$ if $\|\boldsymbol{\nabla r_\theta}\| = 0$. Thus, every type with a non-zero reported gradient has a normalized gradient of the same length (namely, 1 policy unit).

I.B. Motivating the NGA Mechanism

As noted in the Introduction, the Normalized Gradient Addition (NGA) mechanism that we explore in this paper has four key properties. First, it treats all individuals symmetrically. Second, it treats the survey-based utility proxy $r_\theta(\boldsymbol{p})$ as ordinal and non-comparable across individuals. Third, the mechanism is local: its input is the reported gradient at the status quo $\boldsymbol{\nabla r_\theta}$, and its output is an adjustment of policy $\Delta\boldsymbol{p}$. This localness contrasts with typical aggregation mechanisms in the social choice literature, which take as an input the full preference ordering of every type and generate as an output a policy level. A local mechanism has at least two advantages: local policy effects can be estimated empirically more credibly than global effects, and adjusting policy is often more practical than designing policy de novo (cf. Feldstein, 1976). Finally, the mechanism is "first-order strategy-proof" (FOSP), meaning that it makes answering the well-being survey honestly approximately a dominant strategy; we postpone further explanation and discussion of FOSP until section I.D below.

To motivate the form of the mechanism, consider the problem of choosing a local policy adjustment to maximize a social welfare function. Suppose that the utilities $u_1, \dots, u_\Theta$ are interpersonally comparable and directly observed by the social planner. Let $W(u_1, \dots, u_\Theta)$ be a social welfare function, where $W(\cdot)$ is strictly increasing, strictly concave, and continuously differentiable. We constrain the policy change to be at most 1 policy unit to ensure that it is local. The planner's problem is:

$$(1) \qquad \max_{\Delta\boldsymbol{p}} W\big(u_1(\boldsymbol{p}_0 + \Delta\boldsymbol{p}), \dots, u_\Theta(\boldsymbol{p}_0 + \Delta\boldsymbol{p})\big) \text{ subject to } (\Delta\boldsymbol{p})'(\Delta\boldsymbol{p}) \le 1.$$

Letting $\lambda > 0$ be the Lagrange multiplier, the vector first-order condition is $\sum_{\theta=1}^{\Theta} \frac{\partial W}{\partial u_\theta} \boldsymbol{\nabla u_\theta} = \lambda \boldsymbol{\Delta p}$. Since every type is far from its bliss point, the optimal policy change satisfies the constraint with equality. Hence the optimal policy change has length 1 policy unit and direction $\sum_{\theta=1}^{\Theta} \frac{\partial W}{\partial u_\theta} \boldsymbol{\nabla u_\theta}$:

$$(2) \qquad \Delta \boldsymbol{p} = \frac{\sum_{\theta=1}^{\Theta} \frac{\partial W}{\partial u_\theta} \boldsymbol{\nabla u_\theta}}{\left\| \sum_{\theta=1}^{\Theta} \frac{\partial W}{\partial u_\theta} \boldsymbol{\nabla u_\theta} \right\|},$$

where ||·|| is the Euclidean norm. Equation (2) is not feasible for the government agency to implement, however, because the agency does not observe the (true) utility gradients $\boldsymbol{\nabla u_\theta}$ and it cannot calculate the $\frac{\partial W(u_1(\boldsymbol{p}_0),\ldots,u_\Theta(\boldsymbol{p}_0))}{\partial u_\theta}$'s since it does not observe the (true) utility levels.

The Normalized Gradient Addition mechanism, which literally adds up the normalized *reported* utility gradients weighted by the population shares, is:

$$(3) \qquad \Delta \boldsymbol{p} = \sum_{\theta=1}^{\Theta} \phi_\theta \, \widetilde{\boldsymbol{\nabla r}}_\theta \, .$$

The NGA mechanism will not yield exactly the solution in equation (2) unless it happens to be the case that $\frac{\partial W}{\partial u_\theta} = \frac{\phi_\theta}{\|\boldsymbol{\nabla r_\theta}\|}$ for each θ. However, equation (3) can be viewed as an analog of equation (2) that is feasible, given the data available to the government agency. The differences are as follows.

First, the unobservable weights in equation (2), $\frac{\partial W}{\partial u_\theta}$, are replaced by ϕ_θ. Weighting each type by its population share satisfies the symmetry property of the mechanism. Note that all qualitative properties of the mechanism that we discuss in this paper would be the same if different weights were used, as long as they are exogenous to the reported gradients. In particular, an attractive variant of the NGA mechanism might use available data (such as individuals' wealth levels) that may be informative about the unobserved cardinal utility levels to weight worse-off individuals more highly and better-off individuals less highly than their population shares.

Second, the unobservable utility gradients in equation (2), $\boldsymbol{\nabla u}_{\theta}$, are replaced by the normalized (reported) gradients, $\widetilde{\boldsymbol{\nabla r}_{\theta}}$. Given that the length of the reported gradient $\boldsymbol{\nabla r}_{\theta}$ has no meaning, using the normalized gradients helps ensure that all individuals are treated symmetrically.

Third, the policy-change vector has length 1 policy unit in equation (2), while the NGA mechanism changes policy by less than 1 policy unit unless the $\widetilde{\boldsymbol{\nabla r}_{\theta}}$'s are equal across all θ, i.e., all types' reported gradients have the same direction. A more closely analogous variant of the mechanism could move policy in the same direction but require a policy change of length 1:

$$\Delta \boldsymbol{p} = \frac{\sum_{\theta=1}^{\Theta} \phi_{\theta}\, \widetilde{\boldsymbol{\nabla r}_{\theta}}}{\left\|\sum_{\theta=1}^{\Theta} \phi_{\theta}\, \widetilde{\boldsymbol{\nabla r}_{\theta}}\,\right\|}. \tag{4}$$

Equation (3), however—and not equation (4)—satisfies FOSP, and thus the inefficiency of not exhausting the social planner's budget constraint in equation (3) can be viewed as the social cost incurred by having FOSP; we defer further discussion until section I.D.

I.C. An Example and Some Intuition

To build intuition for the NGA mechanism in terms of vector addition, we begin with an example. There are two policy dimensions, the federal tax on distilled spirits and spending on national parks. A 1-policy-unit change in the tax is \$1.25 per proof gallon, and a 1-policy-unit change in federal funding for the National Park Service is \$125 million. There are two types in the population, Young and Old, with population shares $\phi_{\text{Old}} = 2/5$ and $\phi_{\text{Young}} = 3/5$. The government agency estimates that: a \$1 increase in the tax per proof gallon on distilled spirits increases the utility proxy of the Old by 3 units and decreases the utility proxy of the Young by 3 units; and a \$100-million increase in spending on national parks decreases the Old's utility proxy by 2 units and increases the Young's by 3 units. The reported gradients are therefore $\boldsymbol{\nabla r}_{\text{Old}} = \left(\frac{3 \text{ utils}}{\frac{\$1}{\$1.25}\text{units}}, \frac{-2 \text{ utils}}{\frac{\$100 \text{ mil.}}{\$125 \text{ mil.}}\text{units}}\right)$ and $\boldsymbol{\nabla r}_{\text{Young}} = \left(\frac{-3 \text{ utils}}{\frac{\$1}{\$1.25}\text{units}}, \frac{3 \text{ utils}}{\frac{\$100 \text{ mil.}}{\$125 \text{ mil.}}\text{units}}\right)$.

The first step in implementing the mechanism is to calculate the normalized gradients:

$$\widetilde{\nabla\boldsymbol{r}}_{\text{Old}} = \left(\frac{\frac{3\text{ utils}}{\frac{\$1}{\$1.25}\text{ units}}}{\sqrt{\left[\frac{3\text{ utils}}{\frac{\$1}{\$1.25}\text{ units}}\right]^2+\left[\frac{-2\text{ utils}}{\frac{\$100\text{ mil.}}{\$125\text{ mil.}}\text{ units}}\right]^2}}, \frac{\frac{-2\text{ utils}}{\frac{\$100\text{ mil.}}{\$125\text{ mil.}}\text{ units}}}{\sqrt{\left[\frac{3\text{ utils}}{\frac{\$1}{\$1.25}\text{ units}}\right]^2+\left[\frac{-2\text{ utils}}{\frac{\$100\text{ mil.}}{\$125\text{ mil.}}\text{ units}}\right]^2}} \right) = (0.83,-0.55)\ \text{ and}$$

$$\widetilde{\nabla\boldsymbol{r}}_{\text{Young}} = \left(\frac{\frac{-3\text{ utils}}{\frac{\$1}{\$1.25}\text{ units}}}{\sqrt{\left[\frac{-3\text{ utils}}{\frac{\$1}{\$1.25}\text{ units}}\right]^2+\left[\frac{3\text{ utils}}{\frac{\$100\text{ mil.}}{\$125\text{ mil.}}\text{ units}}\right]^2}}, \frac{\frac{3\text{ utils}}{\frac{\$100\text{ mil.}}{\$125\text{ mil.}}\text{ units}}}{\sqrt{\left[\frac{-3\text{ utils}}{\frac{\$1}{\$1.25}\text{ units}}\right]^2+\left[\frac{3\text{ utils}}{\frac{\$100\text{ mil.}}{\$125\text{ mil.}}\text{ units}}\right]^2}} \right) = (-0.71,0.71).$$

Figure 1 illustrates these normalized gradients, each of which is a vector of length 1 that points in the direction of maximal increase in the utility proxy for that type.

The other step of the mechanism is to make the policy change equal to the weighted sum of the normalized gradients, with weights equal to population shares: $\Delta\boldsymbol{p} = \frac{2}{5}(0.83,-0.55) + \frac{3}{5}(-0.71,0.71) = (-0.09,0.22)$. This vector addition, with the resultant policy change from $\boldsymbol{p}_0$ to $\boldsymbol{p}_1$, is illustrated in Figure 2. In natural units, the prescribed policy change is to reduce the tax by $0.09 \times \$1.25$ per proof gallon $= \$0.11$ per proof gallon and to increase spending on national parks by $0.22 \times \$125$ million $= \$25.3$ million.

This vector addition is particularly simple when there is a single policy dimension, $P = 1$. In that case, equation (3) specializes to $\Delta p_1 = \sum_{\theta=1}^{\Theta} \phi_\theta \text{ sign}\left(\frac{dr_\theta}{dp_1}\right)$. In words, the mechanism increases the level of the policy by an amount of policy units equal to the fraction of the population estimated to be better off with an increase in the policy minus the fraction estimated to be better off with a decrease, i.e., the "vote margin" in favor of an increase. Figure 3a illustrates an example with three types that have equal population weights.

The mechanism does not have access to cardinal "intensity of preference" information, but it *does* incorporate information on *relative* policy preferences when $P > 1$. Figure 3b illustrates the same three types as in Figure 3a, except that now there are $P = 2$ policy dimensions. As before, types 1 and 2 prefer an increase in policy 1, and type 3 prefers a decrease. As indicated by the directions of the normalized gradients, however, type 3 cares exclusively about the first policy dimension, while types 1 and 2 care much more about the second policy dimension. Consequently, as more policy dimensions are added, the mechanism can prescribe different changes—indeed, in the example in the figure, the first policy dimension is changed in the opposite direction when the second policy dimension is included. The more policy dimensions that are included, the more information on relative tradeoffs is available to the mechanism. Therefore, ideally, it would be used with as many policy dimensions included as possible.

An attractive feature of the mechanism is that if survey respondents truthfully report their well-being to the government agency, the NGA mechanism will find and implement any Pareto improvements that are available: for any policy-change vector $\boldsymbol{\delta}$ such that $\boldsymbol{\delta}'(\boldsymbol{\nabla u_{\theta}}) > 0$ for all θ, it follows that $\boldsymbol{\delta}'\left(\sum_{\theta=1}^{\Theta} \phi_{\theta} \frac{\boldsymbol{\nabla u_{\theta}}}{\|\boldsymbol{\nabla u_{\theta}}\|}\right) > 0$ and therefore $\boldsymbol{\delta}'(\Delta \boldsymbol{p}) > 0$. We return to this observation in section IV.B (where it appears as Proposition 2).

