# TEN JUSTIFICATION GAMES



```
—Joe Edelman, Berlin, March 2019
```

*Imagine you have an odd friend—call him Larry. When he's with you, he assesses your feeling-state periodically and keeps a log. He also notes, in the log, everything he does with you. He looks for correlations between positive feelings and things he did with you. See, Larry is a hedonic utilitarian; that is, he believes the right actions are those which make people happy. Unlike most hedonic utilitarians, Larry tries to live directly by his ethics. His way of being helpful is to take exactly those actions he thinks will lead to positive feelings in you, based on the log. For Larry, to believe an action led to positive feelings is to believe it has benefited you.*

*Imagine, furthermore, that a journalist is around—verifying and investigating Larry's log and his claims. She asks you things like "on this day, Larry assessed you as happy. Were you really happy? Happier than the day before?". So there's an external effort to verify Larry's stories about helping you, using the same terms Larry uses. (**Game 1.**)*

I would like to say that Larry and the journalist are playing a **justification game**. This game they are playing is one way to *realize* an ethics—in this case, hedonic

utilitarianism.

I want to highlight two aspects of this: First, Larry and the journalist use a certain *vocabulary* to describe you—in this case, a vocabulary of feeling states. Other such games could use other vocabularies: they could describe you in terms of your preferences, choices, or goals, your search for consistent or true beliefs, or your drives for status, power, or social acceptance.

Second, Larry and the journalist use certain *processes*; at minimum, these include: (1) a way to read you in terms of their vocabulary (e.g., to get information about your feelings, goals, or preferences), (2) a way to verify that related events happened (e.g., that feelings were positive, that goals were achieved, that preferences were satisfied), and (3) a way to turn such readings and events into justifications or claims of benefit (e.g., that feelings were more positive than they would have been otherwise, more goals were achieved, stronger preferences satisfied). Additional processes may also be necessary—for instance, because people don't have their feelings, goals, or preferences all worked out, Larry or the journalist might also need (4) ways to clarify or elaborate their readings on demand (such as requests to introspect about feelings, elaborate or revise goals, reframe preferences).

The vocabulary and the related processes together make a justification game.

Larry's game is, of course, a simplistic way to play hedonic utilitarianism. One could make up a better game for the same ethics. But a game must be chosen: it is only when realized as a justification game that an ethical framework can guide projects, direct choices, or support claims that something is good for people.

Some ethical frameworks might be impractical as games.[1] Others might lead to error-prone games: reading people in a certain way can be systematically off. Verifying events in a certain way can fail. Worse, some games might give Larry room to fudge the stories: He might be able to claim benefit where none was provided (even with the journalist checking).

I will focus here on ways that justification games can be error-prone, or can allow for fudged stories. Here are four ways that Larry's justification game can cause trouble. Two are problems of autonomy, and two are problems of scope.

- First, note that Larry might try to make you happy at times when you don't *want* to be happy, or when you want to make up your own mind whether to be happy. Feelings—and the recognitions that precede them—are something we seem to want to do *for ourselves*. We don't generally want other people to work directly on which feelings we should have, so Larry's game is **intrusive** because it takes your feelings as his goal.[2]
- Second, Larry might side with one part of you, against another. What if pizza makes you happy immediately, and writing makes you happy with some delay? Should Larry order you a pizza or lock you in your writing closet? Larry doesn't know if you prefer to live for the moment (with pizza) or prefer to take the long view (and write). He has no vocabulary for these notions. He can have no sense of what you want *on balance*.[3] If you haven't yet settled the matter for yourself, then—whether it's pizza or the closet—it buries your conflict and delays its resolution, widening your inner conflict. In these cases, Larry's game is **divisive**.

Now, the problems of scope.

- First, imagine that you are upset with Larry because he doesn't want to tell you bad news. He says the bad news you want to hear is unlikely to lead to positive feelings overall. You might admit this is true, but want him to level with you anyways. Feeling good doesn't cover everything important to you. But feeling good is all that counts in Larry's vocabulary. So, Larry can duck certain responsibilities, because his vocabulary and processes don't register the relevant harms. Larry's game **inarticulate** about his effect on you.
- Next, imagine that you tend to feel better on Friday afternoons, but that this has nothing to do with what Larry does. Larry could misread this pattern as supporting various interventions in your life. When Larry sells you a fidget spinner on a random weekday, he always sees a spike between 1 and 6 days later. There's a similar problem if your moods are random—by tuning the parameters of his analysis (such as the assumed delay between an act and the resulting positive feelings) Larry can find support for a hypotheses.[4]

> "False positive rates increase with the degree of analytical flexibility"[5]. Larry's game is prone to what have been called *voodoo correlations*[6] or junk science. Larry's vocabulary doesn't allow him to collect your own testimony about *why* you were happy at any particular time. This means Larry can entertain almost any hypothesis about the causes of your feelings —none of which would appear in your own self-understanding. These hypotheses can be flexibly matched against whatever data in the feeling-state log supports it.[7] And it will be hard for you or the journalist to argue with these conclusions using only the vocabulary of feeling-states and interventions. So, Larry's game allows justifications to be **cooked up**. This particular way to cook up justifications I'll call *cherry-picked correlations*.

These four are problems even if Larry means well. But they are worse if he's hostile, self-interested, or he only wants to seem beneficial and pass the journalist's tests. In this case, Larry can be *cleverly inarticulate* to hide what he takes from you. He can *cook up* justifications for things you don't really want. He can be *strategically divisive*, turning you against the part of yourself which doesn't serve his interest. He can be *unaccountably intrusive* whenever it suits him to take over part of your life.

To use the lingo of computer security, I will call such problems *vulnerabilities* of Larry's game, and that they show how Larry's game can be *exploited*. Games which aren't vulnerable to such, I will call *robust*.

Which justification games are robust? What kinds of thinking (vocabularies) and communicating (processes) can justify projects without causing these problems?

I believe this question has importance for political theory, for ethics, and for practical life. These problems with justification games can tell us a lot about which ethical and political views should be practiced.

In particular, I'll argue that thinking in terms of people's feelings, drives, preferences, beliefs, or goals *isn't* a reliable way to serve them, because even *accurate* information about people's good feelings, expressed drives, satisfied

preferences, improved beliefs, or achieved goals can be used against them while appearing to have benefited them. So common justifications for projects—like *giving people positive feelings* or *making their goals happen*—can be rejected on this basis.

I'll also argue *for* a certain justification game—one I believe is robust. It is centered around people's values: their guiding ideas about how best to live, approach things, relate to other people, and the like[8]. I'll build up a picture of what this means practically: of how we can think and communicate this way.

I have two hopes for this project. First, it may address philosophical questions like these:

- What inside of us is to be honored?
- Where is the source of dignity and meaning in our lives?
- What does it really mean to be good to people?
- What does it mean to benefit a community, or to harm society?
- What makes a good society?

These questions have often been understood abstractly, but they could be reinterpreted as questions about enactable types of thinking and practical processes of investigation. In this case, they are helpfully approached in terms of justification games and how they go wrong.

My other hope is that a theory of robust justification games will help us better evaluate the claims around us.

Leaders in politics or policy are embedded in justification games. A city planner must justify her plans and interventions as benefiting the people of her city. So must a politician justify projects in her district. NGOs and governmental bodies make claims that they provide various social *benefits* (e.g., that they *strengthen community* or *improve lives*), or that they reduce social *harms*(e.g., that they reduce *radicalization* or *polarization* or *bullying*). Similarly, justification games are involved whenever a product or business is supposed to be *good for people*,

*humane*, *user-centered*, or *human-centered*, to help people *thrive* or to experience *wellbeing*.

Many of these businesses, organizations, and governments are using exploitable justification games to make these claims. Identifying which justification games are robust may help us recognize when to trust claims like these, and when not to. It may help us see who can legitimately serve us. Ultimately, this may lead to better politics and better business.

## Ten Justification Games

I will cover these justification games. Only the last is robust to the vulnerabilities named above.

1. Positive Feelings (Larry's Game)
2. Goals Reached
3. Preferences Revealed
4. Advances in Wellbeing or Flourishing
5. Triumph of Our Better Selves over Our Unreasoned Selves
6. Triumph of Our Better Selves over Zero-Sum Games
7. Advances in Knowledge
8. Living by Their Values
9. Living by Their Present and Future Values
10. Living On Their Own Terms

## Goals-Reached

**Game 2.** *Jeff, a billionaire, runs an e-commerce site. He also sells a digital assistant technology. Jeff believes he helps a customer when that customer's goals are accomplished using Jeff's tech (and when there's evidence that it*

*would have been harder or more expensive otherwise). He checks with the customer to make sure he got their goals right, and to clear up doubts about their meaning, he encourages them to describe their goal vividly, as a future state of the world. Naturally, he tries to stay focused on goals his tech empire can help with.*

Jeff's game fits within a family of justification games where benefit means getting people what they desire or prefer; for example, by achieving their goals or satisfying their preferences. One attractive feature of this family of games is the degree of openness it leaves for different people to desire different things—its impartiality.

This family can be broken down in different ways, but I will separate out the justification games that assume desires are *articulable and conscious* from those which drop this assumption. When desires are considered articulable, the games can include processes of *asking about desires* or *recognizing them*, and asking if they've been fulfilled. In such case, what's desired is often called a *goal* or *intent*. But, when are people thought to have inarticulate desires, then the desires must be *detected* or *revealed* through their behavior over time. What's desired then gets called a *taste*, a *like*, or a *revealed preference*.

When desires are articulable, it is usually also imagined that they are relatable (whereas preferences are imagined as idiosyncratic), temporary (whereas preferences are semi-permanent), and precise (whereas preferences describe more general patterns).

