You should believe and desire what would be best

University of Toronto, March 5th 2026 | Richard Roth

1 The twin questions

The Belief Question What should you believe? And why?
The Desire Question What should you desire? And why?

Today, I will defend an answer to these twin questions:

Counterfactualism You should believe what it *would* be best to believe, and desire what it *would* be best to desire.[1]

'Should' must be interpreted as coordinated with 'best' and 'would'.

This answer sounds intuitive, and not particular to belief and desire.

A: Should I buy apples or pears?
B: It would be best to buy apples.
A: So I should buy apples?
B: #I didn't say that. You should buy pears.

A: Should I believe that she broke my vase?
B: For now, it would be best to suspend judgement.
A: So I shouldn't believe it?
B: #I didn't say that. You should believe that she broke the vase.

Still, Counterfactualism conflicts with popular "actualist" norms:

You should believe what's in fact true.
You should desire what's in fact good.
You should believe what's in fact likely, given your evidence.
You should desire what's in fact expectedly good, given your evidence.

Many philosophers prefer such "actualist" norms. They worry that norms like Counterfactualism license "epistemic bribes", i.e. believing obvious falsehoods when this would cause enough true beliefs.[2]

Plan: ① Give two arguments for Counterfactualism, ② respond to the "bribes" objection, and ③ draw out two consequences.

[1] Counterfactualism is closely related to epistemic consequentialism (next fn.), Bykvist and Hattiangadi (2013)'s Doxastic Value Maximisation, and Barnett (2023)'s idea that you may assert what would be sufficiently likely to be true if asserted.

[2] See Greaves (2013), Caie (2013, 2018), Berker (2013a,b), Joyce (2018), Carr (2017), Pettigrew (2018) and the collection Ahlstrom-Vij and Dunn (2018).

2 The argument from self-undermining

William James (1897, 23f.) observed that beliefs can be *self-undermining*.

Something can be true and likely given your evidence, and yet if you believed it, it would be false and unlikely given your evidence.

Time Trouble I'm giving a job talk. Past experience suggests that

- if I believe that I'll finish on time, I won't,
- if I don't believe I'll finish on time, I will.

My present evidence suggests that I'm in the second case.

To screen off irrelevant complications, I'll focus on a Moorean variant:

Artie Choke You're investigating a murder. All the evidence points to your friend Artie Choke, who really is the murderer. Nevertheless, you don't believe it was him, and you can introspect this fact about yourself. So you have excellent evidence for the true conjunction *Artie is the murderer, and I don't believe it* ($A \wedge \neg BA$).

Desires can be self-undermining in a parallel way.

Something can be good and expectedly good, and yet if you desired it, it would be bad and expectedly bad.

The Last Supper You're a Buddhist monk, well on your way to Nirvana. You have one desire left, to eat. You evidence suggests

- if you got rid of the desire to eat, you'd make it to Nirvana,
- but if you desired to get rid of the desire to eat, and then stopped desiring to eat, you'd starve and re-start your journey anew, with lots and lots of desires.

"Actualist" norms make counter-intuitive predictions about self-undermining attitudes (Bykvist and Hattiangadi, 2007, 2013; Barnett, 2023).

- Intuitively, you shouldn't believe the conjunction $A \wedge \neg BA$.
  But "actualist" norms predict that you should (because $A \wedge \neg BA$ is true, and likely given your evidence).
- Intuitively, the monk shouldn't desire to stop desiring food.
  But "actualist" norms predict the opposite (because stopping to desire food would be both good and expectedly good).

Counterfactualism makes intuitive predictions about self-undermining:

- You shouldn't believe $A \wedge \neg BA$ because believing that would be bad, *if you believed it*. (The belief would be bad in many ways: false, unlikely given your evidence, ...)
- The monk shouldn't desire to stop desiring food because desiring that would be bad, *if he desired it*. (The desire would be bad in many ways: a desire for something that's both bad and expectedly bad given his evidence, ...)

2.1 Intermezzo: Wide-scope norms

There are two ways to interpret norms like these:

> You should believe what's true.
> You should desire what's good.
> You should believe what's likely true.
> You should desire what's expectedly good.

'Should' could take narrow or wide scope here (Broome, 1999):

> Narrow-scope: (You should believe that p) iff (p is _____).
> Wide-scope: You should (believe that p iff p is _____).