I.D. First-Order Strategy-Proofness

As mentioned above, one property of the NGA mechanism is being first-order strategy-proof (FOSP). To understand FOSP, consider the mechanism-design problem where each type reports her gradient $\boldsymbol{\nabla r_{\theta}}$ to the government agency, and the agency implements a policy change $\Delta \boldsymbol{p}$ that is a function of all the types' reported gradients. Since $\boldsymbol{\nabla r_{\theta}}$ is ordinal, "truthfully reporting" the utility gradient means that $\boldsymbol{\nabla r_{\theta}}$ is equal to $\boldsymbol{\nabla u_{\theta}}$ up to some arbitrary multiplicative scalar, i.e., $\boldsymbol{\nabla r_{\theta}}$ points in the same direction as $\boldsymbol{\nabla u_{\theta}}$. Formally, FOSP means that the mechanism is strategy proof—that is, it makes truthful reporting a dominant strategy—in the game where each type's indifference surfaces are replaced by a linear approximation at the status quo $\boldsymbol{p}_0$. Since any smooth indifference surface is locally linear, the gains from deviating from truthful reporting are second order. Therefore, as long as the policy change is small, truthful reporting is approximately a dominant strategy.

As applied here—where the government agency estimates $\boldsymbol{\nabla r_{\theta}}$ from survey data on well-being—FOSP means that each type has an incentive to make sure that the agency's estimate, $\boldsymbol{\nabla r_{\theta}}$, is equal to $\boldsymbol{\nabla u_{\theta}}$ up to an arbitrary multiplicative scalar. Each type can ensure this by answering the survey honestly. Hence FOSP makes answering the well-being survey honestly an approximately dominant strategy.

To see that the NGA mechanism is FOSP, note that the mechanism decentralizes the policy change: each type's contribution to the policy change is a vector of fixed length, and that type chooses the direction of the vector. Taking the other types' contributions as given, each type θ's most-preferred direction is its utility gradient evaluated at $\boldsymbol{p}_0 + \sum_{\hat{\theta}\neq\theta} \phi_{\hat{\theta}}\, \widetilde{\boldsymbol{\nabla r}_{\hat{\boldsymbol{\theta}}}}$ (i.e., taking into account the summed contributions of the other types). With linear preferences, type θ's utility gradient evaluated there is equal to its utility gradient evaluated at $\boldsymbol{p}_0$. To ensure that its contribution is $\boldsymbol{\nabla u}_{\boldsymbol{\theta}}(\boldsymbol{p}_0)$, type θ chooses to report some $\boldsymbol{\nabla r}_{\theta}(\boldsymbol{p}_0)$ that is proportional to $\boldsymbol{\nabla u}_{\boldsymbol{\theta}}(\boldsymbol{p}_0)$.

Relative to BHKS (Benjamin, Heffetz, Kimball and Szembrot, 2013), in this paper we de-emphasize FOSP for two reasons. First, we suspect that strategic misreporting on well-being surveys would be uncommon even if respondents knew that their responses would be used for policy purposes. Strategic misreporting would require incurring the cognitive costs of formulating a strategy and incurring the psychic cost of dishonest responding, and the gains to any one individual would be small.[2] Second, as we discuss in section IV, although one-shot application of the NGA mechanism is FOSP, iterated application of it—which we believe is of greater interest—is not.

Nonetheless, we still view FOSP as a desirable property for three reasons. First, it is conceivable that individuals *would* strategically misreport once they know that their survey responses will be used for policy purposes (a concern expressed by and Stutzer, 2012). If strategic misreporting occurs, the most likely strategies would be ones that are easy to understand and simple to implement. Other ways of using well-being data for policy are subject to such strategies. For example, much empirical work estimates the effect of a change in policy in monetary units by running a regression of the survey-based utility proxy on the policy and on income, and estimating the change in income required to hold constant the utility proxy under the policy change. If such estimates were used to guide policy, individuals could magnify their weight in policymaking by making their utility proxy less sensitive to changes in income. One way to do this might be to answer the well-being survey questions while focusing on non-income-related aspects of well-being. FOSP rules out such simple manipulation strategies.

Second, as we explain and formalize in section III below, FOSP has an attractive interpretation even if strategic misreporting is known not to be a problem. Specifically, if everyone responds truthfully, FOSP means that the mechanism satisfies a kind of monotonicity

[2] We also note that it would be difficult to disseminate information about how to exploit a possibility for manipulation to a large number of potential survey respondents without that effort becoming known and backfiring. By analogy, if some group encouraged people to respond to employment surveys to skew the unemployment numbers in order to get closer to its preferred monetary policy, the effort would almost surely be discovered quickly and become a political scandal.

property: if type θ's preference is changed, the policy change prescribed by the mechanism cannot change to one that is strictly worse for type θ, as judged by θ's new preference.

Finally, we view FOSP as a feature of the mechanism that has low marginal cost. As noted in section I.B., equation (4) is a mechanism satisfying properties (*i*)-(*iii*) but not FOSP. Figure 4 illustrates that the mechanism in equation (4)—which *guarantees* the policy change to have length 1 policy unit—would in general create an incentive to misreport. Having the additional property of FOSP requires having a smaller policy change—but since the direction of policy change is the same, we view this cost as relatively low. Indeed, in the continuous-time limit of an arbitrarily small policy change, equation (4) and equation (3) become equivalent. And in practice, if the amount of disagreement among types is expected to be large, the government agency could compensate by having 1 policy unit correspond to a larger change in all policies (as long as the size of 1 policy change is exogenous to the gradient reports).

I.E. Choosing the Policy Units

Until now, we have simply assumed that the policies are measured in comparable "policy units." In practice, however, for each policy j, a distance-metric parameter $m_j > 0$ that specifies the amount of change corresponding to "1 policy unit" will need to be chosen. While the qualitative properties of the mechanism do not depend on the specific values of the m_j's, they will matter for the output of the mechanism. An extreme example is shown in Figures 3a and 3b: as m_2 shrinks to zero, the mechanism's prescribed change in policy 1 switches direction.

Because the policy units matter, there may be incentive for agents who have political influence to manipulate the government agency's choice of m_j's. By eliminating degrees of freedom, standardized procedures for determining the m_j's can help minimize the potential for such manipulation. One possibility would be for the agency to conduct a survey to determine what relative m_j's correspond to intuitive judgments of similar-sized changes. Another idea—which we discuss more formally in section IV.C and explain how it would reduce the potential for manipulation—would be to choose relative m_j's such that the types' indifferences surfaces when plotted in policy-unit space are as close to spherical as possible. Doing so would require obtaining non-local information on preferences and hence may not be practical, but would be helpful even if done only imperfectly.

In determining the absolute magnitudes of the m_j's (parameterized in section I.F below as the "step size"), there is a tradeoff. On the one hand, FOSP depends on the mechanism's policy change being small enough that the indifference surfaces are well approximated as linear. On the other hand, the policy changes need to be large enough that it is possible to estimate their effects and that, when the mechanism is iterated, it has a satisfactory speed of progress.

I.F. The General Form of the NGA Mechanism

In describing the NGA mechanism by equation (3), we presupposed that the policies were measured in policy units and that the maximum distance that the mechanism could move is 1 policy unit. Here we will generalize both of those assumptions.

To translate from policies measured in natural units to policies measured in policy units, we define the matrix $\boldsymbol{M} \equiv \mathrm{diag}(m_1, m_2, \ldots, m_P)$. Thus, a policy change $\Delta\boldsymbol{p}$ in natural units is the change $\boldsymbol{M}^{-1}\Delta\boldsymbol{p}$ in policy units. Moreover, if the utility gradient in natural units is $\boldsymbol{\nabla u_\theta}$, then in policy units it is $\boldsymbol{M\nabla u_\theta}$. Similarly, the normalized gradient, redefined in terms of policy units, is $\widetilde{\boldsymbol{\nabla r}}_\theta \equiv \frac{\boldsymbol{M\nabla r}_\theta}{\|\boldsymbol{M\nabla r}_\theta\|}$ if $\|\boldsymbol{M\nabla r}_\theta\| > 0$ and $\widetilde{\boldsymbol{\nabla r}}_\theta \equiv \boldsymbol{0}$ if $\|\boldsymbol{M\nabla r}_\theta\| = 0$.

Given this generalized definition of the normalized gradient, and parameterizing the maximum change that the mechanism can prescribe in policy units by the "step size" $\varepsilon > 0$, the NGA mechanism is:

$$\boldsymbol{M}^{-1}\Delta\boldsymbol{p} = \varepsilon \sum_{\theta=1}^{\Theta} \phi_\theta \, \widetilde{\boldsymbol{\nabla r}}_{\boldsymbol{\theta}} \,. \tag{5}$$

Above and in most of the remainder of the paper, we assume that $\boldsymbol{M}$ equals the identity matrix, which amounts to assuming that policies are already measured in policy units; the exception is when we study the potential for manipulation of the $\boldsymbol{M}$ matrix, which we discuss in terms of a change in the coordinate system (equivalently, a transformation of $\boldsymbol{M}$). We will also generally assume that $\boldsymbol{\varepsilon = 1}$, except when we study the iterated mechanism in section IV, where it becomes useful to state results for $\boldsymbol{\varepsilon}$ sufficiently small.

II. The Voting Interpretation of the NGA Mechanism

In the previous section, we motivated the NGA mechanism as analogous to finding the marginal policy adjustment that maximizes a social welfare function, but which uses only ordinal information on preferences and weights each individual equally. In this section, we provide an alternative motivation: it is a voting mechanism in which each individual allocates a fixed budget of policy units across the policy dimensions.

Specifically, each type θ simultaneously chooses a vector of policy changes, $\widetilde{\boldsymbol{\nabla r}}_{\theta}$, expressed in policy units. We call $\widetilde{\boldsymbol{\nabla r}}_{\theta}$ type θ's *vote*. The vote is chosen subject to a quadratic budget constraint: $\left(\widetilde{\boldsymbol{\nabla r}}_{\theta}\right)'\left(\widetilde{\boldsymbol{\nabla r}}_{\theta}\right) \leq 1$. In words, the budget constraint says that the sum (across the P policy dimensions) of squared policy-unit changes can be at most 1. The policy change is then set equal to a weighted sum of the votes, where each type's weight is its share of the population: $\Delta\boldsymbol{p} = \sum_{\theta=1}^{\Theta} \phi_{\theta}\widetilde{\boldsymbol{\nabla r}}_{\theta}$. The solution concept is Nash equilibrium. We assume—as in the previous section—that the policy units are chosen so that $\Delta\boldsymbol{p}$ is small and that the initial policy vector $\boldsymbol{p}_0$ is not equal to anyone's bliss point.