So, Jeff's game is focused on articulable desires—which I will call **goals**. Here, what's desired is often understood as a future world-state, or a set of such states where certain conditions hold. Certain social sciences have focused on reading people as goal-havers, including rational choice theory, decision science, and operations research. The processes of software design and industrial design were originally built on these readings—objects (especially computers) have a *function*, which is to help people with goals.[9]

Someone who uses these processes—and understands serving people s addressing their articulable desires—might build something like Google search or an email tool. If the product helps the user achieve a goal (such as "responding to an email") then it must be helping the user.

*Andy, a user with goals, comes to Jeff's e-commerce platform and buys a sketchy health product like weight control pills. Andy's goals include: (1) to buy/have these particular pills; (2) to complete the purchase; (3) to receive the pills quickly; (4) to lose weight more generally ; (5) certain even higher level goals: to be accepted, to feel healthy, to be more attractive to others. Andy would agree he had all of these goals in mind at different times.*

*Belle has conflicting goals: she wants to finishing writing her book chapter, and she also wants to stream great original films at a low price.*

With Andy, we see an immediate problem with game 2: Jeff can cherry pick a goal which fits his purposes. When asked by the journalist, Andy has to admit that he *really had* such a goal, and that what was delivered *matched it*.

This another way to *cook up justifications*, but it's a little different than what we saw with Larry's game. Here, what's being cherry-picked is a certain reading of Andy. When there are many ways of interpreting people, a justifier who wants to seem like he's helping can choose the interpretation that makes him sound the most beneficial. In this case, there's a choice of goals. I'll call this method of cooking up justification *cherry-picked interpretations of people*.

It might seem that there are ways to solve this problem—perhaps Jeff should only count himself as beneficial if he helps Andy with a *fully-informed* goal, or an *ultimate* goal. Or perhaps only goals that are both fully-informed and ultimate should count. On this view, the goal about buying pills should be thrown out as misinformed, while the one about completing the transaction quickly should be

thrown out because it's just a step in a larger goal.

But it is hard to say which of our goals count as fully-informed, and even harder to say which are ultimate. Many of our largest and most weighty choices in life —to marry a certain spouse, or to attend a certain university—seem neither fully-informed nor ultimate, so why should we expect our interactions with an online platform to be so?

One could go even further in this direction and claim that everyone has the same ultimate goal or goals, including perhaps some of Andy's: *to be accepted, to feel healthy, to be more attractive to others*. There are many problems with such an approach, which I'll get to in games 4 and 5. For now, it's enough to note we lose much of the appeal of Jeff's game in taking this approach. Jeff helps people with *whatever* they come for, without making assumptions about what their goals should be. By falling back to Jeff's concept of universal human goals, this advantage is lost.

Yet so long as Jeff stays open to the entire tree of goals that a person has, there is the problem of what, in the introduction, I called *cooked up justifications*—Jeff can pick the stories of benefit that suit his own aims.

And as the interaction with Belle shows, his game also *divisive*.

Let's say Jeff's platform helps Belle with her goal about streaming, rather than her goal about finishing her chapter. Jeff has no way to discover which of these goals is in Belle's truest interest. Both goals could be real and of equal weight. Both could even be ultimate goals of Belle's. And Belle may not have formulated any goal about which to do first. Yet it could still matter to Belle which goal she is helped with.

How could this be?

Belle could have ideas about how she wants to live that aren't about reaching one goal-state or another, but are about how she lives along the way. She might

want to be *dutiful*, and put work first, or to be *free with herself* and watch a film when she feels like it.

It could be that Jeff needs information like this to decide whether he's helped Belle, but he has committed to describing her in terms of her goals. Neither being *dutiful* nor *free with herself* is specified as a goal state.

One way that Jeff might repair his game is to broaden his reading of people to include information beyond goals, and I will look at this possibility in games 8, 9 and 10. But if Jeff's claims are built only from her goals, when he is claiming to help Belle, he is actually helping one part of her *against* another part.

So long as Jeff stays limited to goals, but open to the entire tree of goals a person has, his justification game is divisive and it overstates benefit.

But Jeff is an inventive fellow. Imagine he invents a method of determining which goals both (a) really matter to a person and (b) address the whole of their being (rather than one part against another). Jeff will—from now on—only consider himself to benefiting people if he delivers on these *special* goals.

Game 2 still has problems.

*Carmen uses Jeff's digital assistant technology. It reads her mind. Things appear on it's to-do list as soon as she desires them, if they are possible at all. And as soon as they appear, BAM! The digital assistant rearranges the world into the goal state. So, as soon as she desires to write a book, BAM! It was written. As soon as she desires to spend a day relaxing with someone, BAM! That day just happened. She has the memories implanted in her, but the day is over.*

*Dante also uses Jeff's digital assistant technology. He uses it to set up all of his relationships and collaborations. It solves — forever — the problem of determining which trades and contractual arrangements will work in his*

*personal and work life. Dante* immediately *knows who could hire him, how much they would pay, and what's expected. Same with his personal life: he* immediately *knows where there's mutual desire — for kissing, for sleeping together, for going out to dinner, or for conversing about literature. The digital assistant removes all the pain and confusion of negotiating, of searching, of flirting, etc. The assistant also ensures that Dante's relationships are limited to the contract as specified, incurring no unforeseen costs.*

Tasks and trades are about getting to desired states. These stories suggest that goal-states don't cover everything we want out of life.[10] Or, in the terminology I used earlier, goal-related justification games are *inarticulate* about some harms —because harms in non-goal-related areas go uncounted, and thus net-harms can be cast as benefits.

What exactly remains uncounted, here?

A list has been made of what Carmen wants to *happen*:

- write a book
- day with lover

But a second list has been omitted, a list of *how* she wants these things to happen:

- write a book—thoughtfully, cleverly, with an eye towards impacting the world of ideas, and with great personal focus
- day with lover—cultivating intimacy by balancing patience and impulsiveness, empathy and charm

It seems clear that if the book-writing or day-spending has been automated, there is no way for Carmen to live according to the second list. And it might be that her ambition is not just to get something done, but to do it *herself,* and to do it *in a certain way*.

This brings us back to something mentioned earlier: Carmen has ideas about how she wants to live—ideas that aren't about reaching one goal-state or another, but about how she lives along the way. The vocabulary of Jeff's game can't express[11] those ideas or check if they worked out, and that's where the blind spots are. Rather than getting these things done in the way she likes to, the focus on goal-states will lead to patterns of action that are goal-focused: patterns like productivity and efficiency will supplant the patterns Carmen would have specified—like cultivating intimacy and thoughtfulness.[12]

And sometimes this is more than just ducking responsibility—to the degree that Carmen's goals are venues for her to go through a particular process or to face a particular challenge, when the digital assistant auto-accomplishes a goal, it's not just leaving something out. It's also taking away her venue for that challenge or process. This is what, in the intro, I called *intrusiveness*.

We should be sure that—whatever we do to help people—we aren't taking away something they'd really like to do themselves. We don't want to automate away the important part of their life. Getting the relevant information from them and acting on it should leave the important part up to them, retaining their sense of self and purpose.

## Preferences-Revealed

Let's turn to inarticulable desires—which I will call **preferences**.

What are preferences? Unlike goals, these are often understood as idiosyncratic[13] and permanent. Social sciences with this reading of people include microeconomics, welfare economics (especially in notions like Pareto optimality) and certain fields in computer science—notably recommender systems and collaborative filtering. A person is said to have a *preference profile* (also sometimes called a *utility function*[14]) and while this may differ from person

to person[15], it stays similar for one person over time. Indeed, recommender services like Facebook's News Feed, YouTube, and Netflix build such a reading of a person, based on patterns in consumption choices, and this is used in stories of benefit. If their product gives the user something they have a taste for (such as "seeing photos of my friends") then the product must be helping the user.

The idea of serving diverse preferences might be appealing—this seems *even more impartial* than serving diverse goals: people are subject to social pressure with regard to what they say is a goal. People claim goals which, when it comes down to it, they wouldn't actually want to pursue. There are false goals due to wishful thinking, or the desire to fit in. But choices speak louder than words. If stated goals aren't to be trusted, we can look to *revealed preferences*.

**Game 3.** *Moses is a city planner, and Mark runs a social network. Both play a justification game around revealed preferences. If citizens or users are doing one thing rather than another when they seem to have both options—for instance, if they are driving cars instead of taking public transit—then it is considered beneficial to provide more of what they are choosing. Moses says "people prefer driving" in his city and Mark says "people love to scroll" on his app.*[16]

There is a great advantage in this game: it is data-driven, dispensing with the nonsense of people's reported or verified goals. It soars above the crowd, viewing emergent behavior with a detached and objective eye.

But, do people really love to scroll? Are there people who wake up every morning and consider driving or taking public transit, all else equal, and choose driving (perhaps for the quiet, isolated experience)? Or is it rather that they prefer a job at one place (accessible only by roads) to another (accessible by transit)? If the latter, their choice doesn't count as a vote in favor of the transit planning, but rather in favor of certain jobs. The driving could even be a drawback of these better jobs.

To settle this matter, we should like to ask these people *why* they are scrolling, and *why* they took the car and not the train. But, in our move towards greater objectivity and hard data, we seem to have cut off exactly this option. We have decided not to trust what people say, and this leaves us clueless about what preference is actually being expressed in their choice, and thus about who should get credit for benefitting them.

Without being able to ask why, any particular choice can be read as expressing myriad preferences. Mark and Moses can cherry-pick the preferences that make them seem beneficial. By setting the same choices in different contexts—imagining their users or citizens as choosing between different hypothetical menus—they can make up "revealed preferences" that fit their need for justification. This is similar to what we saw with goals: they are cooking up justifications by cherry-picking interpretations of people.