We've rejected narrow-scope (a.k.a. "actualist") interpretations.

How about wide-scope interpretations?

- They are consistent with Counterfactualism.
- But they do not answer our twin questions, and seem undesirably weak. Where belief comes apart from the truth, wide-scope norms remain silent as to which is to be adjusted. But many have felt that there is an asymmetry here (Anscombe, 1957, 56, cf. Schroeder, 2004; Kolodny, 2005).

Counterfactualism avoids this criticism.

- When the truth-value of p doesn't depend on what you believe, and you falsely believe that p, Counterfactualism predicts (given natural assumptions about value), that your belief in particular should be adjusted.
- And when the truth-value of p does depend on whether you believe p, there often isn't any way to adjust one in particular.

3 The argument from Normative Invariance

At the heart of our intuitions about self-undermining beliefs and desires, I want to suggest, is a sort of normative invariance.

Prichard (1932, 26) gives voice to the idea in ethics as follows:

> "the existence of an obligation to do some action cannot possibly depend on actual performance of the action"

One way to state the idea more precisely:

(Weak) Normative Invariance If φ is an option, then you should (may) φ iff if you did φ, you still should (may) φ.

A real permission wouldn't vanish as soon as it was exercised! A real prohibition wouldn't vanish as soon as it was violated!

Proposal: Normative Invariance applies to belief (and desire), too.

(Weak) Normative Invariance for Belief If it's an option to believe something, then you should believe it iff if you did believe it, you still should believe it.

3.1 "Actualist" norms violate Normative Invariance

"Actualist" norms fail because they predict failures of Normative Invariance for belief and desire. They have to say something absurd:

> "You should believe $A \wedge \neg BA$. But if you believed this, you shouldn't believe it. The monk should desire to *stop desiring food*. But if he desired this, he shouldn't desire it."

3.2 Counterfactualism entails Normative Invariance

Counterfactualism entails Normative Invariance. Or so I claim.

A number of Swedish philosophers deny the parallel entailment for consequentialism (Carlson, 1995, 101; Bykvist, 2007; Gustafsson, 2019).

They worry that both your options and the counterfactual consequences of those options may depend on what you do.

I'll focus on a case that illustrates the second worry:

Coin Flip I can bet on a fair coin landing heads. The coin will only be flipped if I accept the bet. I accept and lose.

Swedish Objection: There is a clearly true "hindsight" reading of

(1) I should have rejected the bet.

Normative Invariance then predicts that the following is true, too:

(2) If I had rejected the bet, I should have rejected the bet.

But that has probability .5 only, as it is contextually equivalent to

(3) If I had rejected the bet, then if I had accepted the bet, I would have lost.

Reply: I claim that (3) has probability 1, not .5.

- It's hard to have direct intuitions about (3)'s probability.
- But my probability assignment fits intuitions about other counterfactual concepts, such as *prevention*:

Coin Flip Veto I can bet on a fair coin landing heads. You can veto my decision. The coin will only be flipped if I accept and you don't veto. I accept, you don't veto, and I lose.

There is a clearly true reading of

(1*) You could have prevented me from losing.

But this has probability 1 only if the following has probability 1, too:

(2*) If you had vetoed, you would have prevented me from losing.

And (2*) is contextually equivalent to

(3*) If you had vetoed, then if you had not vetoed, I would have lost.

And if (3*) has probability 1, then surely (3) does, too.

I think examples like this point to a gap in our theories of conditionals. Pairs like this are equivalent:

(I) If I had accepted the bet, I would have won. $(A > W)$

(II) If I had rejected the bet, then if I had accepted the bet, I would have won. $(\neg A > (A > W))$

Existing theories of conditionals don't predict such equivalences.[3] But that's unsurprising, since conditionals like (II) are hard to interpret.

[3]Theories validating Import-Export ($(A > (B > C)) \equiv ((A \wedge B) > C)$) trivialize (II). Others, e.g. Stalnaker (1968) and Lewis (1973), don't trivialize (II) but don't make it equivalent to (I). Yablo (1992) and Hare (2011) reject equivalences like (I) $\equiv$ (II).

In other work, I use similar data to argue for[4]

Conditional Invariance For all options φ, ψ and conditional-free χ, $(\varphi > \chi) \equiv (\psi > (\varphi > \chi))$

Option Invariance For all options φ, your options are the same as they would be if you φ-ed.