This mechanism is easily seen to be equivalent to the formulation in the previous section. Geometrically, choosing one's vote subject to a quadratic budget constraint is the same as choosing a point on or within the unit circle (see Figure 1). Given everyone else's vote, each type θ's optimal vote is to choose the point on or within the unit circle that moves as far as possible in the direction of θ's utility gradient, starting from $\boldsymbol{p}_0 + \sum_{\widehat{\theta}\neq\theta} \phi_{\widehat{\theta}}\, \widetilde{\boldsymbol{\nabla r}}_{\widehat{\theta}}$. Because the policy change is small, type θ's preferences are approximately linear, and therefore his gradient there is approximately equal to his gradient at $\boldsymbol{p}_0$. Because he wants to move as far as possible, his budget constraint binds. Therefore, each type θ's optimal vote is $\widetilde{\boldsymbol{\nabla r}}_{\theta} = \frac{\boldsymbol{\nabla u}_{\theta}}{\|\boldsymbol{\nabla u}_{\theta}\|}$. An alternative way to solve for the Nash equilibrium is algebraically. Each type θ solves:

$$\max_{\widetilde{\boldsymbol{\nabla r}}_{\theta}} u_{\theta}\left(\boldsymbol{p}_0 + \sum_{\widehat{\theta}\neq\theta} \phi_{\widehat{\theta}}\widetilde{\boldsymbol{\nabla r}}_{\widehat{\theta}} + \phi_{\theta}\widetilde{\boldsymbol{\nabla r}}_{\theta}\right) \text{ subject to } \left(\widetilde{\boldsymbol{\nabla r}}_{\theta}\right)'\left(\widetilde{\boldsymbol{\nabla r}}_{\theta}\right) \leq 1.$$

The vector first-order condition is $\phi_{\theta}\boldsymbol{\nabla u}_{\theta} = \lambda\widetilde{\boldsymbol{\nabla r}}_{\theta}$, where $\lambda > 0$ is the Lagrange multiplier, and thus $\widetilde{\boldsymbol{\nabla r}}_{\theta} \equiv \frac{\boldsymbol{\nabla u}_{\theta}}{\|\boldsymbol{\nabla u}_{\theta}\|}$ since the budget constraint binds.

The equivalence of the NGA mechanism to a voting procedure may be seen as providing a justification for the mechanism to the extent that voting is viewed as an attractive way to aggregate preferences. It also implies that the NGA mechanism could be applied by literally implementing the voting procedure.

In our view, however, there are three potentially significant advantages to the technocratic implementation of the mechanism that we emphasize in this paper: a government agency estimates the effects of policy on a utility proxy for each type, and then calculates the policy change prescribed by the mechanism.[3] First, actual voting incurs transactions costs, and these costs would be substantial if the government literally held a referendum on many policy issues on a regular basis.

Second, in order for voters to vote in accordance with their true interests, they need to be well informed and unbiased in their views regarding how the policies will affect them. The technocratic implementation sidesteps this requirement, because the government agency uses its estimates of the actual effects of policy to take into account how each type *would* vote if well informed. Indeed, the mechanism may be especially well suited for application to policies about which citizens have little expertise (such as pollution levels).

Third, in practice voting procedures often fail to take into account everyone's interests due to low turnout. If survey response rate is higher than voter turnout, and if the government agency can adjust statistically for non-response on its well-being survey—admittedly, big "ifs"—the technocratic implementation can reduce this problem.

III. Axiomatic Characterization of the NGA Mechanism

In section I, we summarized the NGA Mechanism and showed that it satisfies four properties: it is (*i*) symmetric, (*ii*) ordinal, (*iii*) local, and (*iv*) first-order strategy-proof (FOSP). In this section, we provide an axiomatic characterization of the mechanism. Doing so clarifies what assumptions, in addition to properties (*i*)–(*iv*), are implicit in the NGA mechanism. It also helps to place the mechanism in the context of the social choice literature, since many of the axioms are local versions of otherwise-standard axioms.

[3] To be clear, we are not suggesting replacing voting by a technocratic procedure—an idea that may sound alarmingly Orwellian—but rather, for certain policies, replacing an existing technocratic procedure (e.g., cost-benefit analysis) with a different technocratic procedure, the NGA mechanism. The discussion below concerns the advantages of implementing the NGA mechanism technocratically rather than implementing it by holding referenda on the policy adjustments of the kind outlined in this section.

We focus on mechanisms that take as an input the vector of individuals' preference gradients and output a change in the policy vector. In doing so, we restrict attention to mechanisms that satisfy (*iii*), the localness property. Accordingly, for any $\Theta \in \{0,1,2,\ldots\}$, we define a Θ-*agent mechanism* to be a function $f\colon (\mathbb{R}^P)^\Theta \times \mathbb{R}^\Theta_{++} \to \mathbb{R}^P$ that assigns, for every profile $\boldsymbol{\nabla r} = (\boldsymbol{\nabla r}_1, \ldots, \boldsymbol{\nabla r}_\Theta)$ of reported preference gradients and every vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_\Theta)$ of strictly positive weights, a vector $f(\boldsymbol{\nabla r}; \boldsymbol{\phi})$ that will represent the incremental change in policy space. (In the case $\Theta = 0$, a 0-agent mechanism is simply a point in $\mathbb{R}^P$.) The $\boldsymbol{\nabla r}_\theta$'s could be directly reported by agents (as in the interpretation of a voting procedure from section II) or could be estimated by the government agency from individuals' survey responses (or other well-being data, as discussed in section I.A. above). The true, unobserved preference gradient profile is denoted $\boldsymbol{\nabla u} = (\boldsymbol{\nabla u}_1, \ldots, \boldsymbol{\nabla u}_\Theta)$. Unlike in section I, here the ϕ_θ's are exogenous weights (not necessarily population fractions), and they do not need to sum to 1.

A *mechanism family* consists of a Θ-agent mechanism for each $\Theta \in \{0,1,2,\ldots\}$. We denote mechanisms with different numbers of agents by the same symbol *f*; there should be no ambiguity. The characterization theorems at the end of this section will show that a mechanism family that satisfies the axioms described below must be the NGA mechanism.

Property (*i*)—symmetry—is formalized by an anonymity axiom.

Anonymity: For any Θ, $\boldsymbol{\nabla r} \in (\mathbb{R}^P)^\Theta$, $\boldsymbol{\phi} \in \mathbb{R}^\Theta_{++}$, and permutation π of the set $\{1,2,\ldots,\Theta\}$, we have $f(\boldsymbol{\nabla r}_1, \ldots, \boldsymbol{\nabla r}_\Theta; \phi_1, \ldots, \phi_\Theta) = f(\boldsymbol{\nabla r}_{\pi(1)}, \ldots, \boldsymbol{\nabla r}_{\pi(\Theta)}; \phi_{\pi(1)}, \ldots, \phi_{\pi(\Theta)})$.

Similar anonymity axioms are common in the social choice literature (except that they usually apply to agents' entire preference profiles rather than just their gradients).

To ensure property (*ii*), we require that the reported gradients that the mechanism takes as an input are ordinal.

Ordinality: For any Θ, $\boldsymbol{\nabla r}$, $\boldsymbol{\phi}$, and vector of strictly positive constants $(k_1, \ldots, k_\Theta)$, we have $f(\boldsymbol{\nabla r}_1, \ldots, \boldsymbol{\nabla r}_\Theta; \phi_1, \ldots, \phi_\Theta) = f(k_1\boldsymbol{\nabla r}_1, \ldots, k_\Theta\boldsymbol{\nabla r}_\Theta; \phi_1, \ldots, \phi_\Theta)$.

The axiom states that the mechanism does not use any information on the length of an agent's reported gradient—and thus only uses information on its direction. (In the social choice literature, an ordinality assumption is generally implicit because the axioms are written in terms of preference profiles, rather than utility profiles.)

To formalize property (*iv*), we use a strategy-proofness axiom that is standard except that it is applied locally, only to the agents' reports of their gradients (which are reports of preferences only up to a first order approximation at the status-quo policy vector $\boldsymbol{p}_0$). Since the mechanism is ordinal, a *false report* for agent θ is defined as a vector $\boldsymbol{\nabla r}'_\theta \in \mathbb{R}^P$ such that for any positive constant k, $\boldsymbol{\nabla r}'_\theta \neq k\boldsymbol{\nabla u}_\theta$.

First-order strategy-proofness (FOSP): For any Θ, $\boldsymbol{\nabla r}$, $\boldsymbol{\phi}$, agent θ, $\boldsymbol{\nabla u}_\theta$, and false report $\boldsymbol{\nabla r}'_\theta$, we have $\boldsymbol{\nabla u}_\theta \cdot f(\boldsymbol{\nabla u}_\theta, \boldsymbol{\nabla r}_{-\theta}; \boldsymbol{\phi}) \geq \boldsymbol{\nabla u}_\theta \cdot f(\boldsymbol{\nabla r}'_\theta, \boldsymbol{\nabla r}_{-\theta}; \boldsymbol{\phi})$.

FOSP says that an agent cannot get a better outcome by lying about her preference gradient than by telling the truth.

Policymakers who anticipate that survey respondents will simply respond truthfully to well-being surveys may not find the incentive-compatibility property of the mechanism to be a compelling motivation in its favor. It is therefore important to note that the FOSP axiom is a normatively desirable property even if agents always truthfully report their preference gradients. Under that assumption, the axiom can be rewritten as stating: For any Θ, $\boldsymbol{\nabla u}$, $\boldsymbol{\phi}$, agent θ, and alternative gradient $\boldsymbol{\nabla u}'_\theta$, we have $\boldsymbol{\nabla u}_\theta \cdot f(\boldsymbol{\nabla u}_\theta, \boldsymbol{\nabla u}_{-\theta}; \boldsymbol{\phi}) \geq \boldsymbol{\nabla u}_\theta \cdot f(\boldsymbol{\nabla u}'_\theta, \boldsymbol{\nabla u}_{-\theta}; \boldsymbol{\phi})$. This property is a kind of "monotonicity" feature of the mechanism: if agent θ's preference is changed, the policy change prescribed by the mechanism cannot change to one that is strictly worse for agent θ, as judged by θ's new preference. This formulation is formally similar to monotonicity properties that have appeared elsewhere in the social choice literature, such as Maskin monotonicity (1999) or strong positive association (Muller and Satterthwaite, 1977).[4]

Beyond properties (*i*)–(*iv*) as formalized above, additional assumptions are needed to have a complete set of necessary and sufficient conditions for the NGA mechanism. A key assumption among them is:

[4] Those properties are in fact equivalent to strategy-proofness in a voting setting with unrestricted preferences over a finite set of alternatives, though they are not equivalent in our setting.

Directional Citizen Sovereignty: For any Θ, $\boldsymbol{\nabla r}$, and $\boldsymbol{\phi}$, $f(\boldsymbol{\nabla r};\boldsymbol{\phi})$ lies in the linear span of $\boldsymbol{\nabla r}_1, \dots, \boldsymbol{\nabla r}_\Theta$. (If $\Theta = 0$, then this axiom means the value of f is zero.)

This axiom states that the policy-change vector prescribed by the mechanism must be a linear combination of the agents' reported gradients (though possibly longer or shorter than the reported vectors). It rules out moving in any direction that has a component that is orthogonal to every agent's reported gradient. In this sense, it requires the mechanism to respect the agents' reported preferences. It can be understood as an axiom that treats the agents' reports $\boldsymbol{\nabla r}_\theta$ as "votes" that the mechanism respects. We call the axiom "directional citizen sovereignty" because it is in the spirit of Arrow's "citizen sovereignty" axiom—which requires that every possible social outcome be achievable by some configuration of individuals' preferences—but is formulated for the context of a mechanism that is local.

The next two axioms deal with how mechanisms within a given family that have different numbers of agents relate to each other.

