Values, Preferences, and Meaning*

Joe Edelman†

10 June 2021 — DRAFT

Abstract

Many fields (social choice, welfare economics, recommender systems) assume people express what benefits them via their ‘revealed preferences.’ Revealed preferences have well-documented problems when used this way, but are hard to displace in these fields because, as an information source, they are private, local, verifiable, fine-grained, battle-tested, and easily-expressed. In order to compete, other information sources (about participants’ values, capabilities and functionings, etc) would need to match this. I present a conception of values which shares many of these features with revealed preference.

Introduction

As an information basis for welfare, optimality, social choice, etc—revealed preference has been much critiqued. All the same, it’s hard to see how to move away from it, without paternalism: if people know best for their own lives, shouldn’t we trust their choices, and avoid imposing “better values” from above?

Suppose people do know best, and have wise values, but that their revealed preferences aren’t the last word on those values. To establish this, you’d want another source of information on people’s values. One that operates at a similar resolution and reliability.

On Preferences

A rich literature in the social sciences covers how revealed preferences, summed up into engagement metrics, lead us astray.

These problems can be summed up as stemming from two sources, often intermingled. First, there can be incompleteness or bias in the option set framed by the preference. Second, outside pressure can make the choice less than a complete expression of the agent’s interests.

Limited or biased options. Alexa buys a car because there’s no local transit. This counts as a preference for cars. Ben wants to move to a different city, but only if his friends also move.¹ Ben and his friends have, unfortunately, revealed a preference for their current city. Carl and his friends want to play tennis more—but they’re choosing individually, from a menu of bookable tennis experiences. It looks like a preference to play tennis separately—even a rivalry² for the same tennis court.

Outside pressure. Dan does something he later regrets, due to social pressure, misinformation, or a manipulative ad. Eli is following a social rule which she knows is not in her best interest.³ Preference!

These problems are not new. In his 1938 paper, Samuelson warned against using revealed preference as a measure of welfare, for presumably these reasons:

I should like to state my personal opinion that nothing said here... touches upon at any point the problem of welfare economics, except in the sense of revealing the confusion in the traditional theory of these distinct subjects.⁴

¹See discussions of the prisoner’s dilemma in e.g., Sen, “Behaviour and the Concept of Preference”; Anderson, “Symposium on Amartya Sen’s Philosophy.”
²Edelman, “Towards ‘Game B’.”
⁴Samuelson1938?

---

*Thanks to B. Gabbai, A. Morris, A. Ovadya, J. Stray, & F. Noriega
†School for Social Design joe.edelman@gmail.com
A similar sentiment was put forth by Arrow regarding social choice, and more recently by Sen, Anderson, etc.

Why, then, are revealed preferences still fundamental to our conceptions of benefit, and to measures of optimality? One reason is their advantages as an information source. These concern the resolution and robustness of the data they provide, the universality of their application:

Resolution:
- **Local.** Preferences are informed by the local situations and priorities of user-citizens. Most alternatives assume outside experts know better than user-citizens what would help them.
- **Fine-grained.** Preferences say subtle things about how a person wants to live.

Robustness:
- **Battle-tested.** We often say we want things, but don’t choose them in the final analysis. Preferences get at our real priorities.
- **Private.** Preferences are often expressed away from the signaling of allegiances, so they’re less influenced by social pressure.
- **Verifiable.** Engagements leave a trail. Who did you actually vote for? What did you purchase?

And universality:
- **Easily-expressed.** You don’t need to be particularly introspective or good with words to act on a gut feeling about what’s right for you.
- **Comprehensive.** Preferences do not limit us to one domain of life—the moral, the self-interested, etc, but represent all-things-considered judgements, and apply to many kinds of choices.

**Attentional Policies**

A revealed preference considers the options as given. As other authors have pointed out, this doesn’t fully reflect how choice works.

What would preference look like, if it were abstracted away from a (possibly biased or incomplete) set of options?