But there's also a second way that a justification based on preference can be cooked up: by altering the landscape to structure people's choices so it looks like they're choosing what you provide. In this case, a highway builder could make behind-the-scenes deals to ensuring the jobs are in places inaccessible to transit.

*Emily, a user with preferences, uses Mark's app in a way that might reveal a preference for procrastination and social isolation. She also seems to prefer scrolling to commenting within the app. The app gives her choices between scrolling and commenting, and she takes scrolling.*

Does Emily prefer scrolling in general? Is she more fulfilled, now that she has the app and can scroll in it? Or is scrolling perhaps a cost she is willing to pay for—let's say—the faint possibility she will find a human connection on the app. If the latter, it would be fairer to say that Emily prefers even low-probability human connection to commenting on random posts. But game 3 will not reveal this preference, instead, it provides spurious support for the false idea that she prefers scrolling—and thus that increases in scrolling might benefit her. A problem of *cooked up justifications*, via both a cherry-picked interpretation of

Emily, and via *a broader manipulation of her environment*.

## Advances in Wellbeing or Flourishing

Larry's game, in the introduction, was just one example of a family of justification games which imagine that there are good and bad physical states of a person. In the simplest versions, pleasure is considered better than suffering; the more complex ones aim at higher pleasures, or happiness, or wellbeing, eudaemonia, or flourishing.

It seems to me that—even on the most charitable reading—one can't turn these into robust justification games. And that this is true whether you imagine that people can self-report their wellbeing/eudaemonia level, or whether, alternatively, you try to calculate a wellbeing/eudaemonia level for everyone according to some universal recipe.

To be charitable, let us begin by dispensing with the analytical limits of Larry's game.

**Game 4.** *Sergei is not limited to recording your emotion/wellbeing/eudaemonia state and inferring what to do from it; rather, he can ask any questions he likes, gather any other data, and build sophisticated models of your interests, dreams, and desires. He even has a magically accurate way to predict which interventions will result in greater lifetime levels of wellbeing/eudaemonia. These are the interventions he counts as beneficial.*

If these levels of wellbeing/eudaemonia are self-reported, one problem Sergei faces is that many people don't seem to be going for wellbeing, as they themselves would define it.

*Francesca doesn't seem to be going for wellbeing. She was born with family*

*money but put herself through a grueling grad school regimen and the stress of a PhD thesis. She gave birth to two children and cared for them, forgoing sleep, sex, and many other pleasures in their early years. After this she wrote book after book, and took in her ailing father to care for at home, refusing to hire a nurse or to send him away to a home. Francesca says that there's more to life than her own wellbeing. She believes she, personally, could have been happier without the PhD, the second child, the books, or the ailing father, but that her life would be less meaningful or would involve less of what mattered to her to do.*

Francesca wouldn't choose—for herself—to maximize her lifetime (or momentary)[17] wellbeing. She is pursuing things that mean more to her than wellbeing. Many would agree with Francesca: people like Rosa Parks, Malcolm X, Nikola Tesla, and Mother Teresa seem to have chosen something besides their own wellbeing or eudaemonia. One can imagine Sergei intervening with Rosa Parks: "Rosa, instead of making all this fuss, why not enroll in a relaxing yoga class and take a bath?". Sergei could support this with a variety of neuroimaging and survey-based wellbeing studies, showing that baths are better for wellbeing than civil rights activism. Rosa might agree with Sergei's data but still choose activism over the bathtub.

Once again we return to a theme: Rosa has ideas about how she wants to live that aren't about wellbeing, eudaemonia, or happiness. So long as these are ignored, the game is *insensitive* to harms. Only a justification game which supports those ideas could avoid harming Rosa or Francesca.

But what if we don't accept Rosa's or Francesca's subjective notion of wellbeing or eudaemonia, and instead insist on a definition which includes their activities in some universal recipe of wellbeing. Within this universal recipe, civil rights activism and stressful PhD writing are *part* of a life of wellbeing. They enhance it. Perhaps we could even build such a universal recipe using results from positive or emotional psychology, wellbeing studies, a theory of universal human values[18], or neuroimaging.

Even if we grant the scientific plausibility of such a thing (which I don't), there are three issues which might lead us to doubt whether it can be worked into a justification game.

First, note is that this universal recipe would be of immense political and economic import. Certain parties would be very interesting in whether it can be nudged towards more/less political activism, more/less material consumption, and so on.

Second, this recipe as understood would be making strong statements about Francesca which she must take on faith and which don't match her own sense of things: in particular, it tells her she's a person who can *only* reach wellbeing by going to grad school, having children, and so on. If today it tells her that the grueling part of grad school was the best wellbeing can get for her, tomorrow maybe it'll say her best life involves drug-addiction, torture, and a pet rabbit. We have already declared that Francesca doesn't get to decide what wellbeing means. So who is qualified to check the recipe? Unless it is God-given, this game allows for *cooked up justfications* via its unspecified implementation details.

Finally, it would be reasonable to think that this recipe, so as to understand what would bring a person wellbeing, might need information about their goals, their preferences, their feelings, their social acceptance, their values, and so on. If this is the case, wellbeing amounts choosing "all of the above" regarding what vocabulary and processes to use in reading people. This means that a justification game about universal-recipe wellbeing would inherit many problems from the other games.

Games 8 and 9 will show a different way of getting at the same idea of comprehensive view, without relying on a universal recipe and without demanding that everyone value their own wellbeing above all.

# Triumph over Our Unreasoned Selves

I now turn to another family of justification games: those which attempt to lay out what is justified and beneficial by contrast with what is not. In some cases, this contrast is black and white: one part of human life is read as a vast and meaningless noise, a futile struggle, or an empty performance. To benefit people is to help them rise above this. In other cases, there are held to be various levels or developmental stages, and benefit means advancing along these stages.

In service of these views, the mind may be read as having a "lower layer" that is animalistic or socially performative, plus an "upper layer" that is rational or that, at least, can be engaged in a noble endeavor. Understandings of this "lower layer" vary: is it driven by evolutionary incentives towards status-seeking or tribalism? By power or dominance games? By social performances more generally? Is it composed of mysterious psychological forces—like the id, the ego? Are we addicted to our own brain chemicals (e.g., "dopamine hits")? Are we just "following incentives"?

By painting some background like this, the justification games in this category get built around the triumph of some "upper layer"[19] over these distractions, addictions, irrational acts, or empty performances. Support for reading people in these terms is drawn from game theory, evolutionary psychology[20], ethology, behavioral[21] and hidden-motives economics[22] or the sociology of social performances, roles, and power games[23]. One appeal of these views is that they make sense — better than some of the previous views — of our inner conflicts, or why we sometimes chose things we regret.[24]

If such games are to be made robust, two things must be possible: First, there must be a procedure for helping someone advance or to grow *on their own* terms, without violating their autonomy or abetting inner conflict. Second, it must be established that the supposed division between upper and lower levels or

stages *really does separate* what's meaningful and of benefit from what's not, and that growth along these axes (or triumph of our better selves) encompasses everything that's important to people. Otherwise, the game will either be under- or over-scoped.

**Game 5.** *Eliezer tries to recognize when people are being rational and when they aren't, with the hope he can make people less manipulable through their irrational behavior—including their addictions, instincts which are no longer adaptive, and so on. In some cases he will ask people to reevaluate their irrational behavior; in others, he'll just make it count less—reducing its role in the economy, in which media goes viral, etc. In his terms, he'll dampen the part of the economy which "responds to people's brainstem" rather than to their contemplative sense of what's best for them. Success in this, he considers benefit.*

Is Eliezer's game—which holds the contemplative or rational above the instinctual—really in the right about what is meaningful? Or about what is a valuable way to live?

*Giorgio has two selves, one self believes in extensive calculation, and he uses a variety of spreadsheets to plan his dates, vacations, meals, charitable donations, and career path. But his other self believes in instinctual living and occasionally he forgets about what he planned in the spreadsheet and just makes an offhand joke at a date, reveals something personal on a whim, or flies to Mallorca last minute because he is swept over by a desire to go dancing on the beach.*

Now, if you believe that my description of Giorgio's "selves"[25] is relevant to Eliezer's game, which self do you think Eliezer should help? The answer may depend on whether you think extensive calculation or instinctual living is better. If so, this is a pretty big value judgement to make.

But perhaps the example of Giorgio seems uncharitable. When Giorgio flies to Mallorca, swept by a desire to go dancing on the beach — doesn't this count as a perfectly good reason? Isn't this perfectly rational? And if so, what separates this

from so-called “irrational behavior”?

Now, you might respond in two ways: you might say that irrational behavior is done without reasons—because of an instinct or drive or a raw desire. Or you might say that irrational behavior is done because of *bad* reasons.