Merging: For any $\Theta \geq 2$, $\boldsymbol{\nabla r}$, and $\boldsymbol{\phi}$, and two distinct agents θ, θ', if $\boldsymbol{\nabla r}_\theta = \boldsymbol{\nabla r}_{\theta'}$, then

$$f(\boldsymbol{\nabla r}_\theta, \boldsymbol{\nabla r}_{\theta'}, \boldsymbol{\nabla r}_{-(\theta,\theta')}; \phi_\theta, \phi_{\theta'}, \phi_{-(\theta,\theta')}) = f(\boldsymbol{\nabla r}_\theta, \boldsymbol{\nabla r}_{-(\theta,\theta')}; \phi_\theta + \phi_{\theta'}, \phi_{-(\theta,\theta')}).$$

(Note that we have written this axiom in a way that presupposes the anonymity axiom; it could be written more generally at the cost of cumbersome notation.) The merging axiom says that two agents with weights ϕ_θ and $\phi_{\theta'}$ that report the same gradient can be treated as a single agent with weight $\phi_\theta + \phi_{\theta'}$. This axiom underlies the assumption, maintained throughout, that individuals with the same preferences can be treated jointly as a "type" in the population. The merging axiom is also crucial in giving meaning to the weights ϕ_θ; indeed, none of the other axioms makes any substantive reference to their magnitudes.

The cancellation axiom states that if two equal-sized groups have exactly opposite preferences, then they cancel each other out.

Cancellation: For any $\Theta \geq 2$, $\boldsymbol{\nabla r}$, $\boldsymbol{\phi}$, and two distinct agents θ, θ', if $\boldsymbol{\nabla r}_{\theta'} = -\boldsymbol{\nabla r}_\theta$ and $\phi_\theta = \phi_{\theta'}$, then $f(\boldsymbol{\nabla r}_\theta, \boldsymbol{\nabla r}_{\theta'}, \boldsymbol{\nabla r}_{-(\theta,\theta')}; \phi_\theta, \phi_{\theta'}, \phi_{-(\theta,\theta')}) = f(\boldsymbol{\nabla r}_{-(\theta,\theta')}; \phi_{-(\theta,\theta')})$.

This axiom captures a property of voting mechanisms. Moreover, it seems natural: when two equal-sized groups have opposing preferences, it would be strange if the net effect favored a direction preferred by one of the groups or a direction preferred by neither group.

The final axiom, which can be understood as a technical assumption, ensures that the mechanism is a smooth function of the agents' gradients, as long as no agent is at a bliss point:

Smoothness: For any $\Theta \geq 2$ and $\boldsymbol{\phi}$, $f(\boldsymbol{\nabla r}; \boldsymbol{\phi})$ is twice-continuously differentiable as a function of $\boldsymbol{\nabla r} \in (\mathbb{R}^P \setminus \{\mathbf{0}\})^\Theta$.

Note that it would not make sense to require the mechanism to be smooth at an agent's bliss point: since the mechanism is ordinal, the length of any agent θ's reported vector $\boldsymbol{\nabla r_\theta}$ has no meaning, so agent θ's input to the mechanism is discontinuous as $\boldsymbol{\nabla r_\theta}$ crosses from positive in a given direction to negative in that direction through the zero vector.

Our characterization theorem states that this set of axioms is necessary and sufficient for pinning down the NGA mechanism uniquely up to an arbitrary multiplicative constant:

Theorem 1. A mechanism family satisfies anonymity, ordinality, first-order strategy-proofness, directional citizen sovereignty, merging, cancellation, and smoothness if and only if $f(\boldsymbol{\nabla r}; \boldsymbol{\phi}) = \varepsilon \sum_{\theta=1}^{\Theta} \phi_\theta \frac{\boldsymbol{\nabla r_\theta}}{\|\boldsymbol{\nabla r_\theta}\|}$ for some positive constant $\varepsilon > 0$.

The axioms determine the direction of movement but not the mechanism's "step size" ε—i.e., how far the mechanism moves policy if all agents report the same gradient vector—which in previous sections we normalized to be 1 policy unit.

All proofs are relegated to Appendix A, but here we provide some intuition for the role of each axiom. It is clear from the discussion in section I that anonymity, ordinality, and strategy-proofness are necessary conditions for the mechanism. Merging is crucial in giving meaning to the weights ϕ_θ.

Cancellation plays two roles in the analysis. First, in conjunction with merging, it pins down what happens when an agent has weight zero. In particular, it allows us to remove any agent with zero weight (meaning that an agent at a bliss point contributes the zero vector to the policy change). Second, cancellation is the only axiom that connects behavior of the mechanism across different weight profiles $\boldsymbol{\phi}$ for which the value of $\sum_{\theta=1}^{\Theta}\phi_\theta$ differs; it ensures that the scale parameter is the same regardless of how many agents cancel each other out.

Smoothness plays an important role in ensuring that the mechanism is additively separable across agents; that part of the proof is a mild extension of work by Barberá, Bogomolnaia, and van der Stel (1998), who also give an example to show in their context that additive separability can fail if smoothness is not imposed.

As noted above, the key additional axiom is directional citizen sovereignty. It is needed to identify the "shape" from which agents can pick their most-preferred vector (that the mechanism adds up across agents). In the NGA mechanism, each agent picks her most-preferred vector from the unit sphere. To generalize from the unit sphere, we could let $S \subseteq \mathbb{R}^P$ be any smoothly convex set containing the origin, and define a mechanism analogous to NGA: for any non-zero $\boldsymbol{\nabla u}_\theta$, let $\boldsymbol{\nabla r}_\theta(\boldsymbol{\nabla u}_\theta)$ be the element of S that maximizes $\boldsymbol{\nabla u}_\theta \cdot \boldsymbol{\nabla r}_\theta$ (and define $\boldsymbol{\nabla r}_\theta(\boldsymbol{\nabla u}_\theta) = 0$ if $\boldsymbol{\nabla u}_\theta = 0$), and then define $f(\boldsymbol{\nabla r}; \boldsymbol{\phi}) = \sum_{\theta=1}^{\Theta}\phi_\theta\, \boldsymbol{\nabla r}_\theta$. This mechanism in general would obey all the axioms, except for directional citizen sovereignty. Figure 5 illustrates an example in the simple case where there is a single agent ($\Theta = 1$) and S is an ovoid. In order for the mechanism to satisfy first-order strategy-proofness, it must be the case that when the agent's reported gradient $\boldsymbol{\nabla r}_\theta$ is equal to his true gradient $\boldsymbol{\nabla u}_\theta$, the mechanism moves in the agent's most-preferred direction. As shown in the figure, however, the agent's most-preferred direction is not the same direction as his gradient: by tilting the direction of the vector contributed to the mechanism toward the "major axis" of the ovoid, the mechanism moves in only a slightly worse direction for that individual, but moves further. But when the agent truthfully reports her gradient, moving in the agent's most-preferred direction violates directional citizen sovereignty, which for a single agent requires contributing a scalar multiple of the reported gradient $\boldsymbol{\nabla r}_\theta$ to the mechanism.

One natural and interesting set of variants of the NGA mechanism that satisfies all the axioms except directional citizen sovereignty are motivated by choosing a different norm for measuring the amount of policy change. Recall from section II that in the voting interpretation of the mechanism, each type θ is conceptualized as choosing its contribution to the policy change $\widetilde{\boldsymbol{\nabla r}}_\theta$ subject to the budget constraint $\left(\widetilde{\boldsymbol{\nabla r}}_\theta\right)'\left(\widetilde{\boldsymbol{\nabla r}}_\theta\right) \leq 1$. This budget constraint can be understood as a constraint on the distance of movement, $\left\|\widetilde{\boldsymbol{\nabla r}}_\theta\right\| \leq 1$, where $\|\cdot\|$ is the L^2 norm (i.e., Euclidean norm) defined by $\|\mathbf{x}\| = \left(\sum_{j=1}^{P} x_j^2\right)^{\frac{1}{2}}$. Alternatively, the budget-constraint distance could be measured by the L^q norm, defined by $\|\mathbf{x}\| = \left(\sum_{j=1}^{P} |x_j|^q\right)^{\frac{1}{q}}$, for any $1 < q < \infty$ (at 1 and ∞, the mechanism would violate the smoothness axiom). As $q \to \infty$, the choice set for $\widetilde{\boldsymbol{\nabla r}}_\theta$ becomes a hypercube. Thus, each type can "vote" on every policy dimension independently, without having to trade off between them. This mechanism would therefore not capture any information on relative intensity of preference across policy dimensions. As $q \to 1$, the choice set for $\widetilde{\boldsymbol{\nabla r}}_\theta$ becomes an orthoplex (the higher-dimensional analog of a diamond in two dimensions, or an octahedron in three). In this case, each type's optimum is a corner solution, effectively voting on only one policy dimension. This mechanism would capture some relative intensity information—specifically, which policy dimension matters most—but for each type would incorporate information on preferences for that one dimension. The value of $1 < q < \infty$ corresponds to a specific tradeoff between how much the mechanism accounts for the relative importance of the dimensions and for the amount of preference information in each dimension.

An argument in favor of the directional citizen sovereignty axiom (the case $q = 2$) is that it makes the mechanism most transparent: a type's contribution to the policy change is in whatever direction most improves that type's well being. For other variants of the mechanism, the government agency would estimate each type's reported gradient and then calculate that type's preferred contribution to the mechanism, which would point in a different direction than the reported gradient. Transparency is often viewed as a desirable property for policy institutions (e.g., Rawls, 1971).

IV. Dynamic Performance of the Iterated NGA Mechanism

The previous sections described and characterized the NGA mechanism for a single marginal policy-adjustment iteration. It is important, however, to understand the dynamic properties of the mechanism because in practice the mechanism would be likely to be applied iteratively: the government agency would estimate the effects of policies on well being, adjust the policies accordingly, and then repeat the process. It is important to know, for example, if the mechanism generates a limit cycle, i.e., at the end of a sequence of iterations one ends up at the starting point, and thus a government using the mechanism is doomed to infinite transition between the same policies. Since the potential for such circularity is well known in voting systems, and since the NGA mechanism can be interpreted as a voting procedure, cycles are a possibility that must be taken seriously.[5]

Much of the social choice literature is concerned with social preferences, which specify a policymaker's complete ranking of alternatives as a function of the preference rankings of each individual. The NGA mechanism does not directly relate to this literature because it is an algorithm that generates an outcome but not a complete ranking of alternatives (except in special cases, such as the one analyzed in Proposition 4, in which the iterated mechanism can be interpreted as maximizing a particular function). Interpreted as a social ordering, the iterated NGA mechanism is only a partial ordering because it only ranks policy vectors along the path that the iterated mechanism takes.

Although the iterated mechanism does not generate a complete social preference, under conditions in which it terminates at a stationary point, it *does* generate a social choice rule: a mapping from the profile of individuals' preference rankings and other parameters to a set of specific policies. Specifically, the iterated mechanism picks out a stationary point, or set of stationary points, as a function of the status quo and the profile of individuals' global preferences.

Perhaps the most well-known result about social choice rules is the Gibbard-Sattherthwaite theorem, which states that any social choice rule satisfying a mild set of conditions cannot be strategy-proof. The iterated NGA mechanism is indeed not strategy-proof, even though it is first-order strategy-proof at each iteration. Our view is that first-order strategy-proofness eliminates many straightforward ways to manipulate the mechanism (see section I.D), and the subtle calculations required to try to manipulate the dynamic path are less likely to be seriously exploited.

Below, we characterize the iterated mechanism analytically, and we further study it by conducting simulations. If it leads to a terminal stationary point, as the mechanism always does in all our simulations, that stationary point is a Pareto optimum. (We show analytically that such a stationary point always exists, but not that the mechanism will always get there.) Furthermore, in our simulations there is usually only one stationary point, and the simulations give insights into what factors make multiple stationary points more likely. We prove that any stationary point

[5] There is also the logical possibility of the mechanism wandering endlessly—something that, like limit cycles, we did not find in any of our simulations, when all the preferences had bliss points within the policy space.

is always within a well-defined sphere surrounding the bliss points of any critical mass of types. Finally, we show analytically the need to worry about what effect the choice of coordinate system for the policy space has on the stationary point, and study this issue through simulations.

