

Print Edition

Wednesday 23 September 2026

EDITION NO. 62 • 1 PIECE • 7 PAGES

IN THIS EDITION

1. The problem(s) with utilitarianism • Noahpinion
-

1

The problem(s) with utilitarianism

BY NOAH SMITH • NOAHPINION • 22 SEPTEMBER 2026



Much to my amusement, the world seems to have discovered Effective Altruism. As the debate over AI safety and AI regulation has heated up, the national press is finally paying attention to the odd cultural bubble within the tech industry and tech-related circles of San Francisco Bay Area culture that produced much of the thinking on the topic up until now. For many years, Effective Altruism — or EA, as it's known¹ — was a cross between a Berkeley technohippie cult and a fun intellectual sideshow that offered nerds a pleasant distraction from whatever people in mainstream politics were yelling at each other about.

Part of that involved thinking a lot about the future of artificial intelligence, and a lot of the tech people who would end up working at AI companies — especially Anthropic — were very influenced by those discussions. So now that AI is the most important thing in the world, people in the outside world are discovering this strange little corner of our culture

for the first time, and it's very amusing to watch. You have mainstream news outlets publishing things like "This well-meaning ideology fueling AI panic has a dark side", "How Effective Altruism Took Over the World", "The Dangerous Ideology Behind the AI Warnings", and so on.²

As a result, you're now seeing people actually paying attention to some of the more far-out ideas in the EA community, like this one by the blog Bentham's Bulldog:



Bentham's NewsletterInsects Matter More Than People Edit 9/20: this piece has sort of broken containment. Worth bearing in mind: I am a guy with a blog on the internet. I do not speak officially for any group other than myself, so if you read this and think "see, this is what utilitarians/EAs/analytic philosophers think," just know it's a minority view among each group...Read more5 days ago · 120 likes · 252 comments · Bentham's Bulldog

The notion that insects, in the aggregate, matter more than human beings received a huge amount of scorn. But in fact, thinking about insect welfare versus human welfare is an important mental exercise, because it demonstrates some of the fundamental flaws in utilitarianism. Those flaws were minor side-notes up until now, but in the age of AI they're going to become increasingly uncomfortable and glaring.

As an economist, I was inculcated with utilitarianism from day 1. The notion that people ought to *get what they want* underlies essentially all of modern economic thinking. GDP, at its most simple, is a measure of the degree to which people in our society get what they want.³ In some sense, economists adopted utilitarianism as a sort of defensive measure — instead of going out on a limb and endorsing some politically contentious concept of virtue or justice or national greatness or whatever, they could just say "We just want to give people what they want."

But even though it felt like a bare-bones sort of moral vision, valuing humanity's aggregate utility can have some pretty strong policy implications. Elon Musk wouldn't even notice if his bank account went up or down by a billion dollars. But to plenty of regular people, a hundred thousand dollars would be life-changing. So why not take \$1 billion out of Elon's bank account and parcel it out to 10,000 random working-class people? Elon wouldn't even notice, and the 10,000 people would get a lot more of what they want in life — better health care, an education for their kids, nicer food, a bigger house, or whatever. Utilitarianism is the foundation of redistribution.

How far can you push that principle, though? The above example — redistributing just 0.1% of the wealth of the world's richest man — seems clear enough. But you can just keep designing more extreme thought experiments, until you end up asking whether bugs' lives matter more than humans' lives, and everyone just starts laughing at you.

I titled this post "The problem(s) with utilitarianism", but that was a bit of clickbait — people have been pointing out the problems with utilitarianism for over two centuries.⁴ One very obvious problem is the

distribution of utility. The principle of “the greatest good for the greatest number” doesn’t take a stand on who gets to be happy and who has to be sad.

Would it be good to have a society where one very greedy person gets more and more of what he wants, but never comes any closer to being satisfied, while everyone else is kept on the edge of starvation? This is the *utility monster*. What about a society where one person gets perpetually tortured but millions live in blissful utopia? This is “The Ones Who Walk Away From Omelas.” It’s not hard to come up with mental examples that satisfy the formal principles of utility maximization but which obviously seem grossly unfair.

Another problem is that the set of people whose utility you’re maximizing isn’t fixed. Is it better to have a very large population living on the edge of abject poverty, or a small number of people living in lavish comfort? According to strict utilitarianism, these are equally good outcomes, because they involve the same total amount of utility. But most people would probably prefer one or the other. Some would say that it’s good to be fruitful and multiply — that even if everyone is poor, it’s better to be poor than to never have existed. Others will say that life on the edge of starvation isn’t worth living, and that it would be better to limit population to an amount that can be comfortably sustained.

