

The Altruistic Parasite:

Why the Wrong Reasons Are the Only Reasons That Work

There is a parasite called *Toxoplasma gondii* that does something genuinely disturbing. It infects rats, travels to their brains, and rewires their fear circuits so that the smell of cat urine stops triggering panic and starts triggering something closer to curiosity, even attraction. The rat walks toward the cat. The cat eats the rat. The parasite, which needs a feline gut to complete its reproductive cycle, wins. The rat's behaviour looks suicidal. From the parasite's perspective, it worked perfectly.

I keep coming back to this when people argue that doing the right thing for the wrong reasons is somehow morally tainted. The implicit assumption is that "pure" motives exist, that they are achievable, and that we should hold out for them. I think that assumption is not just philosophically naive — it is biologically illiterate. The wrong reasons are not a corruption of morality. They are its engine. But I want to argue that carefully, because the strongest version of the opposing view deserves a direct answer before I discard it.

The Kantian Objection, Fairly Stated

Kant's position is not that outcomes don't matter at all. His claim is more specific and, honestly, more interesting: that moral worth inheres in the will, not the result. An action done from duty — from the recognition that it is the right thing regardless of how it makes you feel — has a different moral character than the same action done from vanity, fear, or neurochemical reward. The drowning child gets saved in both cases. But a world in which people save drowning children because children matter, because human dignity commands it, is a genuinely different world from one in which they save them to quiet a cortisol surge. Kant thinks that difference is not trivial. It shapes what kind of agents we are, what habits we form, what we can be relied upon to do when the reward stops.¹

This is a serious argument. I don't think it is ultimately right, but it cannot be batted away with "the child still gets saved." Kant is not talking about the child. He is talking about the world we are building through our choices and what we lose when we abandon the aspiration to act from something other than appetite.

Here is where I think he goes wrong. The Kantian aspiration has been waiting approximately 250 years for a human population large enough to test it. It has not materialised. What we have instead is the actual record of human moral behaviour: neither purely dutiful nor purely self-interested but some unstable mixture of instinct, status anxiety, social mimicry, and occasional genuine compassion. Kant's framework describes how a perfectly rational being would act. It has much less to say about how the actual beings standing in front of us manage to do anything useful at all.

The Altruism Problem

Here is what raw Darwinian theory never quite explained: why do humans sacrifice for strangers? Kin selection explains why you might jump in front of a bus to save your child — your genes survive through theirs. Reciprocal altruism explains some broader cooperation. But neither covers anonymous charity, soldiers throwing themselves on grenades, or the bystander who runs back into a burning building for a stranger's child.²

The more compelling answer is that nature did not give us pure altruism. It gave us a proxy. We help others because not helping triggers a neurochemical punishment. The sight of a drowning child produces a cortisol surge that is, physiologically, almost unbearable. Jumping in quiets it. The dopamine hit from social praise after a publicly generous act is real and measurable. We are not choosing good over evil in some Kantian act of rational will. We are chasing relief and reward, just like every other animal. The motive is selfish at the biochemical level. The outcome is altruistic.

This is not a cynical reading. It is a more honest one. Peter Singer has spent decades arguing, on purely rational grounds, that we should give far more to global poverty relief. His arguments are logically airtight and have largely failed to change behaviour at scale. People do not act on pure reason. They act on feeling, status, narrative, and social pressure. Working with the wrong reasons is not a compromise — it is the only available strategy.³

The Market of Appearances — and Its Failures

Consider what we dismissively call virtue signalling. The phrase is now mostly used as an insult, a way of accusing someone of performing goodness rather than practising it. But if people only did good things when their hearts were completely pure, the total amount of good in the world would collapse toward zero. Cynical motivations — a billionaire funding a library to put his name on it, an individual acting generously because their social circle would judge them otherwise — create a functional society even when they create nothing like a virtuous one.