Consider my choice, with colleagues, to say something witty. At $t = 1$, I choose to say *that* in particular, over doing or saying whatever else I might have in mind. But, at some earlier point $t = 0$, I decided to invest my attention in finding witty things to say. I try witty phrases on in my mind; I test situations for witty reframings. My attentional policy of ‘looking for witty quips and reframes,’ adopted at $t = 0$, is how I assemble the option set for my choice at $t = 1$.

If we conceive of agents as having these **attentional policies** (APs), how would they be revealed?

They could be revealed directly: to find out if someone has an AP $a$ in context $c$, you can put them in $c$ and see what they attend to. Slightly less directly, you can see what options they find. Or, you can test for detailed knowledge of when exactly it makes sense for them to follow $a$. If they really have $a$, they are likely to know the shape of $c$ well.

1. **Fine-grained.** To guide our attention, APs must be precise. “Be honest” is too vague—it doesn’t tell me what to look for. So, a policy like “be honest” is short for a more specific articulation, like “attend to what I feel about each thing we discuss, and let my feeling show,” or “attend to any false impressions the listener might get from my statements, and head them off with a disclaimer.” To have honesty as a policy, I must first have a substantive interpretation of honesty.

---

5Social Choice and Individual Values. 
6Rational Fools.”
7“Symposium on Amartya Sen’s Philosophy.”
8Smaldino and Richerson, “The Origins of Options.”
9By policy, I mean something like “taking out the trash when it’s almost full,” “calling mom on Sundays,” or “running new contracts past the lawyer”—something done regularly, or in a certain context, without a cost-benefit analysis each time. Attentional policies, then, are policies about how to think about a thing, what to pay attention to in a context, or what to look for in selecting an action. APs are necessary due to bounded rationality and social coordination, just like plans, heuristics, norms, intentions, etc. (Bratman, Intention, Plans, and Practical Reason; Simon, “A Behavioral Model of Rational Choice”; Taylor, The Language Animal) They also offer another account for the rationality of collective action—compared to the prisoner’s dilemma arguments in Anderson, “Symposium on Amartya Sen’s Philosophy” and Sen, “Behaviour and the Concept of Preference”. My approach follows Velleman, Self to Self, ch 11, closely.
2. **Local.** These substantive interpretations differ from person to person. APs might account for much of what’s called “personality”: When making friends, are you cautious or bold? When considering a purchase, is the focus on price, quality, or durability? When speaking, do you try to be witty, precise, or down to earth? Often these aren’t just “character traits” a person is born with, but policies adopted for a reason, which work together for that person’s way of life.10

3. **Hard-boiled.** Talking to colleagues at work, I might like to craft my words to be kind, honest, tactful, humble, and inspiring. I might try to be precise in my speech, aware of how each word lands, aware of my own feelings, and transparent with them. Calm and centered, but also passionate. Physically graceful, like a dancer.

This is impossible, since policies compete for my attention. So, choosing to look for witty things to say means I’m not looking for vulnerable things to say, or helpful things, or mysterious things, etc. That’s, more or less, a choice not to be vulnerable, helpful, or mysterious!11

### Justification Shapes

Attentional policies may ameliorate some problems with biased or incomplete option sets. What about problems due to outside pressure? For instance, consider this AP:

I’m careful with my speech at work. My boss fires anyone who speaks imprecisely.

Consider an idealized agent who operates as above: adopting attentional policies ahead of time, and using them to produce options for most choices. Presume also that this agent engages in something like practical reasoning about which APs to adopt, weighing reasons why one policy might be more beneficial than another in a context. (Here reasons are of the form highlighted by12 Velleman,13 etc. Such an agent can be characterized as having a set of recurrent contexts in their life, for each of which an AP has been adopted, for a set of reasons.

Further posit that in some cases, these reasons build on and depend on each other. For instance, as axioms depend on one another in making a theory, or as in steps coming together to make plan. In other cases, reasons are diffuse and independent.

In the example above, speaking carefully is justified by a chain of hypotheses about the consequences. I must speak carefully, or I’ll say things my boss doesn’t like; then, I’ll get fired. We can visualize these reasons as a chain X⟶Y⟶Z.