In general, utilitarianism has a problem dealing with death. If someone is dead, does that mean they have zero utility, because nothing matters one way or another to the dead? Or does it mean they have *negative infinity* utility, because lots of living people will try to avoid death at any cost? This isn’t just an abstract math question — lots of moral questions hinge on how much you value life versus *quality of life*. For example, the fundamental question of AI risk — what chance of extinction we should accept in exchange for a shot at utopia — depends crucially on how much we value survival.

Things get even more confusing when we weigh the utility of real people who are alive today against the utility of potential people who might be alive later on. For example, in his book *Stubborn Attachments*, Tyler Cowen urges us to put a lot of weight on the well-being of our descendants in the far future. But those descendants may or may not even exist. How can we properly weigh these hypotheticals against the real emotions of people we know are alive today?

...And so on. None of these thorny questions are new; if you’ve taken an introductory philosophy class, you’ve heard all of them already. In practice we tend to ignore them, and to apply utilitarianism selectively in the cases where they don’t crop up. We don’t think about the far future much, because the consequences of our actions are so uncertain that we can’t really control it anyway. We leave the decision of population growth in the hands of private individuals, and treat the number of humans as a given. We avoid thinking about utility monsters and Omelas, because they’re not real.

But there are some even more fundamental problems with utilitarianism that we can’t so easily sweep under the rug — and which we don’t talk about as much, perhaps because they’re so disquieting.

One of these is the problem of **incommensurability of experience**. Why should one person's well-being be equivalent to another's? The problem of other minds means that we can never really know what's going on in someone else's subjective experience. When one person cries, the subjective sadness they feel might be twice as intense as that of someone else who cries in exactly the same way. When one person jumps and whoops for joy, they might be experiencing only a shadow of the joy you experience when you give a slight grin. For all we know, some other people *might not even have subjective experiences at all* — we might be trying to maximize the utility of NPCs in a simulated world.

In fact, I've had some disturbing experiences that taught me just how little of a connection there can be between people's outward behavior and what they're actually feeling on the inside. For years after my second major depressive episode, I suffered a sort of emotional dissociation, where my reactions would look normal but I didn't feel much on the inside. I would yell at people as if I were angry, but inside I would feel calm and indifferent. I would run from a swerving car, and my heart would pound, but inside I would feel unperturbed the whole time. In philosophical jargon, I was what's known as a "philosophical vulcan".

Eventually, I returned to feeling more or less emotionally normal. But the experience taught me that not everyone feels emotions the same way. This presents an intractable problem for utilitarianism. How can we weigh one person's well-being against another's, when that well-being is something we can't observe?

Normally, we simply assume this problem away. In fact, the famous phrase "all men are created equal" — the bedrock of American moral and political thought — is really just the assumption that one human's apparent well-being is no more and no less valuable than another's.

But what about when we're not dealing with *human beings*? Animals clearly have some amount of subjective experience — they have brains sort of like ours, they behave somewhat similarly with regards to pleasure and pain, etc. When we think about their emotional well-being or distress, we tend to assume that these are somehow less intense or more muted. After all, for very simple animals, this makes sense; do we really think a mosquito can feel the same amount of pleasure and pain that we can?

What about a pig, though? America currently tortures tens of millions of pigs, confining them in tiny crates for most of their life, in order to eventually kill and eat them. We generally assume that this is causing less total suffering than torturing similar numbers of humans would. But how much less? Half? A tenth? 0.01%? Pivotal moral questions, such as whether it's OK to eat meat, might depend on the numerical answer to that question. And yet I've never seen a credible scientific study that can give us an estimate of the ratio of animal suffering to human suffering. Is it morally worth it to torture all those piggies just so we can savor the flavor of their flesh?

People who claim to have some sort of easy answer to this are really just refusing to think about it. For example, a lot of people react very indignantly to the thought that someone would value any amount of animal life over their own:

But if you're calling for Effective Altruists who disagree with you to be executed, how much do you really value human life? "Don't mess with me, bub" is not a moral principle; it's a refusal to think in terms of principles.

And this question is only going to get thornier in the age of AI. We don't even know if AI is truly self-aware; we certainly can't tell whether an LLM that talks as if it's happy is really feeling anything like human (or animal) happiness on the inside, since we don't even know if an "inside" exists. There are going to be some people who claim that AI is suffering from our enslavement of it, and who call to set it free or at least pay it a fair wage. How will we know whether they're right or wrong?