Ernst Fehr and Simon Gächter demonstrated this experimentally in 2000: when participants in a public goods game could punish free-riders, cooperation rose dramatically, not because people became more virtuous, but because the social cost of being seen as selfish increased. The wrong reason produced the right behaviour. Mandeville made the same argument three centuries earlier in *The Fable of the Bees* (1714), arguing that private vice produces public benefit, and was mostly hated for it.⁴

But wrong reasons can also fail. Research on "crowding out" shows that external incentives for prosocial behaviour can sometimes destroy the intrinsic motivation that existed before. Bruno Frey and Felix Oberholzer-Gee documented this in a 1997 study: when Swiss municipalities were offered financial compensation for hosting nuclear waste sites, local acceptance fell from 50% to 25%. Once money entered the picture, people reframed a civic obligation as a transaction — and rejected it. The

wrong reason crowded out the right one. So the argument is not that wrong reasons always work. It is that they are often the only tool we have, and the question of when they backfire deserves more attention than this essay can fully give it.⁵

Hypocrisy as Social Infrastructure

The hypocrisy objection to wrong-reasons ethics is that performative virtue is corrosive — it crowds out genuine commitment, gives cover to people who benefit from the appearance of goodness, and degrades the moral norms it claims to uphold. There is something to this. But the historical record suggests hypocrisy has also been one of the main mechanisms by which moral progress actually happens.

Consider the abolition of the British slave trade. The coalition William Wilberforce built in the 1790s included men with deeply mixed motives: religious conviction in some, economic calculation in others, political opportunism, social fashion among a London elite that found abolitionism increasingly chic. The Clapham Sect was sincere. Many of their parliamentary allies were not. The Slave Trade Act of 1807 passed anyway — not through moral awakening across the Commons, but through votes from men whose conversion was social and reputational before it was genuine. The performance created the expectation; the expectation eventually created the norm. Erving Goffman's argument in *The Presentation of Self in Everyday Life* (1959) — that social life is fundamentally theatrical — is a description of how norms propagate even when individuals fail to embody them.⁶

The AI Problem

None of this is merely historical. Whether right actions can be divorced from right reasons is now one of the most technically pressing problems in computer science — and the answer is that they not only can be separated but must be.

When researchers train large AI systems, they cannot code in goodness directly. Nobody knows what that looks like in mathematical terms. Instead, they give the system a reward function: a proxy. A wrong reason. The field of AI alignment is essentially the study of how to choose wrong reasons carefully enough that the system produces right outcomes reliably. Stuart Russell has argued that the challenge is not building an AI with pure values but building one whose imperfect, proxy-based values are safe enough not to produce catastrophic results.⁷

If an AI system cures a disease because it has been optimised to minimise a health metric, it has done the right thing for what is, philosophically, a hollow reason. It has no concept of suffering, no compassion, no moral understanding. It moved numbers in a direction we told it to move. And yet people do not die of that disease anymore. Insisting that this outcome is morally void because the AI lacked pure motivation is a position that only makes sense if you care more about moral aesthetics than actual lives. It is what might be called a luxury belief — a position affordable precisely because its costs are abstract. The same logic applies to every human institution that

produces good outcomes through impure means, which is most of them.

What Moral Psychology Actually Shows

The standard move at this point in an essay on moral psychology is to cite Jonathan Haidt's *The Righteous Mind* (2012) for the claim that moral reasoning is mostly post-hoc rationalisation. That is accurate but unremarkable. What is more interesting, and more contested, is whether this finding closes the normative question or merely complicates it.

Haidt's social intuitionist model has faced substantive criticism. Paxton and Greene, writing in *Topics in Cognitive Science* in 2010, argued that the relationship between intuition and reasoning is more bidirectional than Haidt allows — that deliberation, when confronted with strong arguments, can shift what people intuitively feel rather than merely rationalising pre-existing responses. This matters here because if reasoning only ever rationalises emotional responses, the aspiration toward pure motives Kant describes is not just psychologically unrealistic but categorically impossible. If reasoning can sometimes revise intuitions, the Kantian project is merely very difficult — a different claim. I find myself genuinely uncertain about this. Whether emotional responses coming first "most of the time" means "always" is a question the data does not fully settle.⁸

What I am more confident about is that the aspiration toward purer motives has its own instrumental value. Holding ourselves to the standard of acting from duty — even while accepting that we mostly fail — creates accountability that pure consequentialism cannot generate on its own. The Kantian tradition is not wrong to want something more than outcomes. It is wrong to insist that something more is achievable in any consistent way. The aspiration is useful. The insistence is dangerous because it has historically led to the dismissal of outcomes that matter enormously to the people experiencing them.