IV.A. Defining the Iterated NGA Mechanism

In defining the iterated mechanism, we re-introduce the step-size parameter $\epsilon > 0$ from section I.F, which we assume to be very small in order to approximate a differential equation.[6] The iterated mechanism is defined as a sequence of policies, $\boldsymbol{p}_t$ for $t = 0,1,2, \ldots$, such that

$$\Delta \boldsymbol{p}_t = \epsilon \sum_{\theta=1}^{\Theta} \phi_\theta \, \widetilde{\boldsymbol{\nabla r}}_\theta(\boldsymbol{p}_t).$$

We also adopt the Nash-equilibrium assumption from section II because we would like to allow for the possibility that the current policy is in the neighborhood of the bliss point of one of the types. In the context of the one-shot mechanism, we assumed away this possibility as non-generic. In some of our simulations discussed below, however, the path of the iterated mechanism enters the neighborhood of a bliss point, and in some simulations and some analytic results, the iterated mechanism converges toward a bliss point.

One issue of concern is that type θ's optimal reported gradient direction is especially sensitive to everyone else's reported gradients once current policy $\boldsymbol{p}_{t-1}$ is sufficiently close to her bliss point. More precisely, when the distance is less than ϵ, *any* direction can be optimal, depending on everyone else's report. Figure 6 illustrates that type θ's preferred direction starting from the period-$(t-1)$ policy vector, $\boldsymbol{p}_{t-1}$, is not the same as that type's preferred direction after everyone else's report has been contributed, $\boldsymbol{p}_{t-1} + \epsilon \sum_{\hat{\theta} \neq \theta} \phi_{\hat{\theta}} \, \widetilde{\boldsymbol{\nabla r}}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{p}_t)$. Another issue arises when everyone else's report puts the resultant policy within $\phi_\theta \epsilon$ of type θ's bliss point. In that case, type θ's preferred vector would bring the mechanism exactly to type θ's bliss point—and the preferred length of the type's vector would be the distance to that bliss point, which may be shorter than the length $\phi_\theta \epsilon$ prescribed by the mechanism as it was described in sections I and III.

[6] While we continue to work in discrete time, a natural alternative approach would be to define the iterated mechanism by the differential equation $\dot{\boldsymbol{p}}_t = \sum_{\theta=1}^{\Theta} \phi_\theta \, \widetilde{\boldsymbol{\nabla u}}_{\boldsymbol{\theta}}(\boldsymbol{p}_t)$ and consider its stationary points. We are currently exploring this approach.

Therefore, as in the voting interpretation of the mechanism from section II, we allow each type θ to choose not only its direction but also the length of its vector within the interval $[0, 1]$ and assume that the directions and lengths are chosen in accordance with Nash equilibrium. In addition, we assume that the agents are assumed to be myopic in the sense that at each iteration they play a Nash equilibrium as if the mechanism were one-shot, not taking into account their dynamic incentives. We also assume that no two types with different preferences share a common bliss point.[7]

This set-up has three attractive theoretical features. First, it specializes to the mechanism as described in previous sections in the case where the current policy vector $\boldsymbol{p}_{t-1}$ is not in the neighborhood of any bliss point: in that case, each type's (approximate) dominant strategy is to choose its preferred direction from $\boldsymbol{p}_{t-1}$ and its maximal length of 1. Second, in the case where the current policy vector is in the neighborhood of type θ's bliss point, at the "myopic Nash equilibrium," all other types continue to play their dominant strategy (i.e., truthful reporting) as in the previous sections, and we only need to adjust the analysis for type θ, who chooses her optimal contribution to the mechanism starting from $\boldsymbol{p}_{t-1} + \epsilon \sum_{\hat{\theta} \neq \theta} \phi_{\hat{\theta}} \, \widetilde{\boldsymbol{\nabla r}_{\hat{\boldsymbol{\theta}}}}(\boldsymbol{p}_t)$. Third and relatedly, as $\boldsymbol{p}_{t-1}$ moves from one side of type θ's bliss point to another, this modification avoids a discontinuity in the mechanism's prescribed policy change $\Delta \boldsymbol{p}_t$ that would occur if the length of type θ's contribution were held fixed. With our set-up, there is no discontinuity because the length of type θ's preferred vector converges to 0 as the result from everyone else's choices on that iteration, $\boldsymbol{p}_{t-1} + \epsilon \sum_{\hat{\theta} \neq \theta} \phi_{\hat{\theta}} \, \widetilde{\boldsymbol{\nabla r}_{\hat{\boldsymbol{\theta}}}}(\boldsymbol{p}_t)$, gets closer to his bliss point.

As a practical matter for a government agency implementing the iterated mechanism, dealing with a type near its bliss point could generate potential complications. To maintain incentive-compatibility on the well-being survey at each iteration, the agency would need to implement the myopic Nash equilibrium. The agency may be able to detect when a type is near its bliss point by observing that the type's preferred policy direction is becoming especially variable from one iteration to the next. In order to execute the equilibrium policy-change contribution of that type, the agency would need to estimate more than first-order information from the well-being data, including for example curvature of the indifference curves. However,

[7] Myopic Nash equilibrium would still be defined if the current policy were the bliss point of more than one type, but the equilibrium might not be unique. Since multiple types sharing a bliss point is non-generic, we assume away the possibility to avoid the complexity of dealing with equilibrium selection.

we note that if the type near its bliss point has any concave, cardinal utility function underlying its ordinal preferences, then dealing correctly with near-bliss-point situations is relatively unimportant, regardless of what smooth social welfare function is used to aggregate it with the utility functions of others. Near a type's bliss point, small changes in policy have only a second-order effect on the agent's utility, which implies that the effect on the agent of incorrectly accounting for these preferences is second order and that the agent has little incentive to respond untruthfully to the well-being survey.

Returning to the theory, a stationary point of the iterated mechanism is defined as a policy vector $\boldsymbol{p}^*$ such that when $\boldsymbol{p}_{t-1} = \boldsymbol{p}^*$,

$$\Delta \boldsymbol{p}_t = \mathbf{0}.$$

We discuss the possibilities of non-existence of a stationary point (e.g., a limit cycle) and non-uniqueness below in section IV.C.1.

IV.B. Analytic Results

In this section we provide analytic results. Some of these results describe the one-shot behavior of the NGA mechanism (which apply to each step along the policy trajectory), and some describe what this behavior implies for the properties of the iterated mechanism's stationary points. Hylland and Zeckhauser (1980) previously proved several of the results about stationary points; we credit them with "HZ" in the statement of these results.

The first result states that a stationary point exists.

Proposition 1 (HZ). There exists a stationary point of the iterated mechanism.

The proposition follows from application of a standard fixed-point theorem.

Proposition 1 does not rule out the existence of multiple stationary points or the possibility that some initial policy vectors may lead to limit cycles instead of converging to the stationary point. We have not found any *general* conditions that rule out these possibilities, but as we discuss in section IV.C below, in our simulations we find no instances of limit cycles and find multiple stationary points only rarely.

The next result describes an attractive feature of the NGA mechanism at each step: if some direction of policy change would make all groups better off, then the mechanism will find and implement change in that direction, in the sense of having a positive dot product with that direction.

Proposition 2. If the current policy vector $\boldsymbol{p}_t$ is Pareto dominated by any other policy $\boldsymbol{p}'$, then as long as the step size ϵ is sufficiently small, $\Delta\boldsymbol{p}_t \cdot (\boldsymbol{p}' - \boldsymbol{p}_t) > 0$.

This is equivalent to saying that the mechanism delivers Pareto improvements when they are available, but these will often be combined with movements in a direction orthogonal to the Pareto improvement. (These orthogonal movements typically will not themselves be Pareto improvements, and may mean that some types lose out from the overall policy adjustment, which is therefore not always a Pareto improvement. For example, the mechanism does not wait until after all Pareto improvements have been exhausted to also favor the interests of large majorities over the interests of small minorities.)

An immediate corollary is that if policies are adjusted by iterated application of the mechanism, the policy adjustments will not stop until a Pareto optimal point is reached.

Corollary to Proposition 2 (HZ). If $\boldsymbol{p}^*$ is a stationary point, then $\boldsymbol{p}^*$ is Pareto optimal.

Pareto optimality is generally viewed as an attractive feature of a social choice rule.[8]

If the Corollary to Proposition 2 is viewed as the iterated NGA mechanism's analog of competitive equilibrium's "first welfare theorem," then the next result is its analog of the "second welfare theorem": fixing the set of types and their preferences, for any Pareto optimal policy vector, there is a profile of population fractions that makes that vector a stationary point.

Proposition 3 (HZ). Fix the number of types Θ and their utility functions $u_1(\cdot), \ldots, u_\Theta(\cdot)$ (but not their population fractions). For any Pareto optimal policy vector $\boldsymbol{p}$, there exists a vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_\Theta)$ of strictly positive population fractions such that as long as the step size ϵ is sufficiently small, $\boldsymbol{p}$ is a stationary point.

[8] More generally, Proposition 2 implies that no limit cycle is Pareto dominated in the sense of having a point that Pareto dominates every point on the limit cycle.

An intuition for this result comes from the analogy between the NGA mechanism and social-welfare maximization from section I. In that analogy, the population fractions are the counterpart of the social welfare weights. Just as social-welfare maximization can deliver any Pareto optimum if the weights are chosen appropriately, the iterated NGA mechanism can do so if the population fractions and starting point are chosen appropriately.

A trivial case for the behavior of the mechanism is when one type θ comprises more than half the population. In that case, the mechanism always moves in the direction of θ's bliss point.

Proposition 4. If $\phi_\theta > \frac{1}{2}$ and the current policy vector $\boldsymbol{p}_t$ is not a bliss point for type θ, then as long as the step size ϵ is sufficiently small, $\Delta\boldsymbol{p}_t \cdot \boldsymbol{\nabla u}_\theta(\boldsymbol{p}_t) > 0$.

To understand this result, note that even if every other type prefers policy to change in the exact opposite direction, type θ's contribution to the direction of policy change is big enough to outweigh everyone else. It follows that when the mechanism is iterated and always moving in a direction preferred by type θ, it ends up at type θ's bliss point.

Corollary to Proposition 4. If $\phi_\theta > \frac{1}{2}$, then there is a unique stationary point $\boldsymbol{p}^*$, and it is equal to type θ's bliss point.

The voting interpretation of the mechanism provides a straightforward intuition: a fully unified majority with identical preferences gets what it wants.

This corollary implies that when there are two types in the population, the generic stationary point is the bliss point of the larger group.[9] Therefore, settings with only two types in the population are not generally useful for illustrating how the iterated mechanism would behave in real-world settings, which will typically involve many types. For this reason, our simple examples in what follows largely focus on settings with three types.

Contrary to the assumption of the last result, it would probably be much more common in practice that no type comprises a majority of the population. In that case, one might conjecture that the iterated mechanism would induce movement toward the types' bliss points collectively

[9] Hylland and Zeckhauser (1980) stated the result for the case of two types.

in some sense. This conjecture is not true without some restriction on the types' preferences; *any* policy vector at an arbitrary distance from the bliss points can be made a stationary point if the types' preferences can be chosen arbitrarily. Figure 7 illustrates an example with four types in which the stationary point occurs far from the bliss points. We will discuss this example after presenting the next pair of formal results.