Contrast that with these:

I recently opened up to a friend about a struggle of mine. Since then, the relationship feels more intimate, and stronger; it’s easier for me to think about what to say; my friend is unexpectedly helpful. Now, I can’t imagine a good life that doesn’t include being honest with friends.

Here, honesty leads to many benefits, not chained together. We can visualize these reasons (which point in many directions) as a star ⋆.

When reasons for a policy form a chain ⟷, I’ll say it’s narrow-justified, or adopted for narrow-benefits. When the reasons are a star ⋆, I’ll say broad-justified or broad-benefits.

I’ll only call something broad-justified if it’s adopted for bountiful, redundant, and untracked reasons. By bountiful, I mean I haven’t listed them all. I’ve only started listing benefits of honesty and expect to discover new ones. By redundant, I mean I’d continue being honest if any one benefit (such as “my friend is unexpectedly helpful”) turned out mistaken. By untracked I mean that, when being honest, I’m not tracking whether the benefits I’ve named happen in each case. I just focus on being honest.14

---

10 This characterization of an identity as composed of APs shares something with Anderson, “Symposium on Amartya Sen’s Philosophy”; and Velleman, Practical Reflection.

11 At least, at work. I have an idea when to try for wittiness: maybe at work, but not in a fight with my wife.

12 Levy?

13 Practical Reflection.

14 This argument descends from Boyd, “How to Be a Moral Realist.”
In the intro, I mentioned two kinds of “values”\(^\text{15}\): social visions and meaning nuggets. Here’s a social vision:

**Social visions.** Sometimes people mean *visions of what’s right for everyone, or for a group*—what a family should be like, how a father should behave, what a nation should be like, etc. On this definition, values would include things like inclusiveness, freedom, feminine dress-codes, etc.

**Meaning nuggets.** Other times, people mean *things that feel right and meaningful when you do them*—such as being vulnerable, taking stage, being creative, etc.

If you want meaning nuggets, without social visions\(^\text{16}\), select only BAPs.

Andrew believes a pervasive dishonesty is undermining democracy and civil society. For this reason, he tries to spread honesty—denouncing lies, and setting an example of honesty wherever he goes.

Andrew thinks honesty will save democracy, and that he should spread it by being honest himself. This is a chain of hypotheses ⟷. In general, social visions create narrow-justified APs (NAPs), not broad-justified ones (BAPs).

Selecting only BAPs will also filter out other things—what’s done just to keep our jobs, fit in with a friend group, achieve specific goals, or get good sensations—what other philosophers call instrumental goods.\(^\text{17}\)

### Wisdom and Meaning

Questions of welfare are intimately wrapped up in the deeper questions of what it is we should hold as good and honorable in human life. I cannot tackle these questions here, but I want to convey an intuition about why a person’s BAPs might reasonably be worth honoring, at least as much as their preferences, goals, feelings, opinions, and the like.

I’ll capture this intuition in a story about the life-cycle of a BAP.

\(t = 1\). Brenda sips her morning tea, watching a bird on the feeder. Something shifts in her; she sees the bird shares a great project with her. She and the bird are explorers and representatives of what it is to be alive.

\(t = 2\). Over time, thinking of herself as “an explorer of what it is to be alive” becomes a new kind of curiosity for Brenda. It comes up when she “does animal things” (in the woods, overcome by emotion, plunging into cold water). She notices more about her environment, and about how she feels.

\(t = 3\). Brenda’s become more explorative and bolder. Her attention shifts: when in nature, etc, she no longer focuses on being an explorer, but on balancing exploration with other factors, like self-care. Being “an explorer of what it is to be alive” is still, in a sense, something she does. It’s still important to her. But it’s become automatic.

At \(t = 1\), Brenda might say her bird-moment is a

\(^{15}\)Values are often considered as evaluative criteria or attitudes (Chang, "‘All Things Considered’", Velleman, *Practical Reflection*). Here, I treat them as policies, but I think these definitions are interchangeable. An evaluative attitude or criterion can be viewed as something a person does when making an evaluation or choice.