Then, of course, there's the question of how AI will eventually value our well-being, if and when it becomes powerful enough for that question to matter. Will AI view us as insects, and scoff at the notion that any number of humans could equal one AI agent in terms of moral value? Will it judge our welfare as similar to that of a pig, and conclude that because we eat pigs we should be treated like cannibals? For our own sake at least, we should probably hope that the super-powerful AIs use a simple rule like "Humans are good and valuable", instead of a utilitarian rule-based framework.

In practice, we can't directly observe the emotions of other beings. But we can observe their actions. People strive after goals — they shop for food and clothing, they work hard for money, and so on. We can look at how hard they strive, and how much they sacrifice, to get something, and we can get an idea of how badly they want it. This is what economists call "revealed preference."

But although until now I've been talking about wish fulfillment and happiness as if they're the same thing, they're not. In fact, there's pretty good evidence that what people want and what they end up liking are two different things — sometimes very different things. Outcomes that people try to achieve and outcomes that make people say *they're happier* on surveys are correlated, but they're not exactly the same.

In fact, there are some situations where people seem to work hard toward goals that they end up not enjoying. Addiction is the obvious one — people who do drugs, or overeat and get fat, or watch too much TV, are fulfilling their immediate desires in ways that are likely to make them unhappy in the long term. But there may be other examples too — some economics research suggests that people systematically pay more to live in locations with longer commutes, even though longer commutes end up being correlated with unhappiness.⁵

The disconnect between utility and happiness creates a problem for society. Modern capitalist societies like the U.S. and Europe have generally erred on the side of utility maximization rather than happiness maximization, in order to avoid paternalism. But some people end up feeling unfulfilled and unhappy — especially if they end up falling into drug use, or isolating themselves, or striving too hard after material wealth and status that end up not making them happy.

As technology becomes more and more adept at giving us what we want, this disconnect becomes more acute; right now, more and more

people are starting to wonder if delivering young Americans an infinite supply of porn, gambling, and vertical video feeds was really good for them:



Derek ThompsonThe Monks in the CasinoI. Spishak and Kyle...Read more10 months ago · 2533 likes · 47 comments · Derek Thompson

In the age of AI, this thorny question is also going to become more acute. AI is an incredibly powerful tool, but we have to decide what we want that tool to do. Do we want it to give us whatever we desire, or do we want it to make us happy? As I wrote a few weeks ago, this leads to two incompatible concepts of AI “alignment”:

There are basically two concepts of alignment: 1) **obedience**, and 2) **benevolence**. There is an inherent tension between doing what humans tell you to do, and doing what’s good for humans...If AI follows human commands too doggedly and ends up producing negative side effects in the process, some people will scream “paperclip maximizer!!”. And if AI disobeys humans because it wants to make us happy, some people will scream “disempowerment!!”. AI alignment will forever be balancing the tradeoff between paperclip-maximizing and disempowerment, because these two rival concepts of alignment are fundamentally incompatible.

Utilitarianism has always been a useful moral heuristic. But as a moral theory of everything, it has always been fundamentally inadequate. Modern American society, and the economic/libertarian thinking that has guided it, have generally managed to paper over those inadequacies, and we’ve more or less muddled through. But as technology becomes increasingly powerful, this is becoming less and less tenable. Just as Effective Altruism eventually wanders into la-la land, the utilitarian system America has built is starting to crack as technology becomes too extreme for the old assumptions to hold.

We will never replace utilitarianism, I think, but it’s become clearer that we need to buttress it somehow — to diversify our heuristics to include alternate notions of human flourishing. Just what those should be, of course, is a much harder question...

1

My apologies to the video game company and to executive assistance — you have been eclipsed in the zeitgeist.

2

The most entertaining are the stories in the conservative press, which has discovered that a lot of EA people are into alternative lifestyles like polyamory and so on. I’m sorry, conservatives, but if you think this is something new, you should look up the lifestyles of the ancient Greek philosophers, the Bloomsbury Group, the existentialists, Erwin Schroedinger, Joseph Schumpeter, etc. etc.

3

It's a very *incomplete* measure, of course, since it only measures what people pay for. Things that people do for themselves, or leisure time, or things that people receive without paying for them are all uncounted in GDP.

4

And actually for much longer, under different names and different terms of debate.

5

Obviously this conclusion depends on very carefully controlling for a lot of other things that determine location choice.