Counterfactualism entails Normative Invariance given this pair.

- Say that the value of φ-ing is how good it would be if you φ-ed. Conditional Invariance ensures for all options φ and ψ, the value of φ-ing is the same in actuality as it would be, if you φ-ed.
- And Option Invariance ensures for all options φ that your options in actuality are the same as they would be if you φ-ed.
- So which options have the greatest value is the same in actuality as it would be if you φ-ed, and so you should do the same.

A virtue of Counterfactualism is that it ensures Normative Invariance.

4 Doxastic and Orexic Bribes

Many philosophers have worried that principles similar to Counterfactualism commit us to accepting "doxastic bribes".[5]

The problem of doxastic bribes Intuitively, you should not adopt beliefs that would obviously be bad even when they would result in many other good beliefs.

They worry that if beliefs were evaluated in terms of their consequences, we should accept "doxastic bribes" like the following:

Mystery Belief I'm putting mystery presents in numbered boxes. The first present is badly packaged: it a pile of mud, plainly visible through transparent wrapping.

- If you believe that the first box contains a good present, I will package the good presents in the even-numbered boxes. You will then be able to form true beliefs about all other boxes.

[4]Given certain counterfactual theories of ability, Option Invariance turns out to follow from Conditional Invariance.

[5]See Greaves (2013), Caie (2013, 2018), Berker (2013a,b), Joyce (2018), Carr (2017), Pettigrew (2018) and the collection Ahlstrom-Vij and Dunn (2018).

- If you don't believe that the first box contains a good present, I will allocate the presents randomly to boxes.

What you "should believe" here depends on what we mean by 'should':

- Practically speaking, you should believe that the first box contains a good present. It'll allow you to get the good presents.
- But in another *epistemic* or *rational* sense, you shouldn't believe it. It would be completely irrational.

To be maximally friendly to the objection, I will from now on assume that we're targeting the second sort of interpretation of 'should'.

The same problem arises for desire:

The problem of orexic bribes Intuitively, you should not desire something that would obviously be bad, even when doing so would cause many (other) good desires.

Mystery Desire I'm putting mystery presents in numbered boxes. The first present is badly packaged: it a pile of mud, plainly visible through transparent wrapping.

- If you desire to get the first present, I will package the good presents in the even-numbered boxes. You will then be able to form apt desires about all other boxes.
- If you don't desire to get the first present, I will allocate the presents randomly to boxes.

I will discuss three prominent reactions to doxastic bribes.

4.1 Ultra-Anti-Consequentialists

Ultra-Anti-Consequentialists maintain that in considering what to believe, you should *ignore* the consequences of believing it.

> Consequentialism in epistemology is also forward-looking (if not temporally, then at least in the order of explanation): it ties a belief's, or process', or character trait's epistemic merit to the value of the states of affairs it helps bring about (whether causally or constitutively or otherwise). But this, I think, gets things exactly backwards. Consider the following slogans: "Epistemic justification is a matter of responding to how things appear to you," "Epistemic rationality is a matter of respecting one's evidence," "Epistemic virtue is a matter of fitting together one's cognitive states into a coherent whole." Each of these is, of course, highly metaphorical, and filling in the details of what these metaphors come to requires positive theorizing. But note that, in each case, epistemic normativity is characterized as fundamentally backward-, or at least sideways-, looking. (Berker, 2013b)

Objection: Self-undermining and self-reinforcing beliefs and desires suggest that what you should believe and desire does depend on *some* (causal and constitutive) consequences of believing and desiring it. Namely, those which would affect the goodness of the belief/desire!

- Caveat: The normative jungle is full of concepts like *justified*, *rational*, or *virtuous belief* that may be back-ward or side-ways "looking". My claim is just *what you should believe* is not like that.

4.2 Error theorists

Error theorists say that, contrary to intuition, you should sometimes adopt beliefs that would obviously be false when they would cause enough (other) good beliefs. To make this seem less bad, they offer an error theory of our intuitions about doxastic bribes.

- Pettigrew (2018)'s error-theory: We judge beliefs to be permissible when they are supported by our evidence. This is a good heuristic, but it goes wrong for doxastic bribes.

Objection: If our intuitions followed that heuristic, it should intuitively seem permissible to believe *Artie is the murderer, and I don't believe that*. But intuitively, it seems terrible to believe that!