\(^{16}\)Social visions often feel meaningful, but only because they ride on a companion BAP. E.g., my social vision of *spreading values-articulacy* rides on a BAP: responding deeply to the world-situation as I find it. To verify that the meaning comes from the BAP, I can check for vison-but-not-BAP meaningless times, and vice versa.

\(^{17}\)Or: what’s broad-justified constitutes the good life; what’s narrow-justified is done to reach the good life, from a position outside it. This makes some geometric sense, if reasons are steps in a topological space. In any case, I hope broad vs narrow is more defensible than intrinsic vs instrumental.
new idea about something broadly-beneficial. But she doesn’t see how to repeat it, or further explore it.

At $t = 2$, Brenda has a new BAP. She probably doesn’t have a phrase for it, like “being an explorer of what it is to be alive.”

My intuition that BAPs are worth honoring is based on the familiarity of this story. Experiences like Brenda’s are precisely those we call meaningful$^{18}$. And we often feel like those experiences (or their distillation into policies that comes after them) amount to a gain in wisdom$^{19}$.

Broad justifications are harder to communicate—they’re made of many data points, usually collected via experience living a certain way. That’s why “life wisdom” mostly comes from experience. No matter how dog-eared your Kahlil Gibran book is, you haven’t collected all those diffuse benefits.

![Figure 2: This chart shows how we often refer to stories like Brenda’s](image)

---

$^{18}$At $t = 3$, “greeting the world as an explorer of what it is to be alive” has ceased to be meaningful. It may become meaningful again, if she loses her way—for instance, if she gets too busy with work, or loses touch with her curiosity.

$^{19}$Knowledge towards narrow benefit is know-how; well-informed broad justifications are wisdom. I.e., wisdom could be defined as “knowing from experience which policies are broadly beneficial.” Meaning is the frontier of wisdom—the part we still need attention for. Or just: Meaning, the first derivative of a wise, good life.

### Conclusion

That relevance to “meaning” and “wisdom” is one reason to augment preference information with BAP-information in the calculations of social choice and welfare.

Concretely, we can use BAPs to augment or replace preference information, by asking people what they attend to in various contexts$^{20}$, and filtering out the NAPs.

Then, instead of asking if users engage with a product, designers can ask if it makes space for them to be vulnerable, to be “explorers of what it is to be alive,” or whatever their BAPs are.

Although this gets collected individually$^{21}$, it seems likely that BAPs will reduce problems of atomization. They capture what sort of social life individuals find meaningful—information often missing from preferences due to coordination problems. And BAPs contain wisdom, also often lost in preference, about the smooth operation of social life. For instance, a scientist’s BAPs might include intellectual humility, passionate pursuit of the truth, etc. These aren’t just meaningful for scientists, they keep the institution of science on the rails.

Diverse problems stem from engagement-maximizing systems. Depression, media clickbait, isolation of the elderly, obesity, over-consumption, political polarization, bullshit jobs—all stem from a gap between preference and what people find meaningful and wise.

But to close this gap will require further exploration:

1. One next step is to explore BAPs empirically. Are they more or less stable than preferences? How often are they regretted? How independent are they from social pressure, framing effects, or the specifics of survey language?

2. BAPs are not so easily-expressed as preferences. Can interfaces for assisted introspection bring the articulacy requirements of BAPs

---

$^{20}$E.g., what do they attend to while socializing? Is it how much space they give the people around them? Is it saying things carefully, so that their friends aren’t upset?

$^{21}$And thus leave some problems with preference-based measures unsolved—such as agnosticism about distribution, and difficulty with interpersonal comparison.
closer to those of preferences? Can BAPs be inferred from other data? Can such data collection mechanisms be made robust?

3. A final step will be to make social choice mechanisms that augment preferences with BAP-information. For instance, by taxing BAP-negative outcomes, or subsidizing BAP-positive ones.

References


———. “Towards ‘Game B’” https://www.notion.so/humsys/Towards-Game-B-dab1d85ec0f14c7991930c07a0a3b488, 2020.


Laws and auditing systems exist to avoid preference fraud (collusion, cooked books, snake oil, voting fraud, etc). Can BAPs be similarly secured? In Edelman, “How to Operationalize Metaethics.” I argue they can be more secure.