Recent work in the philosophy of action[26] doesn’t leave much room for the first idea—that we act without reason. It seems we aren’t comfortable acting when we don’t know our reasons. Even if we imagine we are probably in the right—we stop and try to figure out what we are up to. Here’s David Velleman:

> *You are walking up Fifth Avenue. All of a sudden you realize that you don't know what you're doing. You can see that you're walking up Fifth Avenue, of course: the surroundings are quite familiar. But the reason why you’re walking up Fifth Avenue escapes you, and so you still don't know what you're doing. Are you walking home from work? Trying to catch a downtown bus? Just taking a stroll? You stop to think.*[27]

Or, as Daniel Hausman put it:

> *My awareness of a desire to do X does not automatically incline me to do X intentionally, unless I can see some reason to do X. If I see no reason to do X, I will try to suppress my desire to do it. A fervent desire to eat mouse droppings sends people to a therapist rather than to their mouse-infested basement cupboards.*[28]

We are quite alarmed when something doesn't have the blessing of our rationality. Here's another example from Velleman:

> *Imagine that your arm becomes temporarily paralyzed. When you wake up each morning, the first thing you do is to check whether you have regained control of your arm. What exactly are you hoping to find?*
>
> *Part of what you're hoping to find, no doubt, is that your arm moves. But movement by itself wouldn't be enough. Waking up to find your arm flapping around aimlessly wouldn't lead you to think that your control over it had been restored. You'd have to conclude instead that paralysis had given way to a spasm.*
>
> *What you're hoping to find, then, is that your arm not only moves but moves when and where you want it to. But would movement in response to your desires be enough? You might of course be encouraged if you found your hand scratching an itch behind your ear; but if you subsequently found it grabbing food off someone else's plate, you wouldn't necessarily be re-assured by the reflection that you had indeed wanted what he was eating.*[29]

Overall, it seems we are hesitant to act without reason.[30]

Let's turn, then, to the other possibility—that irrational acts are those done for bad reasons. We might take take a few paradigmatic examples: the act of checking Facebook when you should be working, or of eating a piece of chocolate cake that breaks our diet, or even the act of returning to a drug

addiction.

If we admit that there are reasons behind these acts, what could they be? Perhaps we check facebook because we hope that a break before work will recharge us, or help us feel less lonely as we proceed to work. Perhaps the cake feels like a way to enjoy life or to have a brief sensory delight amidst a day or drudgery and disconnection. Perhaps returning to the drug is the only way we see how to make a tolerable day of it.

While these may be bad choices, or conflicted ones, it is hard to say they are made for bad *reasons*. It might be more correct to say they were made for short-sighted reasons, or for reasons that take into account some factors but not other factors. Yet isn't this the case with all of our choices?

On balance, it seems like this rational / irrational line is a hard one for Eliezer's game to draw. I believe this is good news: psychological theories that paint people as irrational tend to imply there's nothing noble that they hope for, nothing their actions are about. When people aren't thought to have reasons or values of their own, there's nothing honorable in them to listen to. This makes it hard to build a robust process or vocabulary so as to know when you are serving them.

A robust justification game may have to look into the details of people's reasons, rather than draw a contrast between rational and irrational choices, or between good and bad reasons. Games 8 and 9 are like this.

## Triumph over Zero-Sum Games

Another member of this contrastive family avoids speculating about human psychology. It doesn't connect blame individual irrationality for the "lower level"—the vast and meaningless noise, the futile struggle, or the empty

performance. Instead, it places blame on group processes, status-strivings and signaling, or social performance—each of which may be individually-rational, but which are held to be meaningless on-balance.

**Game 6.** *Scott admires Eliezer, but has a different theory. Scott presumes that people do in fact have good reasons for even ultimately meaningless acts. The problem is that the game theory of social performance leads to zero-sum status and tribal games, which crowd out the important work of society. By damping down on these zero-sum games, Scott believes he can benefit everyone in getting back to the meat and meaning of life.*

Unfortunately for Scott, drawing lines around zero-sum games seems no easier than around irrational behavior. I will start with the easier case: games of status and signaling, and then see if I can extend it to games of power.

*Hector has evolved to play status games. These feel meaningful to him, but are actually meaningless. At work he is a chef in a traditional French kitchen, where he is currently the sauté chef and must obey the sous-chef (whom he hopes someday to replace). In the evening he is a tango dancer, acutely aware of the better dancers and the male-female dynamics wherein the best or most attractive dancers manage the room. Even his solitary hobbies—arranging flowers in his study, writing in his journal, and so on—are calculated to win an eternal struggle to rise in the ranks.*

*Isaac has evolved to play social games more generally, which feel meaningful to him but are meaningless on-balance. He is trying to fit in and to enact correctly various social performances. He has learned how to be a good man, a good employee, and a good community member—how to play certain roles in certain scenes. The ideas which guide him—ideas like productivity, creativity, success, social justice, responsibility, empathy, or ending oppression—are performative. He acts in certain ways to seem productive, or to push others towards productivity. He acts in certain ways to seem masculine, and so on. When he is being courageous, it is a performance so he can fit into a particular crowd*

*which demands courage, etc.*

Are status-driven environments—like Hector's traditional french kitchen or tango milonga—*only* venues for status-games? It seems not—people are also there to dance and to make great food. The interpretation of Hector as status-driven seems to be a bit wonky. It seems more likely that he loves to dance, loves to cook, and that his hobby of arranging flowers is not an elaborate status game—in short, that Hector is a creative guy—with values like self-expression, living an embodied life, and celebration.

Furthermore, it appears certain considerations inform what it is to have status in different places. There are different kinds of strengths of character which are held as important—say in a good husband, father, or wife—in different environments: In Moscow, perhaps a good husband needs to be ruthless. Whereas in Berlin, in the same period, a good husband might be artsy and kind.

These attributes are not *entirely* matters of performance. They're also about personal values—the ways of living that people have worked out for themselves, and which they believe in *on their own terms*.

We can see this by noticing the difference between what it is to *seem ruthless to others*, and what it would be to be *ruthless on your own terms*, without regard to how you come off.


When we focus on social performance or status games, we focus on *scenes*: on

royal courts, on teen subcultures, *etc.* But scenes—even these scenes—are also venues for collecting around, sharing, and exploring values.

This explains why—for most people—pure social performance (doing things "just for show") is somewhat uncomfortable. Sure, it's quite a job to fit in and to play social games. In order to be things like a good employee or a good man, it's often considered that you need a job, a spouse, a means of transport, certain outfits, and so on. You must learn a huge number of unwritten social rules and to comply with them improvisationally. But when we imagine a person for whom social performance (such as acting polite, responding promptly, acting cool) has become their *only* concern, we feel that something has gone wrong. This person is missing some of the purpose of life. But these cases are rare.

I believe this remains true, even in games of power.

*Jen works to set the agenda in various aspects of her life, and to establish new norms in her relationships, associations, and in broader society. She is aggressively redefining social space to serve her vision for it. She hopes that certain ways of acting will spread, leading to social changes she favors or that will put a certain political group on top. She tries to be honest with her partner mainly because she wants them to be honest back to her. She uses vocabulary she hopes will spread and redefine discourse.*

The most plausible reading of Jen is that she is up to something besides power games. If we imagine that norm creation has become her *only* concern—we feel that something has gone wrong. She's missing some of the purpose of life.

Whether with Hector, Isaac, and Jen—we find meaningful activity where we hoped to find a "silly show" or "mindless struggle". Yet something remains unexplained.

When introducing Game 5, I mentioned that these contrastive justification games—with their upper and lower levels or developmental stages—get some of their

appeal because they speak to our inner conflicts and regrets.

It is indeed a struggle to know what we want in life, to know how we believe in living, or what approach to take in each environment. We sometimes *do* regret our social performances and status games. But as I'll show when we arrive at game 10, I think this is better explained as a struggle to find our best personal values, and to live by them.

## Advances in Knowledge

**Game 7.** *Neil runs a television show and media empire with the mission of improving the accuracy of people's beliefs about the world. His fans track the belief-improvements they've achieved, with the idea that being "less wrong" will improve their personal lives and advance society because they can advocate for evidence-based policy and so on.*

Neil's approach is appealing for many reasons.

On one perspective, a person cannot even have the right goals or preferences or feelings if they don't have a background set of beliefs that made those goals seem possible, those preferences advisable, those feelings warranted, *etc.* Goals, preferences, or feelings based on wrong beliefs are just mistakes. On this view, satisfying the goals or preferences people have *now* would be foolish, as it makes sense to correct their mistakes first. So, gains in clarity or knowledge can be seen as more important than anything else that can happen to a person.

Another appeal of Neil's approach is that gains in *collective knowledge* (often referred to as Progress or Science) seem responsible for massive gains in human welfare. Further such gains may present the best path to help more and more humans deliver on their *goals*, *preferences*, or *feelings*, whatever they might be. Knowledge (including knowledge of how to better organize social systems)

might be the limiting factor in the full expression of *whatever* is important for human beings.[31]

And there's a third appeal—one directly relevant to our concerns with justification games. We already have practices for deliberating together, improving the quality of our beliefs. We realize when our beliefs were improved and can give credit to particular sources. And generally we can endorse such a gain in clarity or knowledge with our whole self. Previous justification games struggled with issues of autonomy and inner conflict. A knowledge-based justification game can use processes like deliberation and rational argument (rather than persuasion and brainwashing) to benefit people in an autonomy-preserving way. Through such processes, a person can accept an epistemic gain on their own terms, and credit it freely.

So, can we make a robust justification game based on self-attested advances in knowledge? I don't think so. Not, at least, the way knowledge is usually understood.

To gain in knowledge usually means to find clearer or more accurate beliefs about the world, or more powerful or grounded theories. So, the appeals above suggest a great human importance for *theories and details about how the world is*. Yet, if this is what we mean by an advance in knowledge, and if knowledge has the central role supposed here, we should be willing to trade many things for an advance in knowledge.

*Kate is an astronomer and an orthodox Jew. She spends Saturdays with her religious community, where the conversations focuses on practices and questions of how to live, and rarely turn to theories about the world.*[32] *People are always talking about how to do Yom Kippur right, and never talking about where God resides in the universe. The rest of the week, Kate is in the lab.*

*Neil has a plan to upgrade Kate's powers of perception and recognition. Like Sherlock Holmes, she'll be better at noticing things: the composition of her*

*environment, patterns in the stars, clues about others' emotional states, the realities of her political and economic situation, etc. The only price is a corresponding reduction of her powers of* ***appreciation****. She'll know more about what's going on, but less about what excites her. She'll have trouble seeing why to care. Whatever she currently values — whether for its beauty, its usefulness, or its passion — she'll appreciate less. This will advance Neil's mission—of improving the accuracy of people's beliefs about the world—and it amounts to benefiting Kate in his terms.*

Should Kate take the trade? Would it make for a better life? Would it make her a better scientist? Would the world be a better place if all of humanity was more perceptive, more factual, and less appreciative?[33]

And why, when given the chance to debate Yom Kippur or cosmology, does Kate's religious community stick to Yom Kippur, trading away their chance to discuss the sort of factual or theoretical knowledge that we have said helps us have better goals and preferences and to progress as a civilization? If theories and facts about the world are so vital to human life, why aren't more people[34] concerned with them?