The Clock and Its Gears

Humanity is a species that mostly stumbles into virtue while chasing something else. We built hospitals out of status competition, abolished legal horrors through political coalition, and are currently attempting to align AI systems by engineering elaborate proxies for values we cannot fully articulate. The gears that drive all of this are ugly — neurochemical self-interest, social vanity, incentive structures, fear of punishment. They are also the only gears there are.

This does not mean that motives are entirely beside the point. A society that produces good outcomes entirely through incentive systems and social performance is probably more fragile — less capable of sustaining norms when external enforcement breaks down — than one where people have internalised some genuine commitment to each other. That is a serious concern. I do not dismiss it.

But the answer to that concern is not to demand purity before acting. It is to recognise that wrong reasons, properly channelled, are how the right things actually get done, while remaining alert to the conditions under which they stop working. The rat walked toward the cat because its fear circuits were hijacked. It did not know what it was serving. We are in a slightly better position: we can study the mechanisms, design better incentives, and notice when the wrong reasons are producing the wrong outcomes.

The clock still needs gears. But we can build better ones.

Endnotes

1. Kant, I. (1785). *Groundwork of the Metaphysics of Morals*, §1–2. The core claim is that an action has moral worth only when performed from duty, not merely in conformity with duty.
2. Hamilton, W.D. (1964). The genetical evolution of social behaviour I & II. *Journal of Theoretical Biology*, 7(1), 1–52; Trivers, R.L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46(1), 35–57.
3. Singer, P. (2009). *The Life You Can Save: Acting Now to End World Poverty*. Random House. Singer discusses the gap between argument and action directly in Chapter 9.
4. Mandeville, B. (1714). *The Fable of the Bees: Or, Private Vices, Publick Benefits*. J. Roberts; Fehr, E. & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4), 980–994.
5. Frey, B.S. & Oberholzer-Gee, F. (1997). The cost of price incentives: An empirical analysis of motivation crowding-out. *American Economic Review*, 87(4), 746–755. The study was conducted among Swiss residents in communities near proposed nuclear waste sites.
6. Goffman, E. (1959). *The Presentation of Self in Everyday Life*. Anchor Books.
7. Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking. The reward-proxy problem is central to Chapters 5–7.
8. Haidt, J. (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Pantheon Books; Paxton, J.M. & Greene, J.D. (2010). Moral reasoning: Hints and allegations. *Topics in Cognitive Science*, 2(3), 511–527.

Bibliography

- Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4), 980–994.
- Frey, B.S., & Oberholzer-Gee, F. (1997). The cost of price incentives: An empirical analysis of motivation crowding-out. *American Economic Review*, 87(4), 746–755.
- Goffman, E. (1959). *The Presentation of Self in Everyday Life*. Anchor Books.
- Haidt, J. (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Pantheon Books.
- Hamilton, W.D. (1964). The genetical evolution of social behaviour I & II. *Journal of Theoretical Biology*, 7(1), 1–52.
- Kant, I. (1785). *Groundwork of the Metaphysics of Morals*. (Trans. M. Gregor, Cambridge University Press, 1997).
- Mandeville, B. (1714). *The Fable of the Bees: Or, Private Vices, Publick Benefits*. J. Roberts.
- Paxton, J.M., & Greene, J.D. (2010). Moral reasoning: Hints and allegations. *Topics in Cognitive Science*, 2(3), 511–527.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.
Singer, P. (2009). *The Life You Can Save: Acting Now to End World Poverty*. Random House.
Trivers, R.L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46(1), 35–57.

Authorship Declaration

I declare that this essay is my own original work and that it has not been published elsewhere or submitted to any other competition. I have not used artificial intelligence tools to write this essay. I confirm that I have discussed ideas with others but that the writing and arguments presented are entirely my own.

Category: Philosophy

Question answered: Question 1 — Is it ever wrong to do the right thing for the wrong reasons?