To formalize our restriction on preferences, we define a measure of "maximum obliqueness." The obliqueness of a point $\boldsymbol{p}$ on an indifference curve is the angle (measured in radians) between $\boldsymbol{p}_\theta^{\text{bliss}} - \boldsymbol{p}$ and $\boldsymbol{\nabla u}_\theta(\boldsymbol{p})$. The *overall maximum obliqueness* α is defined as the supremum of this angle over all types and over every point for each type. In words, the overall maximum obliqueness α measures how far anyone's utility gradient can deviate from pointing directly toward her bliss point. For example, if all types had spherical preferences that could be represented by the Euclidean distance $u_\theta(\boldsymbol{p}) = -\left\|\boldsymbol{p}_\theta^{\text{bliss}} - \boldsymbol{p}\right\|$ (or any monotonic transformation of that distance) the overall maximum obliqueness α would be zero. In contrast, the example in Figure 7 shows a case where the types' indifference curves have very high maximum obliqueness: in some places the gradient is almost orthogonal to the vector pointing to the bliss point. What will be needed for the proposition below is that the overall maximum obliqueness never reaches orthogonality nor asymptotes to it: $\alpha < \frac{\pi}{2}$. The most commonly used functional forms for preferences with bliss points will satisfy this requirement. For example, it is satisfied by any set of preferences that are homothetic around the bliss point, including quadratic preferences (defined by ellipsoidal indifference surfaces centered on a bliss point).[10]

To further facilitate stating the proposition, we define the *bliss-point circumsphere* as the smallest sphere that circumscribes the bliss points. It has center $\boldsymbol{p}_C \equiv \text{argmin}_{\boldsymbol{p}} \max_\theta \left\|\boldsymbol{p} -\right.$ $\boldsymbol{p}\theta$bliss and radius r≡max$\theta\boldsymbol{p}C-\boldsymbol{p}\theta$bliss.

Proposition 5. Suppose that there is some $\alpha < \frac{\pi}{2}$ such that the overall maximum obliqueness is α. If $\|\boldsymbol{p}_t - \boldsymbol{p}_C\| > \frac{r}{\cos\alpha}$, then as long as the step size ϵ is sufficiently small, $\Delta\boldsymbol{p}_t \cdot (\boldsymbol{p}_C - \boldsymbol{p}_t) > 0$.

[10] To see why, note that concavity of $u_\theta(\boldsymbol{p})$ implies that for any $\boldsymbol{p}$, the angle between $\boldsymbol{p}_\theta^{\text{bliss}} - \boldsymbol{p}$ and $\boldsymbol{\nabla u}_\theta(\boldsymbol{p})$ is less than or equal to $\frac{\pi}{2}$. Note also that there is a maximum angle on the unit sphere around a type's bliss point (because the unit sphere is a compact set), and this maximum is strictly less than $\frac{\pi}{2}$. For a type with homothetic preferences, the maximum angle on the unit sphere is also the maximum angle on any sphere around the bliss point, and it is therefore the maximum over all $\boldsymbol{p}$.

Proposition 5 states that if the maximum obliqueness over all types is small enough and if the current policy vector is outside a sphere that is an expansion of the bliss-point circumsphere by a factor of $\frac{1}{\cos\alpha}$ (out from the center of the bliss-point circumsphere), then the mechanism will move toward (in a positive-dot-product sense) the center of the bliss-point sphere. The size of this expanded sphere is as small as possible if all types have Euclidean-distance preferences, in which case $\alpha = 0$ and $\cos\alpha = 1$. The radius of the expanded sphere, $\frac{r}{\cos\alpha}$, is finite as long as $\alpha < \frac{\pi}{2}$.

Proposition 5 immediately implies that any stationary point (and any limit cycle) must lie within this region.

Corollary to Proposition 5. If the maximum obliqueness is $\alpha < \frac{\pi}{2}$, then any stationary point $\boldsymbol{p}^*$ is within a sphere of radius $\frac{r}{\cos\alpha}$ around the center of the bliss-point sphere, i.e., $\|\boldsymbol{p}^* - \boldsymbol{p}_C\| \leq \frac{r}{\cos\alpha}$.

The example in Figure 7 satisfies the conditions of this corollary because the types' preferences are quadratic, but the stationary point is far from the center of the bliss-point sphere because the preferences have high maximum obliqueness. Figure 7 also illustrates, however, that having a stationary point that is far from the bliss points in policy space does not necessarily mean that the types are on a much lower indifference curve. In fact, by converging to a stationary point that is far from everyone's bliss point in policy space, but close to the major axis of everyone's elliptical indifference curves, the iterated mechanism succeeds in putting everyone on a relatively high indifference curve.

Proposition 6 extends Proposition 5 to handle the case when there are a few outlying types with bliss points far from the other types, giving conditions under which the mechanism will move toward the center of the circumsphere of the bliss points of a subset of "inside types."

Proposition 6. Let *J* be a subset of types whose bliss points are all within a circumsphere with radius r, and center p_c and let $\omega \equiv \frac{\sum_{j \notin J} \phi_j}{\sum_{j \in J} \phi_j}$. Suppose $\alpha < \frac{\pi}{2}$ is the overall maximum obliqueness for the subset of types J and $\omega < \cos(\alpha) < 1$. Then if $\|\boldsymbol{p}_t - \boldsymbol{p}_C\| > \frac{r}{(\sqrt{1-\omega^2}\ )\cos(\alpha) - \omega\sin(\alpha)}$, and the step size ϵ is sufficiently small, $\Delta\boldsymbol{p}_t \cdot (\boldsymbol{p}_C - \boldsymbol{p}_t) > 0$.

A corollary follows immediately:

Corollary to Proposition 6. Suppose $\alpha < \frac{\pi}{2}$ is the overall maximum obliqueness for the subset of types J and $\omega < \cos(\alpha) < 1$. Then any stationary point $\boldsymbol{p}^*$ is within a sphere of radius $\frac{r}{(\sqrt{1-\omega^2}\ )\cos(\alpha) - \omega\sin(\alpha)}$ around the center of the circumsphere for the bliss points of the subset of types J.

Note that $\omega < 1$ is only possible if the "inside" group is a strict majority. The importance of Proposition 6 and its corollary is in showing that the mechanism does not allow a small group of "holdouts" to prevent a substantial majority from getting what they want. How close the mechanism is guaranteed to get to the center of the circumsphere of that majority depends on both how big a majority it is and on the degree of disagreement within that majority, where the level of disagreement is indicated by both the size of the sphere needed to contain the bliss points of everyone in that majority and the overall maximum obliqueness in that majority.[11]

IV.C. Assessing the Limitations of the Iterated Mechanism

In this subsection we explore some of the iterated mechanism's limitations, with a combination of analytic results and simulations. We first explore the possibility of limit cycles and multiple stationary points, and then the possibility of manipulation of the mechanism's outcome through the choice of policy units.

IV.C.1. The Possibility of Limit Cycles and Multiple Stationary Points

As noted above, Proposition 1 proves the existence of a stationary point but does not rule out that there may also be additional stationary points or limit cycles. We have not proven that

[11] Proposition 2 and its corollary are not special cases of Proposition 6 and its corollary, since with one fully unified type, the maximum obliqueness does not matter. The overall maximum obliqueness in the majority matters in Proposition 6 because it indicates the degree to which the gradients of different types within that majority might cancel each other out, even when their bliss points are very close to one another. By contrast, individuals within a single type never cancel each others' gradients out since their preferences are identical.

the iterated mechanism converges to a unique stationary point under *general* conditions, but the next pair of results provides a specific, highly restrictive condition. Specifically, we assume that each type has quadratic preferences, $u_\theta(\boldsymbol{p}) = -(\boldsymbol{p} - \boldsymbol{p}_\theta^{\text{bliss}})\boldsymbol{\Omega}(\boldsymbol{p} - \boldsymbol{p}_\theta^{\text{bliss}})$, where $\boldsymbol{\Omega}$ is a positive definite matrix common to all types.

Proposition 7. Suppose that each type has preferences $u_\theta(\boldsymbol{p}) = -(\boldsymbol{p} - \boldsymbol{p}_\theta^{\text{bliss}})\boldsymbol{\Omega}(\boldsymbol{p} - \boldsymbol{p}_\theta^{\text{bliss}})$, where $\boldsymbol{\Omega}$ is a positive definite matrix common to all types. If $\boldsymbol{p}_t \neq \boldsymbol{p}_\theta^{\text{bliss}}$ for all θ, then the function $W(\boldsymbol{p}) \equiv -\sum_\theta \phi_\theta \left\| \boldsymbol{\Omega}(\boldsymbol{p} - \boldsymbol{p}_\theta^{\text{bliss}}) \right\|$ is strictly concave and differentiable at $\boldsymbol{p}_t$, and as long as the step size ϵ is sufficiently small, $\Delta\boldsymbol{p}_t \cdot \nabla W(\boldsymbol{p}_t) \geq 0$, with strict inequality unless $\boldsymbol{p}_t$ is the global maximum of W.

Proposition 7 essentially says that there is a concave function, $W(\boldsymbol{p})$, such that the mechanism moves policy in a direction that increases its value; in other words, it is a Lyapunov function. Proposition 7, together with an argument in the proof of the following corollary to handle cases when the iterated mechanism hits bliss points or ends up at a bliss point, implies that the iterated mechanism will converge to a unique stationary point.

Corollary to Proposition 7. If each type has preferences $u_\theta(\boldsymbol{p}) = -(\boldsymbol{p} - \boldsymbol{p}_\theta^{\text{bliss}})\boldsymbol{\Omega}(\boldsymbol{p} - \boldsymbol{p}_\theta^{\text{bliss}})$, where $\boldsymbol{\Omega}$ is a positive definite matrix common to all types, then there is a unique stationary point that is the unique maximizer of the function $W(\boldsymbol{p})$.

An interesting special case is when the matrix $\boldsymbol{\Omega}$ is the identity matrix—the case where each type has Euclidean-distance preferences. In that case, $W(\boldsymbol{p})$ is simply minus the weighted sum of distances to every bliss point: $W(\boldsymbol{p}) \equiv -\sum_\theta \phi_\theta \left\| \boldsymbol{p} - \boldsymbol{p}_\theta^{\text{bliss}} \right\|$. In this case when the matrix $\boldsymbol{\Omega}$ is the identity matrix, $W(\boldsymbol{p})$ is in fact a social welfare function in the sense that it can be written as a function of the utility functions: $W(\boldsymbol{p}) \equiv -\sum_\theta \phi_\theta \sqrt{-u_\theta(\boldsymbol{p})}$. This is a well-behaved social welfare function *except* that it is not smooth, since it has kinks at bliss points. In view of this lack of smoothness we view this "social welfare function" more as a useful mathematical device for describing the behavior of the mechanism than as something with independent normative significance. In this special case, the iterated mechanism can be understood as hill-climbing and ultimately maximizing this "social welfare function."[12]

Simulation experiments. In order to explore a wider range of cases than quadratic preferences that are identical except for their bliss points, we conducted several simulation experiments. We allowed types to differ in the shapes of their preferences, but we maintained the assumption for all types in all simulations that preferences are quadratic. We did so because quadratic preferences are a second-order Taylor approximation to any smooth preferences, are among the simplest preferences to study that have a bliss point, and (despite their simplicity) generate substantial richness and variety in simulation results.

Across simulation cases, we varied the number of policy dimensions (2, 3, 4, or 25), the number of types (3, 4, or 10), and their population fractions and preferences, as described next. The specific numbers and probability distributions we used are arbitrary, but the different cases we explored had particular purposes, as we explain below.[13] In each simulation case, we aimed to reach 400 simulation runs, but we exceeded this number in a few cases.

The population fractions were either all equal or randomly chosen. For this and other random features of our simulations (described below), all variables are i.i.d., and we usually use the exponential distribution with parameter value equal to one defined by the density function $f(x) = e^{-x}$. We use this "unit-mean exponential distribution" because it has no additional parameters, is constrained to be positive, and has substantial variance. To choose the population fractions randomly, we drew a unit-mean exponential random variable x_θ for each type θ, and then we normalized the realized values to obtain the population fractions: $\phi_\theta = \frac{x_\theta}{\sum_{\theta'=1}^{\Theta} x_{\theta'}}$.