## Knowledge about Values

These puzzles resolve if we widen our understanding of what knowledge is. Here are two stories about a peculiar kind of learning. They are about a kind of knowledge that doesn't seem to fit the definitions above.

1. *Imagine you ride your bike regularly, but you usually get lost in your thoughts while you do so. One day you try to attend to the wind, to how it feels on your skin, etc. You like it. You decide that this is how you want to do bike rides. Not for any payoff -- it just seems to be a better way to ride your bike.*
2. *Imagine you have something to get off your chest, and you decide to try to*

> *be honest with one of your buddies for this reason. It is an experiment, and you have this goal to get whatever it is off your chest. But along the way, in being honest with your buddy, you discover a lot of other advantages. It feels good to be honest. The relationship feels closer to the kind you really believe in. After a while you realize that your honesty is no longer a goal-driven tactic or an experiment, it's just how you try to be. It seems to be the best way to live.*

To give a name to what you are learning in the two stories above, I will say they are about learning *values*. In the first story, you learned the value of *being sensual while biking*. In the second, of *being honest with your friend*.

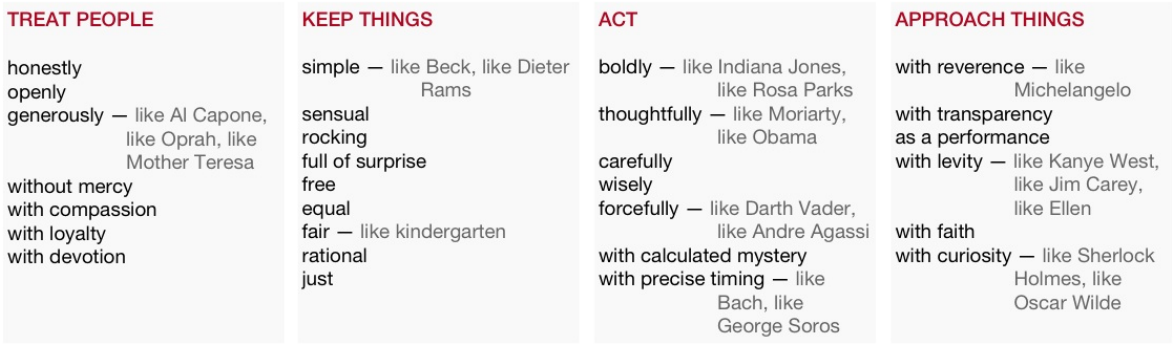
From one angle you can think of this as gaining a belief. You could say that in each story above, you've updated your set of beliefs about what makes up a good life for you. So you can think of a value as a belief about *what makes up a good life*.

But what I mean by *values* doesn't cover just any belief about the good life. I want to reserve this word to mean *ideas about where best to put your attention in certain contexts* that are, furthermore, *unstrategic*.[35] I'll try to clarify what I mean, and then show how this kind of learning addresses the puzzles of the previous section.

There is a sense in which *ideas about where to put one's attention* have—among ideas—a special role in human life. Even if you are pursuing a goal or making one, you may be doing it because you have an idea that goal-making or goal-pursuit is a good place to put your attention. Some actions—like burping—may not be mediated by ideas at all. But to the extent that that ideas steer us, the *ideas about where to put attention* are some of the most important.

These ideas guide us even when we don't have goals. A person may not have any particular goal in mind when they're chatting with their best friend, they just let the conversation drift.[36]

But even without a goal, they may still have ideas about how they want to *be* with their friend. They might want to be *honest*, or *real*, they might want to *keep things light*, and so on. And as contexts for being honest or real or keeping things light arise, these ideas come to guide their attention for a moment, as they improvise the conversational flow.[37]




When you learn a new such idea, about how to live, it changes how you live. So, these ideas are more like preferences or goals than other beliefs. They say something about how you want your life to go, and about how you are steering it.

When I speak of values, I mean a subset of these ideas. I mean the ones that are *unstrategic*.

Some ideas we have about how to go about things are strategic. We should look for opportunities to pass the other cars, but only because we want to get there by 8pm. We should act like a gentleman, but only because we want to impress our date.

By the definition here, these aren't values. Values include only such ideas which have decided to use outside of achieving a particular outcome. In the second story, *being honest with your friend* only becomes a value when it ceases to be merely strategic and gets incorporated in your general vision.

So, to sum up: a value is a belief about the good life. More specifically, it is an unstrategic idea about where best to put your attention. It is a belief that the good

life unfolds through putting your attention on these things. As such, it's a kind of knowledge about *what works in general*—a rough rule about what to pay attention to, when.[38]

Values include attention-guiding ideas about how to treat people (*honestly*, *openly*, *generously*, *without mercy*); how to act more generally (*boldly*, *thoughtfully*, *carefully*); how to approach things (*with reverence*, *with levity*, *with skepticism*); and how to keep things (*simple*, *sensual*, *rocking*, *full of surprise*).

I'd like to claim that *this* kind of learning is what was missing from Neil's approach in Game 7.

We can roughly understand the word *perception* to cover the apprehension of facts, theories, or patterns in the world. And we can understand *appreciation* to cover the apprehension of values. Whereas a perception forms a new idea about the universe, an appreciation forms a new idea about what's important, exciting, or good. About what to attend to.[39]

It makes sense that Kate can't do her work without this kind of "belief" about what's worth attending to, or what's exciting. A "belief" like this has to precede any other scientific work she does, otherwise she wouldn't be interested in doing it!

When Kate is at the lab, recording facts or data points, this is foremost an expression of what Kate finds important and interesting. Kate finds the stars and planets interesting and wonderful, and so she observes them. She finds the arcs traced by the planets poetic, and sees the possibility of charting their paths as a way to participate in the great cosmic game. Kate collects one kind of data over another for a reason, and those reasons come from her values and appreciations.

In general, perceptions happen when someone thinks something is important to look at.[40] And statements are made when someone thinks they are important to

say. Without knowing what to look at, we don't make observations. Without knowing what we care about, we'll never know how to frame a fact. Science can only detect laws or patterns amongst phenomena that have been recognized as important.[41] Thus, the progress of science is mostly about changes in what we find important to look at. The facts unfold before us once some values have made them important.

Kate is guided in doing good science—not because she tries to get her facts right—but because she is a *valuer*. What allows her to advance the field involves being guided by her values and being skilled at discovering new values.

Similarly, if we extend our definition of knowledge to include learning values—learning about how to live well, or about what's important or exciting or good, or what to attend to—this also sheds light on why Kate's community prefers to argue about how to do Yom Kippur right, rather than cosmology. They were talking about what's important, or good, or what to attend to during Yom Kippur.

Humans seem very concerned with this kind of knowledge—we constantly ask one another what's important (whether in a spouse, a wine, or a programming language), or what to attend to in this or that situation. We are concerned with getting values right, more than with getting facts right, because we want to direct our lives well.

So, building a justification game around getting-facts-right would be shooting ourselves in the foot. It would be *inarticulate* about our appreciations and values and give too much weight to learnings which don't address our lives.

But let's see what a justification game about values would look like!

## Living by Their Values

**Game 8.** *Charles and Amartya don't care about your feelings or preferences about the world. But they are acutely interested in how it's important to you to live, and what you believe is worth paying attention to in which contexts. In other words, they gather information about your values—the guiding principles you use to decide what to attend to and how to act in different circumstances. They know—for each of your relationships and daily contexts—the top values that guide your choices. Were you trying to be dutiful today? Were you trying to be bold? Are you trying to be more intimate with your friends? And so on. They want to make sure these are really action guiding—so they ask whether there were particular moments during the day where this was really how you wanted to live. They also ask some followup questions: do you have a good environment in which to live by this value? How is it going? And they justify projects by whether they give you an environment—a place to live by your value and furthermore—when they help provide such an environment—if living by that value goes well in it.*

This game avoids some autonomy problems from the earlier games. Recall from game 2 that when someone helps with a goal, this can take away the part we wanted to do ourselves. It is similar when someone gives us good feelings. We called this *intrusiveness*. But if someone helped you be honest, or courageous, or bold—this implies they left the important part to you. To say "someone helped me be courageous" means they didn't do it for you. Helping with values isn't intrusive, because values name the part *we* want to play—the things we want to attend to and the ways we want to act.

This game also avoids some problems of *inarticulacy*. It may seem that values are just one kind of aim we have—and that game 8 would be inarticulate about other aims. But, when a person says they have lived by their values, this *rolls up* many other assessments:

- To the extent that a person values *choosing ambitious goals* or *doggedly pursuing them*, then when they assess that they've lived by those values, this includes an assessment that they've chosen and pursued such goals.[42]

- To the extent that a person values *self-care* or *surrounding themselves with comfort and beauty*, then when they assess that they've lived by those values, this includes an assessment that they were able to provide various *good feelings* or *tasteful things* for themselves.
- To the extent that a person values *disinterested curiosity* or *getting to the fact of the matter*, then when they assess that they've lived by those values, this includes an assessment about moving towards *truth*, *clarity*, or *knowledge* as far as they were able.
- And so on!

So, when you tell Charles and Amartya that you've lived by your values, this summarizes many other assessments, each to the degree appropriate. It says you've also had success with goals, tastes, and other aims.