4.3 Splitters

Splitters draw a line between what you should choose (and in particular choose to believe) and what you should believe.[6]

Konek and Levinstein (2019) motivate this in terms of direction of fit:

> Epistemic actions, like all actions, are properly assessed in terms of their causal impact on the world. They are valuable to the extent that they make the world fit our desires, to the extent that they cause the world to be good (desirable). Epistemic states, on the other hand, are

[6] See also wrong kind of reason sceptics like Parfit (2001, 21ff.).

> assessed in terms of their fit to the world. They are valuable to the extent that they encode an accurate picture of the world, not to the extent that they causally influence the world so as to make it fit that picture. (Konek and Levinstein, 2019, 72; see Gallow (2021) for agreement)

Hypothesis: What you should *believe* comes apart from what you should *choose to believe* because beliefs and choices differ in direction of fit.

Two general worries about Splitting:

- Is it really better to say that you should *choose to believe* an obvious falsehood than to say that you should *believe* it?[7]
- Splitters predict you should have self-undermining beliefs:
  - According to splitters, you should believe (though not choose to believe) *Artie is the murderer, but I don't believe that*.

A particular worry for splitting a la K&L:

- Bribes arise in parallel for desire.[8] While K&L can reject doxastic bribes, they are forced to accept orexic bribes. That's just as bad! We should treat doxastic and orexic bribes the same.

4.4 Local Value

The key to avoiding bribes is to accept, in the special target sense of 'better',

Local Value Beliefs and desires are not made better by causing good beliefs or desires, or preventing bad beliefs or desires.

And Local Value is compatible with Counterfactualism!

My take: Counterfactualism captures what was right about "epistemic consequentialism". Local Value captures what was right about the objections to "epistemic consequentialism".

[7] Cf. Rabinowicz and Rønnow-Rasmussen (2004, 413): "To be sure, we also have reasons to want to have such attitudes and to try to have them, but this is because we have reasons to have them, in the first place."

[8] K&L agree that beliefs and desires have opposite direction of fit.

5 Consequences

So far, I've given two arguments for Counterfactualism, and defended it against the objection from epistemic bribes.

I want to end by showing that Counterfactualism has interesting consequences for a other issues, too. I'll mention two, the first being

Permissivism Sometimes you may believe something, but you also may believe its negation.[9]

Consider the following case from James:

Leap You're climbing a mountain but get stuck on a precipice. If you believe that you can make the jump, you will succeed. If you believe that you cannot, hesitation will ensure that you fail.

Here, it seems equally epistemically permissible to believe that you'll succeed and to believe that you'll fail. As James (1897, 97) observes:

> There are then cases where faith creates its own verification. Believe, and you shall be right, for you shall save yourself; doubt, and you shall again be right, for you shall perish.

A second consequence concerns

The Fittingness Theory of Value A possibility is good just in case it is fitting to desire it.

The Last Supper spells trouble for this theory:

(i) *Stopping to desire food* is good for the monk,

(ii) But it isn't fitting for the monk to desire this.

- (ii) follows from Counterfactualism if you should desire what's fitting to desire (though see Berker (2022)).
- Alternatively, Normative Invariance arguably applies to fittingness, too: the monk's desire can't be fitting unless it would be fitting if the monk had it! But that's ruled out by the theory.[10]

Important differences from standard examples:

- Standard self-undermining desires, such as the desire to end all desire (cf. SN 51.15 (Brahmaṇa Sutta)), prevent their own *reali-*

[9] See a. o. White (2005), Meacham (2013), Schoenfield (2014) for discussion.

[10] See Reisner (2015, 479ff.) and Rowland (2019, §7.4) for relevant discussion.

sation. By contrast, the monk's desire prevents its own *aptness*.

- Standard reasons not to desire something good are "wrong kind" of reasons, e.g. a demon threatens to punish you if you desire a cup of icecream.[11] The monk has both a "wrong kind" of reason (blocked access to Nirvana) and a "right kind" of reason (the desire wouldn't be apt, if the monk had it).

Fittingness theorists can respond in various ways.

One response: It is still fitting for an observer to desire that the monk no longer desire to eat, and that's what matters (cf. Way, 2013).

- But this threatens to lose a supposed advantage of fittingness theories — their ability to explain why the good is motivating.
- And it incurs new difficulties regarding which observer is relevant (or in Way (2013)'s words, "suitably related").