Quadratic preferences have homothetic ellipsoidal indifference surfaces that can be fully characterized by the location of the bliss point and the directions and lengths of the two axes. For each type, the location of the bliss point was random: the direction of the bliss point from the

[12] Under what other circumstances could the behavior of the iterated mechanism be consistent with hill-climbing a social-welfare function? What is needed is for each group to have preferences with a representation that makes the normalized gradient at every point equal to the gradient before normalization. That equality is equivalent to a nontrivial partial differential equation that would generate a family of indifference surfaces from a single, initially specified, smooth indifference surface. We have not studied this problem in depth, but in our preliminary investigations, we have not been able to find any cases where the whole set of indifference surfaces appears relatively simple other than the case of spherical indifference surfaces.

[13] Upon publication, we will make our simulation code publicly available via the journal's website to facilitate further simulation experiments.

origin was drawn from the uniform distribution on the unit sphere, and the distance of the bliss point from the origin was drawn from the unit-mean exponential distribution. The direction of the first axis was drawn from the uniform distribution on the unit sphere, and the direction of the second axis was drawn from the uniform distribution on the unit sphere in the $(P-1)$–dimensional subspace that is orthogonal to the direction of the first axis, and so on for the other axes in higher dimensions.

For the lengths of the axes, we conducted two different types of simulations. In the "non-extreme" simulations, the length of each axis was drawn from the unit-mean exponential distribution. In the "extreme" simulations, the length of each axis was drawn from a mixture distribution: the unit-mean exponential distribution with 50% probability and the $\frac{1}{10}$-mean exponential distribution (that has probability density $f(x) = 10e^{-10x}$) with 50% probability. Preferences in which one axis is much longer than the other are extreme in the sense that the agent cares much more strongly about movement that is orthogonal to the shorter axis.

Some of our simulation-design decisions had particular rationales. We view the simulations with random population fractions and "non-extreme" axis lengths as more realistic, but we studied equal population fractions and "extreme" axis lengths because we anticipated that these cases would make limit cycles or multiple stationary points more likely. Intuitively, equal population fractions of types with preferences symmetric to each other would imply that if there is one non-centered stationary point, then there must be another stationary point because the vector field induced by the mechanism would be symmetric. And "extreme" ellipsoidal indifference curves (planes, in the perfectly extreme limit) would correspond to nearly lexicographic preferences, implying that a type' utility gradient shifts direction nearly discontinuously whenever its currently highest-ranked policy preference is satisfied. Indeed, prior to running simulations, we constructed examples of multiple stationary points by hand based on these intuitions.[14] In addition, we view the cases with $P > 2$ policy dimensions as more

[14] The key example with extreme preferences that we constructed and analyzed by hand had infinitely elliptical ellipses that looked like straight lines forming an isosceles triangle with an obtuse 178 degree angle at the apex, and bliss points for the different types at the center of each "side." In this setting, the lexicographic aspect of the preferences means that when it starts below the long base of the triangle, the mechanism with an infinitesimal step size goes straight up until it hits one of the upper sides of the triangle, after which it proceeds to the bliss point on the upper side it hits first. (Thus, two of the three bliss points are both stationary points.) Note that that example has mirror symmetry. Another highly symmetric example that we specified by hand, but analyzed with a simulation, had four equal-weighted types with bliss points in a rectangle. The major axes of the elliptical indifference curves had a

realistic, but the two-dimensional cases enabled us to graph the simulation output and build intuition.

Examples of the results from our simulations are in Appendix B.

Results on limit cycles. Table 1 shows, across the different types of simulation experiments, the frequency at which limit cycles were observed. We observed zero instances of limit cycles.

We conducted roughly 20 additional simulation runs with the specific purpose of trying to generate highly spiral policy trajectories that may be most likely to generate limit cycles and also found none in these cases. Specifically, we assumed equal population fractions for each type and arrayed their bliss points in a radially symmetric manner in a circle. We varied the ratio of the length of the major axis to the length of the minor axis. Each type's elliptical indifference curves had a major axis at a given angle relative to the vector from the center of the circle to the bliss point. We ran simulations with 15-, 45-, and 75-degree angles. We observed no cycles. Figure 8 shows one example from a simulation with 120 types.

While we caution that our simulations cover only a very small part of the spectrum of possible preference profiles, we tentatively conclude from these results that limit cycles may not be particularly problematic in practice.

Results on multiple stationary points. Across our simulations, Table 2 shows the frequency at which multiple stationary points were observed. We had anticipated that the "equal weight extreme" simulation cases would be the most likely to have multiple stationary points. It turned out that the "random weight extreme" cases had a similar likelihood to the "equal weight extreme" cases, and these two generally had higher frequencies than the "non-extreme" cases. However, multiple stationary points were always reasonably uncommon. In particular, whenever there were 10 types or at least 4 policy dimensions, the frequency of multiple stationary points was never more than 2%.

We suspect that multiple stationary points are not a major limitation of the iterated mechanism, for two reasons. First, as noted, they are pretty rare in our simulation results. Second, even though such path-dependence is theoretically unappealing, we conjecture that

fixed angle with the ray from the center of the rectangle to the relevant bliss point. Varying the angle and the ratio of the length of the sides of the rectangle quickly yielded cases with multiple stationary points.

people would nonetheless be willing to accept a policy-change sequence that starts at the status quo and along which they view each step as an improvement.

IV.C.2. Manipulation of the policy-distance matrix

As we have emphasized, we believe that the biggest limitation of the NGA mechanism is the potential for manipulation via choice of the policy units. In section I, we illustrated how such manipulation might work for the one-shot mechanism. For the iterated mechanism, the next result provides conditions under which a given type θ, if it determines the policy-distance matrix $\boldsymbol{M}$, can manipulate the mechanism to get its bliss point $\boldsymbol{p}_\theta^{\text{bliss}}$ to be a stationary point.

The key condition is that type θ is *pivotal in at least one direction at its bliss point*: there exists $\boldsymbol{h} \in \mathbb{R}^P$ defining a hyperplane $H = \{\boldsymbol{p} \in \mathbb{R}^P : \boldsymbol{p} \cdot \boldsymbol{h} = 0\}$ such that $\boldsymbol{\nabla u}_{\hat{\theta}}\big(\boldsymbol{p}_\theta^{\text{bliss}}\big) \notin H$ for all $\hat{\theta} \neq \theta$, and

$$\left| \sum_{\hat{\theta} \in \Theta_-} \phi_{\hat{\theta}} - \sum_{\hat{\theta} \in \Theta_+} \phi_{\hat{\theta}} \right| < \phi_\theta$$

where $\Theta_- = \big\{\hat{\theta} \in \Theta : \boldsymbol{\nabla u}_{\hat{\theta}}\big(\boldsymbol{p}_\theta^{\text{bliss}}\big) \cdot \boldsymbol{h} < 0 \text{ and } \hat{\theta} \neq \theta\big\}$ is the lower half-space defined by H, and Θ_+ is the analogously defined upper half-space. In words, there is a separating hyperplane between the gradients of the other types at θ's bliss point such that the difference between the fraction of the population preferring to move to one side of the hyperplane and the fraction preferring to move to the other side is smaller than θ's fraction of the population. Note that type θ is pivotal in at least one direction at its bliss point if *any* such hyperplane can be found.

Proposition 8. If a type θ is pivotal in at least one direction at its bliss point, then there is some (positive definite) policy-distance matrix $\boldsymbol{M}$ such that type θ's bliss point is a stationary point.

Intuitively, type θ can "divide and conquer": by shrinking distances in the direction orthogonal to the hyperplane, the marginal utilities in that direction can be made extremely big (relative to the marginal utilities in other directions), so that the other types expend virtually all of their votes fighting over which way to go in that direction, and then type θ can cast the deciding vote to remain at the status quo. Figures 9a and 9b illustrate the proposition in a situation with three equally weighted types, where type 3 is pivotal in at least one direction at its bliss point. In

Figure 9a, type 3 cannot reach its bliss point, but Figure 9b shows a transformation of the coordinate system such that type 3 reaches its bliss point (and chooses not to overshoot).

The intuition underlying the proposition makes clear why, as noted in section I.E., if policy units are chosen to make types' indifference curves closer to spherical, then the scope for manipulation is reduced: even if a given type is pivotal in at least one direction at its bliss point, exploiting that position requires choosing policy units such that the indifference curves become highly oblique. More formally, a criterion such as choosing policy units to minimize the overall maximum obliqueness (or some measure of average obliqueness) of the preferences would make the mechanism harder to manipulate.

While Proposition 8 demonstrates the potential for extreme manipulation, it assumes that one type has complete control over the policy units. A more realistic political-economy worry is that some type could influence (but not fully control) the policy units. Also, it is unlikely that that type would have an equal opportunity to manipulate in any vector direction it chose; the opportunity to manipulate in a substantial way might be only along a limited number of coordinate axes. To shed some light on how problematic such influence could be, in the remainder of this section, we report simulation experiments that examine how variation in one of the policy-dimension units affect the stationary points of the iterated mechanism. While it is difficult to know how much influence would be realistic, we examine changes (both increases and decreases) by factors of $\frac{3}{2}$ and 3, magnitudes that intuitively appear to us rather large.

Simulation experiments. We ran the same simulation cases as described above in section IV.C.1, except that we omitted the "extreme" axis cases. The one new feature is that after conducting each simulation run, we ran an otherwise-identical simulation except that we changed the policy units by multiplying the $P = 1$ policy-dimension coordinate by $\frac{1}{3}, \frac{2}{3}, \frac{3}{2}$, or 3 before executing the iterated mechanism appropriate to the new coordinate system. (Note that the decreases by the first two factors will not necessarily generate the same results as the increases by the second two factors because when $P > 2$, making one dimension longer relative to the $P - 1$ other dimensions is not the same as making it shorter relative to them.) Then we undid the coordinate transformation to make the terminal stationary point chosen by the iterated mechanism after the coordinate transformation comparable to the terminal stationary point chosen by the iterated mechanism with the original coordinate system. We then assessed the distance d between the two terminal stationary points.

In order to report results in units that are comparable across simulation runs, we normalized d in two ways. First, we divided all distances by the average distance between bliss points in 2-dimensional, 3-type simulations: 1.554. Then, for each simulation run, we divided by the ratio of the average distance between bliss points for that run to the mean over all simulations we conducted with the same number of types and dimensions of the average (within-run) distance between bliss points. Within a given simulation case, this procedure of normalizing the distances avoids reporting especially large and small distances merely as a result of a layout in which bliss points happened to be especially far apart or especially close together. And using the distance between bliss points in the 2-dimensional, 3-type simulations as the numeraire gets closer to making the normalized yardstick comparable to the distance between *neighboring* bliss points, as opposed to an average distance between bliss points that are more likely to be far apart for at least some pairs of bliss points when there are many types.

Results on changes in the policy-distance matrix. Figure 10 shows the distribution of normalized distances from these simulations, in each case ordered from largest to smallest across the 100 runs. In all simulation cases, it is quite rare for the normalized distance to exceed 1, i.e., to move the average distance between one bliss point and the next bliss point over, and in many simulation cases, the normalized distance is close to zero in the majority of runs.

Three patterns are apparent from comparing results across simulation cases. First, as expected, the normalized distances are generally larger in the simulation runs with the more extreme dimension rescaling factors of $\frac{1}{3}$ and 3 compared to the less extreme factors of $\frac{2}{3}$ and $\frac{3}{2}$. Second, as the number of policy dimensions increases from $P = 2$ to 3 to 4, the normalized distances tend to be smaller. Finally, for the simulation cases where the types have equal weight, as the number of types increases from $\Theta = 3$ to 4 to 10, the normalized distances tend to be smaller. The effect of increasing the number of types is less clear in the random-weight cases, but these cases may be expected to generate noisier results because some of the types will end up with relatively small weights, and thus the effective number of types will be smaller, and because random weights lead to a wider space of parameter values being explored, which increases the variance in distance.