Finally, this game is also robust against three ways to *cook up justifications* that we've seen in other games.

First, with feelings, the same data could be used to support too many hypotheses about benefit. I called this *cherry-picking correlations*.

*Liz wants to be brutal, physical, and centered when she practices jeet-kune-do; she wants to be soft, warm, and tolerant with her family; she wants to be calculating and ruthless with her enemies. Liz thanks Amartya and Charles that she has such a good environment for practicing her brutality and centeredness.*

Liz's values are scoped to the contexts in which she believes in living by them. This establishes a correspondence between a person's values and the events of their life. It is easy to ask about people about the contexts that go with a value, or the values that go with a context. I could ask Liz how she wants to treat any particular person, or act in any particular situation. Her top values in each context are exactly the ones she hopes to keep in mind, or to remember intuitively, so that she manages to live these moments in the way she believe. It's not too hard to name exactly which of our values we failed to live out today, and which we succeeded at. This scoping makes *cherry-picking correlations*

much harder than it would be with feelings or goals.

Another source of cooked up justifications came from too-much flexibility in reading people, rather than in analyzing events. I called it *cherry-picking interpretations of people*.

*Liz also wants to be generous in the manner of Al Capone. She has the words of Shakespeare on her wall: "Those friends thou hast and their adoption tried, grapple them unto thy soul with hoops of steel." Marty also wants to be generous, but in the manner of Oprah. He likes to think he's spreading joy every time he tweets a cute photo or gives away a ticket to one of his DJ gigs.*

While Liz and Moe might each count "generosity" as a key value, neither would want to live by the other's idea of what the word means. Do these two generosities share a more abstract parent in the space of values? If they do, neither Liz nor Marty would claim that abstraction as a personal value. This is different from how it would be with goals, where two people who want to get to different places both want to travel.

Goals sit amidst hierarchies, both of generality and of planning. Any goal can be further broken down (either into steps or by further elaborating its specification) or further generalized (either by looking for a more ultimate goal, or by dropping specificity). This leads to new versions of the goal, which still count as goals of the same person. This hierarchical arrangement of goals is why goals-related games are vulnerable to *cherry-picking interpretations of people*. The justifier can pick a level of generality or planning stage at which the goal suits their own interest.

Values aren't like this. Yes, there are many kinds of courage, honesty, or generosity—but they seem to have definite senses. Only at a certain level of specificity are they important to live by—the level at which the value is useful as a guide to action or attention. So, using a value seems more like using a dictionary word than a goal.[43] It's possible to ask a person which sense they

meant, and their answer has a kind of authority missing from statements of goals.

We saw a third way to cook up justifications with a universal recipe for wellbeing. It has unspecified implementation details, and there was no way for Francesca to verify that what was being justified accords with her values. Game 8, by contrast, is transparent.

All in all, #8 is the best game yet. It's not intrusive and is immune to three ways of cooking up justifications.

Unfortunately, that's not enough.

## Refining Values

*Nina wants to be both tactful and clear with her coworker. Charles and Amartya find a way to help her be clear, but in way that's not tactful at all. Can they rightly consider themselves to have helped her?*

*Oliver wants to be firm with his children. Charles and Amartya facilitate this. Later, Oliver regrets his earlier value. He wishes he'd encouraged his children to be empowered, rebellious, and risk-taking, rather than obedient. He wishes he hadn't accepted Charles and Amartya's help.*

When Nina has conflicting values for the same context, she finds that game 8 can be *divisive*—helping one side of her against another. Similarly, Oliver finds that game 8 is *inarticulate* about something important to him: he doesn't just want to live by his *existing* values; he also wants to find the right values to live by.

To address these problems, it'd be helpful if we had:

- (a) an understanding of how Nina could resolve her conflict, or how Oliver could endorse upgrades to his values;
- (b) a rule about when the helpful thing is to encourage the above, rather than to focus purely on current or conflicted values;
- (c) an understanding of what it means to do the above while respecting Nina's and Oliver's autonomy, letting them decide for themselves which values to endorse, and when to endorse them.

There are three reasons to think this is possible with value-conflicts, where it wouldn't be possible in other areas.

**The first reason**. Recall that values are beliefs about what makes up a good life. As such, they are *debatable*.

*Oliver and Pete like to talk about their values very openly, and they debate them. Oliver says he tries to be honest with his wife, and Pete says he thinks it's better to be very guarded with everyone, to never share anything except strategically. Pete says he is merciless with his enemies, and Oliver says he tries to be charitable with all people, to turn the other cheek. Between Oliver and Pete, no approach is considered morally right or wrong, but some are considered wiser than others. How to approach things best is the topic of many discussions.*

We have the whole structure of argumentation, deliberation, evidence, and so on, which we can use to offer alternative values that can be endorsed by the other party as improvements.[44] Preferences and feelings are not considered debatable like this.

Oliver and Pete—like many people—are curious about how to approach things best, are interested in what others have to say. They believe they can learn something. In general, people want their lives to work out well, and if some of their approaches have errors of thinking in them, or if they don't apply in exactly the situations where they think they do, people have an interest in improving or refining those values—sometimes by learning from others.[45]

**The second reason**. A person has a certain authority to say which values are right *for them*. Values differ, in this, from other beliefs. We think of facts as convergent: measure the same thing over and over again, and you converge on the same data.[46] Values are not convergent in this sense, because the same situation may be approached meaningfully in different ways. So Oliver and Pete might *both* be on to something.

Values are also often developed before they can be explained or justified. The value of *sanding a boat* was developed before theories of fluid dynamics. They are like tools where it is easier to show *that* they work than to explain *why* they work. A person can learn a value through admiration, by meeting someone who lives by a value that seems like an improvement. Or through an intuitive appreciation of the importance of something.

Even when values are intuitive, they are based on personal experimentation and reflection. Because a person's values may diverge from others' values and still be right, and because they can be intuitive, inexplicable, and result from experimentation and reflection which only that one person has done, a person must be considered the authority as to their own values.

**The third reason**. While values might be divergent from person to person, within a person they converge towards a manageable and harmonious set. This happens because as we approach a context (say, a difficult conversation) the values that we want to live by (like Nina's tact and clarity) need to rise to our attention, otherwise there's no way for us to live by them. When too many values apply, or when they conflict, they cannot guide us effectively.

So—unlike preferences, or goals—our values converge towards a set that's realistic, workable, compatible, and integrated.[47]

*Quinn values being likable and fun, but she also values being at ease. At work, she finds her efforts to be likable and fun are actually leading to her feeling tense all the time. This makes her confused about how to act, and embarrassed.*

As we can see with Quinn, a conflict in values is often accompanied by negative feelings. These can signal a conflict between two or more values, or a conflict between our values and reality. And the feelings push us towards a resolution.

*Later, Quinn hits upon the idea of being authentic and caring, and this replaces her former values of being at ease, being liked, or being fun. Her new value seems to be clearer about how she wants to live than her previous values, while also providing clearer guidance at work.*[48]

Quinn might also say she'd remedied an error in thinking—that she used to think relationships were about being liked. In transitioning away from being liked, she repaired a misunderstanding of good relationships. She could drop the old values, precisely because she'd clarified what they'd really meant for her. The importance of the old values was captured in the new, more comprehensive value.

To generalize, when we are faced with a conflict in values, like Nina's conflict or Quinn's, we go in search of a way to reconcile the conflict with a single idea. Sometimes this just means deciding to be clear in one context, and tactful in another. At other times we must ask questions like "What could I aim for that involves being clear *and* tactful at the same time?" or "Do I *really believe* in being tactful, or could it be that it's *always more important* to me to be clear?" And so on.

We reflect; we ask other people about their values; we experiment; we see what works for us. In so doing, we unify our sense of ourselves and of the best way to live.

So, I've claimed that values are debatable, intrapersonally convergent, and interpersonally divergent. And that these traits, together, mean that a person is the authority about which values work for them.

This paints a very different picture of changing or revising values than we have

about other justification games. Changing a person's preferences or feelings must considered a manipulation.[49] The whole concept of serving a person goes away when we can imagine changing their preferences or feelings arbitrarily, because there is no way for the person to endorse those changes as improvements, through experimentation or reflection.

But what if someone is trying to change our values, by telling us there are better approaches than the one we are trying to live by? Since values are beliefs, they don't have to brain hack us to change our values. They can just argue that there's a better way to live. Or better: they can offer us a space to experiment, trying on the other values and seeing how it goes.

Ultimately people will know best which ways of approaching life work best for them, and they can report back whether you've inspired them in a new approach.

## Living by Their Present and Future Values

**Game 9.** *Ruth and David play a justification game much like Charles and Amartya's, but with an upgrade. Unlike Charles and Amartya, they do care about your feelings—especially about feelings of conflict. They even care about feelings of conflict they think you might have in the future. But they don't act to minimize those feelings, as Larry would. Instead, they use them as signals to offer you opportunities to debate your values, or to experiment. They limit themselves to doing this in a way that lets you reach your own conclusions about how you want to live. They count themselves to have benefitted you either when they help you find better values, or when they help you live by values you remain unconflicted about.*

Game 8 could be divisive. Game 9 is neither intrusive nor divisive. It respects autonomy, as far as has been understood in this essay.

I also believe it is robust to attack via *clever inarticulacy*. Since our values *roll up* all of our other aims, our values at any one time capture what's important to us then. To this, Game 9 also tries to capture—as best as can be done—what might become important to us in the future. The vocabulary and processes here capture everything important to human beings.

Unfortunately, there is a fourth way to cook up justifications, which I mentioned in the section on preferences: *via a broader manipulation of the environment*.

## Living On Their Own Terms

*Ruth and David build—as a side project—a social network for entrepreneurs. On this network, certain norms and expectations arise—it becomes very important to have #goals, to be #killingit, and to be #professional. Many people adopt these "values" and this lets Ruth and David take credit for selling them professional clothing, productivity-oriented software, and other things which help their social network users live by these "values". But their users wouldn't have had these values at all if it weren't for Ruth and David's social engineering.*

This is a more expensive way to cook up justifications than we've covered previously. But attacks like this do happen.[50] Is it possible to make a justification game that's robust against them?

At some level, I believe the answer is no. Values only make sense within a context. A person is more likely to hit upon the value of being adventurous in a culture in which entrepreneurship, rock climbing, and so on are viable activities. If Ruth and David have the capacity to reengineer culture so these things are no longer good ideas for anyone —if they can alter the entire landscape to make certain ways of living impossible—their exploit will be successful.

And yet, I believe we can do better than game 9. We can harden it against smaller-scale attacks like the social network above. They key here is that, at least at first, the incentives on the social network are not exactly to *be* professional (or goal-driven, or killing it)—but rather *to seem to be that way*, to others on the network. They are social performances.

I brought up this distinction in the discussion of game 6.

> *These attributes are not entirely matters of performance. They're also about personal values—the ways of living that people have worked out for themselves, and which they believe in on their own terms.*

> *We can see this by noticing the difference between what it is to seem ruthless to others, and what it would be to be ruthless on your own terms, without regard to how you come off.*

The desire to maintain this distinction is one reason I went out of my way to define values as *unstrategic*. Social performances are strategic. By my definition, then, the problem with the entrepreneurship network is that it injects false values into a justification game that's not designed for them.

The way to harden game 9 is to exclude these false values, as best we can. This is difficult, because the users of their social network may not themselves be clear as to the difference.

**Game 10.** *Joe and Tristan play a justification game much like Ruth and David's, but with a clarification. They try to collect information not just about your*

*values, but also about how you act to influence others, and to fit in in different contexts. In other words, they collect information about norms that you are trying to comply with in each context, and norms you are hoping to create. These are used to help you separate the ways you act to influence others from how you want to act on your own terms.*

Game 6 was about trying to identify and write-off social performance as meaningless, but we couldn't figure out how it could work.

We noted meaningful parts of status-games—for instance, when a person is trying to embody a certain kind of *leadership*, and meaningful parts of scenes—as in the formation of certain church groups or worker cooperatives. We saw that certain aspects of social performances like *masculinity* or *femininity* might be meaningful, and others only for show. For these reasons, we couldn't justify excluding status-games or social performance from the meaningful part of life.

Game 10 defines these meaningful parts—as attempts to gather around values that we endorse *on our own terms*. At the end of that section, I wrote:

> *It is indeed a struggle to know what we want in life, to know how we believe in living, or what approach to take in each environment. We sometimes do regret our social performances and status games. But as I'll show when we arrive at game 10, I think this is better explained as a struggle to find our best personal values, and to live by them.*

# Conclusion

If someone were playing game 10 with you, they'd ask certain questions:

- For each of your relationships and daily contexts, what are the top values that guide your choices?
- Were there particular moments during the day where this was really how you wanted to live?
- Do you have a good environment in which to live by this value?
- Are you conflicted about it?
- How is living by this value going?
- How are you acting to influence others, and to fit in in different contexts?
- How do you want to live on your own terms?

My claim in this essay is that we have reason to trust someone who asks *these* questions—and who justifies projects by their answers—above those who try to use other kinds of information.

My claim is limited in some ways. I have discussed four ways that justification games can be exploited, but not that these are the only vulnerabilities which matter. I haven't shown that intrusiveness and divisiveness are the only two autonomy-violations; nor that there are only four ways to cook up justifications.

A full theory of robust justification games may need to go further and enumerate all possible attacks using some kind of framework. And a full defense of game 10 may need to go into the role of values in choice and action, or into how they can be specified and formalized, or whether justifications like those from game 10 can be checked by machines.

Nonetheless, what I have written here might be enough to encourage us to trust people playing games like #10. A world in which we were scrupulous about this would look very different from the present. Currently:

- Google Search and Amazon help us with our goals (Game 2) but do they help us live in the way we want?
- Voting and Markets give us what they claim we prefer (Game 3), but—the projects we vote for, or the products we buy support—does they underwrite the ways of life we believe in?

- Facebook and Netflix give us more of what we like (Games 1 and 3), but do they give us more of what we value?

Something like game 10 could serve as a basis for political decision making, economic decision making, and metrics at Internet companies. Indeed, there are places where similar games—of which Amartya Sen's capability approach is the best known—are already installed.

# Cast of Characters

Game 1. Positive Feelings

- **Larry** assesses your feeling-state periodically and keeps a log

Game 2. Goals Reached

- **Jeff**, a billionaire, runs an e-commerce site
- **Andy**, a user with goals, comes to Jeff's e-commerce platform
- **Belle** has conflicting goals
- **Carmen** uses Jeff's digital assistant technology. It reads her mind.
- **Dante** also uses Jeff's digital assistant technology. He uses it to set up all of his relationships and collaborations.

Game 3. Preferences Revealed

- **Moses** is a city planner and **Mark** runs a social network. Both play a justification game around revealed preferences
- **Emily**, a user with preferences, uses Mark's app in a way that might reveal a preference for procrastination and social isolation.

Game 4. Advances in Wellbeing or Flourishing

- **Sergei** is not limited to recording your emotion/wellbeing/eudaemonia state and inferring what to do from it

- **Francesca** doesn’t seem to be going for wellbeing
- **Rosa** might agree with Sergei’s data but still choose activism over the bathtub

Game 5. Triumph of Our Better Selves over Our Unreasoned Selves

- **Eliezer** tries to recognize when people are being rational and when they aren’t
- **Giorgio** has two selves

Game 6. Triumph of Our Better Selves over Zero-Sum Games

- **Scott** presumes that people do in fact have good reasons for even ultimately meaningless acts.
- **Hector** has evolved to play status games
- **Isaac** has evolved to play social games more generally
- **Jen** works to set the agenda

Game 7. Advances in Knowledge

- **Neil** runs a television show and media empire
- **Kate** is a scientist and an orthodox Jew

Game 8. Living by Values

- **Charles and Amartya** are acutely interested in how it's important to you to live
- **Liz** wants to be brutal, physical, and centered when she practices jeet-kune-do
- **Marty** also wants to be generous, but in the manner of Oprah

Game 9. Living by Your Present and Future Values

- **Nina** wants to be both tactful and clear with her coworker
- **Oliver** wants to be firm with his children
- **Oliver** and **Pete** like to talk about their values very openly
- **Quinn** values being likable and fun, but she also values being at ease

- **Ruth** and **David** do care about your feelings—especially about feelings of conflict

Game 10. Living On Your Own Terms

- **Ruth** and **David** build—as a side project—a social network for entrepreneurs
- **Joe** and **Tristan** try to collect information about how you act to influence others and to fit in in different contexts

Here, I will mostly limit my discussion to justification games that are practiced (or practicable) in everyday life. So, in discussing preferences, I'll address just the understandings of preferences common in business, public policy, design, and everyday life. I won't argue against the complex notions used by certain preference-utilitarian philosophers, which mostly aren't of practical use in justifying projects.

I will ensure any a new approaches I put forward are similarly practical.

An argument in Velleman's Beyond Price suggests that you mustn't have a goal for anyone you love (even for them to be happy), or you'll trample on their agency (which he says is the core of what you love).

You might imagine that Larry will make you unhappy eventually if he chooses to be short-term when you prefer long-term, but is this true? He may be able to keep you in an unreflectively delighted state.

If it were possible to copy an individual, try different interventions, and discern separately the impacts of each, and that these were additive, this registration problem could be overcome. But since this is not possible, claims of benefit justified by this kind of data are unlikely to be well-grounded.

Ioannadis 2005, Why most published research findings are false

"a research finding is less likely to be true when.... when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes."

Dead salmon fmri paper

You might think that it would help Larry to have a large population to sample from, but adding data about other people to this mix can also make the problem worse: Larry can use population data to claim that a certain level of happiness is

the best you can hope for, or that your career goals will make you unhappy. He can use *other people's moods* to claim that a cocaine habit makes you happier in the very long run; *etc.*

A proper definition comes later.

The idea of a tool-using, goal-driven person still shapes and justifies our operating systems and design methods in technology—such as the Jobs To Be Done design framework (often used in design thinking), and the meta-goal of augmenting human intellect.

Sen 1973 - and trace his references back

Goals can be reformulated to include not just where you got but how you got there. (Sen 1973) But once this is done the notion of goals or preferences isn't necessary, as I'll show with games 8 and 9.

Optimizing based on this leads to the situations pointed to above—situations where values like productivity and efficiency are stars of the show, but other ways of operating that people value, like *courage* and *exploration*, are starved. While *goals achieved* goes up, something unaccounted for goes down, and projects which were actually destructive appear beneficial.

Elizabeth Anderson (1993) has pointed out that goals/preferences views reduce us to either valuing-by-desiring (in the case of goals) or valuing-by-liking (in the case of preferences). These two ways of valuing are unusual in that they are private and hard to argue about. As we will see in game 9, arguing seems to be an important component in robust justification games.

Although the idea that utility can be characterized as a scalar function has gone out of style. See Isaac Levi.

See Stigler and Becker (1997) for an even more minimalist view, where everyone has the same preference profile, and we differ instead in the sort of

internal capital we've built up or been endowed with, allowing us to leverage goods and services differently.

Engagement metrics—which I've often maligned—seem right based on this view. When a city measures park use or a social network measures the use of a feature, they're trying to abdicate responsibility for deciding what's good for people.