Given that in practice, we imagine that the numbers of types and policies would both be at least as large as those we have examined in our simulations, we view our simulation results overall as somewhat reassuring. For example, in the case with $(P, \Theta) = (4{,}10)$ and equal weights, across all simulation runs the normalized distances are never more than 0.25 and almost always smaller than 0.10.

V. Discussion

In our analysis, we have treated all relevant policies as if they were control variables that could be adjusted quickly, with quick results. In practice, for many policies, it will take time to assess the extent to which individuals became better or worse off as a result of a policy change. In addition, in choosing between more frequently taking a smaller step size vs. less frequently taking a larger one, there is a tradeoff between having the gradients be better approximations to the indifference surfaces and providing adequate time for the assessment of delayed effects. Moreover, step sizes and frequency both need to be large enough to have a reasonable speed of progress toward the eventual destination of the iterated mechanism.[15]

While the mechanism could in principle be applied to policies with delayed effects, such as policies affecting climate change, it would have to proceed somewhat differently than we have described throughout the paper. Much like cost-benefit analyses of such policies are conducted today, predictions about delayed policy effects on the state of the world would be used in conjunction with estimates of how such effects would matter for the utility proxy (here interpreted as a measure of instantaneous utility) to generate an estimate of policy effects on expected lifetime (discounted) utility. These estimates of effects on expected lifetime utility would be the inputs to the mechanism. After the policies are adjusted, the agency would formulate new predictions about the effects of further policy adjustments, and these would be used in the next iteration.

Throughout this paper, we have assumed that the policy space is $\mathbb{R}^P$, but in practice, some policies are bounded (e.g., government spending on parks cannot realistically fall below zero). With a bounded policy space, the analysis would carry over without modification as long as the policy vector always remains in the interior of the space. An extension of the mechanism

[15] While we have emphasized the issue of manipulation of the coordinate system for the policy space, another area where there is a danger of manipulation is those who benefit from the status quo lobbying for a slower pace of progress toward the terminal stationary point.

to the case where the policy vector reaches an edge of the space is yet to be worked out. Furthermore, while simple constraints among policies are still consistent with the existence of a representation of the policy space by $\mathbb{R}^P$, complex constraints among policies would require adjustments to how the mechanism is specified that we have not worked out.

In our analysis, the application of the mechanism is limited to policies that can be varied continuously. In rare cases, it might nonetheless be possible to apply the mechanism to discrete policies by convexifying the policy space, i.e., making the policy variable the fraction of time at each discrete policy level, or the probability of each discrete policy level. In most cases, however, such convexification would be impractical. It remains an open question whether the mechanism could be extended in some other way to accommodate discrete policies.

One challenge to implementing the NGA mechanism in practice is identifying the types in the population. In theory, different individuals are of the same type if they share the same local indifference surface over policies. In practice, sharing exactly the same indifference surface is unlikely, and the government agency would need to settle for partitioning the population into approximate types. One theoretically unjustified yet practical way to proceed would be to treat subgroups defined by a standard, small set of observables (such as sex, age, etc.) as types. But if there is heterogeneity in policy effects within these groups, then estimating the policy effects requires treating the well-being survey responses as interpersonally comparable within each group—i.e., it requires the assumption that within a group, respondents' use of the well-being survey-response scale is uncorrelated with the policy effects. This is precisely the assumption that an ordinal approach to aggregation is intended to help avoid. To minimize such heterogeneity, the agency would ideally conduct a pilot study, tracking well-being and measuring a large set of observables in the sample, estimating policy effects in this sample for different groups, and then try to identify a smaller set of observables that could be measured regularly and used to characterize types. Alternatively, type membership could be treated as unobservable and the agency could use statistical methods to partition the population into types.

On the other hand, on the theoretically justified extreme of the tradeoff between theoretical justification and practicality, every individual could be treated as her own type (with random sampling of the individuals surveyed so that the measured preferences are representative of the population). In practice, however, it would be impossible to implement the NGA mechanism this way using well-being data because, as long as the mechanism adjusts the policy levels at every iteration, local policy effects cannot be estimated at the individual level. Treating individuals as types would require constructing reported gradients from contingent-valuation-type surveys that directly elicit local policy preferences instead of from estimated policy effects on well-being data. But such contingent-valuation surveys crucially rely on the assumption that the population is well informed regarding the effects of policy—another assumption that our approach is intended to avoid. In summary, identifying the types in a practical yet theoretically justified manner remains an open challenge.

There are three ways in which the NGA mechanism could become a reality. The story we have carried through this paper—a government agency with comprehensive power that not only computes the NGA mechanism but also executes policy changes—is the first and least likely.

The second is a government agency with limited powers implementing policy changes in a few dimensions within its scope. Because the agency would need to have the legal authority to adjust policies regularly (without legislative approval) and based on a technocratic procedure, it would presumably focus on regulatory policies. In addition to neglecting people's preferences over many other dimensions of policy, as we noted before, such a limited implementation would entail loss of the information on preference intensities from measuring marginal rates of substitution over many dimensions of policy.

The third and perhaps most likely way in which the NGA mechanism could be realized is through a government agency addressing a broad range of policy dimensions, but with powers limited to data collection, statistical analysis, and moral suasion on the rest of the government. In the U.S., a model might be the Congressional Budget Office, which strives to be an impartial arbiter giving input into a wide range of legislative decisions. Because any power this agency had would rest on its credibility, it would have a strong bureaucratic incentive to make a good-faith effort to do its statistical job carefully. Since the agency would be charged with estimating the effects of policy, it could also play a major role in encouraging policies to be adopted in ways that enable credible econometric identification of their effects.

We have described the NGA mechanism as aggregating the effects of policy on a utility proxy, for example one derived from well-being survey data. However, the mechanism could be applied whenever the government agency has access to estimates of the marginal rates of substitution across policies for different types in the population. For example, if the population is well-informed about policy and its effects, contingent-valuation surveys that ask respondents to make tradeoffs between alternative policy adjustments could be used to obtain such marginal rate of substitution estimates. Estimates of the compensating variation of policies obtained with standard revealed-preference methods could also serve as an input into the mechanism. Indeed, policy effect estimates obtained through different methods could be used in combination, as long as it is possible to calculate a gradient vector for each type.

Thus, the program we have discussed in this paper, while still far from addressing all the theoretical and practical questions—and hence still far from being ready for immediate adoption—provides a broadly applicable procedure based on democratic principles for aggregating estimates of the local effects of policy. It may achieve its greatest potential when used with a utility proxy derived from well-being measures that track a broad set of aspects of well being, including sense of purpose, quality of social relationships, and freedoms. Because such data may capture more of what matters to people than standard economic indicators alone do, a procedure for guiding policy based on these data may over time lead to lasting improvements in well being.

References

Barberá, Salvador, Anna Bogomolnaia, and Hans van der Stel. 1998. “Strategy-proof probabilistic rules for expected utility maximizers.” *Mathematical Social Sciences*, 35(2), 89–103.

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Alex Rees-Jones. 2012. “What Do You Think Would Make You Happier? What Do You Think You Would Choose?” *American Economic Review*, 102(5), 2083–2110.

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Alex Rees-Jones. 2013. “Can Revealed-Preference Tradeoffs Be Inferred From Happiness Data? Evidence from Residency Choices.” *American Economic Review*, forthcoming.

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Nichole Szembrot. 2013. “Aggregating Local Preferences to Guide Marginal Policy Adjustments.” *American Economic Review Papers and Proceedings*, 103(3): 605–610.

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Nichole Szembrot. 2014. “Beyond Happiness and Satisfaction: Toward Well-Being Indices Based on Stated Preference.” *American Economic Review*, 104(9): 2698–2735.

Chung, Hun, and John Duggan. 2014. “Directional Equilibria.” University of Rochester mimeo, February.

Feldstein, Martin. 1976. “On the theory of tax reform.” *Journal of Public Economics*, 6, 77-104.

Ferrer-i-Carbonell, Ada, and Paul Frijters. 2004. “How important is methodology for the estimates of the determinants of happiness?” *Economic Journal*, 114, 641–659.

Fleurbaey, Marc, and Didier Blanchet. 2013. *Beyond GDP*. Oxford: Oxford University Press.

Frey, Bruno, and Alois Stutzer. 2012. “The Use of Happiness Research for Public Policy.” *Social Choice and Welfare*, 38(4): 659–74.

Hylland, Aanund, and Richard Zeckhauser. 1980. “A mechanism for selecting public goods when preferences must be elicited.” Harvard University mimeo, December.

Maskin, Eric. 1999. “Nash equilibrium and welfare optimality.” *Review of Economic Studies*, 66(1), 23–38.

Muller, Eitan, and Mark A. Satterthwaite. 1977. “The equivalence of strong positive association and strategy-proofness.” *Journal of Economic Theory*, 14(2), 412–418.

Rawls, John. 1971. *A Theory of Justice*. Cambridge: Harvard University Press.

Stiglitz, Joseph E., Amartya Sen, and Jean-Paul Fitoussi. 2009. *Report by the Commission on the Measurement of Economic Performance and Social Progress*.

Table 1: Frequency of Cycles

	3-Type				4-Type				10-Type			
	Random Weight	Equal Weight	Random Weight Extreme	Equal Weight Extreme	Random Weight	Equal Weight	Random Weight Extreme	Equal Weight Extreme	Random Weight	Equal Weight	Random Weight Extreme	Equal Weight Extreme
P=2 Policy Dimensions	0.00%* (0/400)	0.00% (0/416)	0.00% (0/412)	0.00% (0/400)	0.00% (0/400)	0.00% (0/405)	0.00% (0/402)	0.00% (0/145)	0.00% (0/410)	0.00% (0/450)	0.00% (0/402)	0.00% (0/400)
P=3	0.00% (0/400)	0.00% (0/410)	0.00% (0/480)	0.00% (0/400)	0.00% (0/463)	0.00% (0/405)	0.00% (0/420)	0.00% (0/416)	0.00% (0/450)	0.00% (0/407)	0.00% (0/400)	0.00% (0/400)
P=4	0.00% (0/400)	0.00% (0/430)	0.00% (0/400)	0.00% (0/400)	0.00% (0/400)	0.00% (0/410)	0.00% (0/400)	0.00% (0/400)	0.00% (0/400)	0.00% (0/400)	0.00% (0/430)	0.00% (0/420)

Table 2: Frequency of Multiple Stationary Points

	3-Type				**4-Type**				**10-Type**			
	Random Weight	**Equal Weight**	**Random Weight Extreme**	**Equal Weight Extreme**	**Random Weight**	**Equal Weight**	**Random Weight Extreme**	**Equal Weight Extreme**	**Random Weight**	**Equal Weight**	**Random Weight Extreme**	**Equal Weight Extreme**
P=2 Policy Dimensions	0.75% (3/400)	0.24% (1/416)	0.97% (4/412)	2.25% (9/400)	1.00% (4/400)	5.43% (22/405)	2.49% (10/402)	7.95% (33/145)	0.24% (1/410)	0.00% (0/450)	2.00% (8/402)	1.75% (7/400)
P=3	0.75% (3/400)	1.46% (6/410)	1.25% (6/480)	1.25% (5/400)	0.43% (2/463)	1.23% (5/405)	1.43% (6/420)	2.64% (11/416)	0.00% (0/450)	0.25% (1/407)	0.50% (2/400)	0.75% (3/400)
P=4	0.75% (3/400)	0.23% (1/430)	1.50% (6/400)	1.00% (4/400)	0.75% (3/400)	073% (3/410)	0.75% (3/400)	1.25% (5/400)	0.00% (0/400)	0.00% (0/400)	0.70% (4/430)	0.24% (1/420)