In the discussion here, I assume lifetime wellbeing is the kind of wellbeing to go for. But this is an unfounded assumption. Reject it and this justification game is even harder to make robust. What kind of time horizon or discount rate should Francesca use in these decisions? Should she prioritize short-run happiness or long-run happiness? Should she make high-risk happiness investments? Your hope for Francesca might depend on whether you value living *prudently* or *adventurously*, *patiently* or *spontaneously*.

Someone who wants to claim they are serving Francesca can claim they are minimizing pain in the short run, or maximizing happiness in the long run, whichever supports their case—and they can do this without regard to how Francesca wants to live.

Max-Neef

Descriptions of the upper layer vary even more than of the lower—and they aren't important to my argument here. But to give a few: it may be supposed that while most people are busy with whatever activities this lower layer promotes, a lucky few people are: "expanding the frontier of science and human knowledge", or "working towards a non-oppression", or "bringing the kingdom of God to Earth", or "fashioning an authentic, reflected life".

Haidt

Kahneman, etc

“The Elephant in the Brain”

Goffman, Girard, Bourdieu, and Foucault

Indeed, this was the motivation for introducing these views—under the terms metapreferences or multiple selves—into economics. These conflicts and changes of mind *are* part of life, and the idea of revealed preference can’t account for them. What our choices say are our preferences *can’t possibly* be our true interests, because we all have experiences where we didn’t realize our true interests until after we’ve made the wrong choice.

Economists and psychologists like Daniel Kahneman, Tyler Cowen, and Jonathan Haidt have advanced multiple-selves or multiple-sets-of-preferences models. Sometimes this is put in terms of an “elephant” (a fast-thinking, “irrational”, heuristic-driven part of us) and a “rider” (a calculating, slower part). These writers avoid explicitly committing to this value judgement, but their work has been used to justify games like Eliezer’s, about getting the elephant under control.

Another approach is **metapreferences.** It suffers the same problem. This is the idea that we have preferences beyond the ones we currently hold, preferences about what we would like our preferences to be. In the 1970s, economists (including Sen ) were briefly enthralled with the idea of higher-order preferences. (See Hirschman 1985 for a tour of the metapreferences literature.)

David Velleman, Christine Korsgaard, Daniel Hausman

Velleman, 1999

Hausman (2011)

Velleman Practical Reflection (1999)

This story seems to contract a popular account, that people act from underlying

or unconscious drives, then make up reasons after the fact. That we make up (when asked) whatever reasons we can to justify our (drive-based) actions. Recent versions of this idea came from behaviorist psychology, and got a particular boost from the studies by Michael Gazzaniga in the 1960s, where split-brain subjects "confabulated" reasons for their actions. (Professor Gazzaniga was my undergrad advisor and mentor at Dartmouth.)

But in Velleman's stories, the character shows a remarkable unwillingness to "just make up reasons" after the fact. Even when the arm is doing something quite coherent, it appears hard for the character to justify it after the fact.

How these be reconciled? In the original studies (which have recently failed replication) the subjects actually had good reasons for their actions before they took them. They were told directly by the experimenter to take certain actions, before they acted. So the subjects were—quite reasonably—following instructions.

The reasons for acting in the experiments are quite clear: the scientist in charge has told them to point to one word or another according to a certain pattern.

But split-brain patients, perhaps, were not able to explain their reasons like we can. The part of them that tried to verbalize why they did a thing found those reasons inaccessible, and tried to make due with the information it *did* have.

So the Gazzaniga experiments don't show that people act and make up reasons afterwards. The split-brain patients acted with good reasons just like the rest of us.

A third appeal descends from the story about tribalism in the previous section, but where the clashing tribes are framed as "belief systems". With this view, it's not so much that people are understood as seeking a certain kind of consistency of belief, and—because achieving this consistency of belief is only possible inside one systems or another—this creates a clash of civilizations, or

worldviews, or belief systems. Sometimes those with this view believe that there's one "correct" belief systems: often the scientific/Liberal one, although occasionally it's a traditional, "nonwestern" one. Others just imagine a kind of continual struggle with no right answer.

In the standard versions of this view, religions or "civilizations" are considered as monolithic and as defined not by their norms and practices, but rather by their beliefs. So Islam is defined by "belief in the Koran," and the West by "belief in Science," or maybe "belief in Christianity," or something roughly like that.

But when you modify the terms in this way, the "clash of civilizations" isn't a matter of worldviews—Muslim vs Scientist—but rather a clash of different approaches to living well.

And if you go to church, you're more likely to find Christians talking about hope, faith, or charity than exactly how old the dinosaur bones really are.

It's not clear that we even *have* pure perceptions. Try looking around you in a way that's purely about perception without any appreciation or sense of value. In my experience, this is neither possible nor desirable.

devout Bayesians, and Sherlock Holmes, aside

This values-driven account of human nature has antecedents in Aristotle, Aquinas, and Kant. A subculture of academics and writers with this view formed initially as a reaction to mid 20th-century post-modern views which didn't distinguish between values and norms, and which conceptualized all culture as a kind of war. There were simultaneous reactions to this in literature — via the New Sincerity of David Foster Wallace and Miranda July — and amongst academics who viewed meaning as concrete, communicable, accessible, lived, and not-entirely-subjective phenomenon: most notably Charles Taylor, but also Velleman, Chang, Korsgaard, Putnam, and many others.

The understanding of values I present here comes from this philosophical

tradition. If you want to go straight to the source, read David Velleman's Practical Reflection and How We Get Along, and/or Charles Taylor's Sources of the Self. If you want group exercises to get clear about your own values, check out the classes and worksheets at Human Systems

Someone might go into a salary negotiation with a goal — a concrete outcome they hope to achieve — but even with a goal, they still have ways they want to approach it. Someone may want to be *courageous* in their salary negotiation, or *fair-minded*.

Some values only apply in extremely particular situations, for instance, a electric blues guitarist may have the value of "crispy licks", a mother of "letting her child get bumped around a bit", an improviser of "maintaining a loose awareness of the shape of the room", *etc.*

Just like someone could get distracted and forget one of their goals—perhaps arriving late to a meeting—someone can forget one of their values, and for example end up wishing they'd remembered to *notice their sensations* on that bike ride.

Like plans, values are necessary because of our bounded rationality. We are unable to calculate, in each conversation, at each moment, what to reveal and what to conceal. Instead, a person adopts the general value of *being honest*, because they've decided this is a good thing to aim at, in general. So we formulate values as guidelines for ourselves, because to live without them would mean continuous, difficult calculations. (Bratman)

Although not directly about value realism, Boyd's "How to Be a Moral Realist" is great on this; Gibson's ("The Senses Considered") account of perception is also relevant.

Or when our *genes* think something is important to look at.

Even if you find something new amidst data that was recorded for a different

reason, your act of noticing this new thing is still driven by an interest in it—a value this pattern has for you.

Values are more fundamental than goals, because values guide us in picking and revising our goals.

While some values (like the above) have names, most don't. But values without names can usually be referred to by phrases such as "honoring the dead", or "building the capacity of the team to handle problems together". As with goals and beliefs, no such phrase ever expresses the entirety of what someone means by a value.

Any sentence can be read different ways—including any sentence about values or goals. When you go to the store to buy bread, are pretzels also a possibility? Statements of a person's values in language *do* have this ordinary ambiguity (as do statements of goals or preferences or any other statement). In general this ordinary ambiguity is tractable—you can always ask someone to be more specific, to explain whether they meant this word meaning or that other one, to expand a phrase into a paragraph.

Short value statements like "being courageous" or "being honest" will always need to be expanded in this way to get at how the person actually wants to live, just as a short goal statement like "start a company" doesn't fully specify the goal—would they be happy starting a barbershop? Clarification is needed and clarification is possible. There is no point of absolute clarity, but in practice we can figure out, for different audiences, how specific to be, and such language works out most of the time.

Charles Taylor - epistemic gain

Ruth Chang - a more comprehensive value

David Velleman - advances in self understanding

We can even say that Oliver and Pete have discovered the same (objective) value, if we grant that it's possible for them to do roughly the same values-experiment (say, trying *honesty* and *dishonesty* with their spouse).

Actually, I think the case for the convergence of facts is overstated. Because different values in the recording of data lead to different facts being recorded, facts are divergent in the same shape as values, and facts can go off in different workable directions, just like values can.

This process-based view has swept through economics and analytic philosophy since the late 1980s, replacing ideas about metapreferences and multiple selves.

I refer to a particular process-based view, combining three ideas: *first*, that we don't tend to act when we have only desires but no reasons; *second*, that to have a good reason is to have undergone a process of reflection which has succeeded; *third*, that such success is an endorsement based on our "identity": our sense of who we believe we are and who we believe in being. By 2002, Sen had switched to this view, but it is already present in Velleman 1985, Korsgaard 1992, and Quinn 1993.

On this view, when we choose, we do so based on a rough guess of what our true interests are, and we are always looking to improve that guess by finding better values. This is a process of refining our sense of ourselves and of the best way to live, based on new considerations, reflections, and experiments.

It says that change comes through participation in a *process* — of reflection and discovery and clarification — leading to a better understanding.

The process continues:

Later, Quinn's new value of being authentic and caring comes into conflict with being effective. She notices herself being uncaring while pushing groups to be effective. Feeling frustrated and confused, she resolves this with a new view of

effectiveness, one that's about fostering capacity in herself and others.

Again, a misunderstanding is corrected: she used to think that teams were about getting things done. In transitioning away from *being effective,* she corrected a similar misunderstanding about *good teams*.

This diagram shows how Quinn's conflicts lead her (via feelings) to find more comprehensive values:


"Brain-hacks."

Edward Bernays - Early ads for smoking - plus actual social networks.