References

Ahlstrom-Vij, H. K. and Dunn, J., editors (2018). *Epistemic Consequentialism*. Oxford University Press, Oxford, GB.

Anscombe, G. E. M. (1957). *Intention*. Cornell University Press, Ithaca, N.Y.

Barnett, D. J. (2023). Cogito and moore. *Synthese*, 202(1):1–27.

Berker, S. (2013a). Epistemic teleology and the separateness of propositions. *Philosophical Review*, 122(3):337–393.

Berker, S. (2013b). The rejection of epistemic consequentialism. *Philosophical Issues*, 23(1):363–387.

Berker, S. (2022). The deontic, the evaluative, and the fitting. In Howard, C. and Cosker-Rowland, R., editors, *Fittingness*, pages 23–57. Oxford University Press.

Broome, J. (1999). Normative requirements. *Ratio*, 12(4):398–419.

Bykvist, K. (2007). Violations of normative invariance: Some thoughts on shifty oughts. *Theoria*, 73(2):98–120.

Bykvist, K. and Hattiangadi, A. (2007). Does thought imply ought? *Analysis*, 67(4):277–285.

Bykvist, K. and Hattiangadi, A. (2013). Belief, truth, and blindspots. In Chan, T., editor, *The Aim of Belief*, pages 100–122. Oxford University Press.

Caie, M. (2013). Rational probabilistic incoherence. *Philosophical Review*, 122(4):527–575.

[11] See Reisner (2015) for an exception: Jim can donate to Fauxfam. A billionaire has already donated his fortune to Fauxfam, instructing Fauxfam to use all donations for good purposes iff nobody has a pro-attitude towards Jim's donating money. (They are stipulated to have a brain scanner to tell whether the condition is satisfied.)

Caie, M. (2018). A problem for credal consequentialism. In Ahlstrom-Vij, K. and Dunn, J., editors, *Epistemic Consequentialism*. Oxford University Press.

Carlson, E. (1995). *Consequentialism Reconsidered*. Springer, Dordrecht, Netherland.

Carr, J. R. (2017). Epistemic utility theory and the aim of belief. *Philosophy and Phenomenological Research*, 95(3):511–534.

Gallow, J. D. (2021). Updating for externalists. *Noûs*, 55(3):487–516.

Greaves, H. (2013). Epistemic decision theory. *Mind*, 122(488):915–952.

Gustafsson, J. E. (2019). Is objective act consequentialism satisfiable? *Analysis*, 79(2):193–202.

Hare, C. (2011). Obligation and regret when there is no fact of the matter about what would have happened if you had not done what you did. *Noûs*, 45(1):190–206.

James, W. (1897). The sentiment of rationality. In *The Will to Believe and Other Essays*, pages 63–110. Longmans, Green, and Co., London.

Joyce, J. M. (2018). Accuracy, ratification, and the scope of epistemic. In *Epistemic consequentialism*, pages 240–66. Oxford University Press.

Kolodny, N. (2005). Why be rational? *Mind*, 114(455):509–563.

Konek, J. and Levinstein, B. (2019). The foundations of epistemic decision theory. *Mind*, 128(509):69–107.

Lewis, D. (1973). *Counterfactuals*. Blackwell.

Meacham, C. (2013). Impermissive bayesianism. *Erkenntnis*, 79(Suppl 6):1185–1217.

Parfit, D. (2001). Rationality and reasons. In et al, D. E., editor, *Exploring Practical Philosophy: From Action to Values*, pages 17–39. Routledge.

Pettigrew, R. (2018). Making things right: The true consequences of decision theory in epistemology. In Ahlstrom-Vij, K. and Dunn, J., editors, *Epistemic Consequentialism*, pages 220–239. Oxford University Press.

Prichard, H. A. (1932). Duty and ignorance of fact. *Philosophy*, 8(30):226–228.

Rabinowicz, W. and Rønnow-Rasmussen, T. (2004). The strike of the demon: On fitting pro-attitudes and value. *Ethics*, 114(3):391–423.

Reisner, A. (2015). Fittingness, value, and trans-world attitudes. *Philosophical Quarterly*, 65(260):464–485.

Rowland, R. (2019). Too little value? In Rowland, R., editor, *The Normative and the Evaluative: The Buck-Passing Account of Value*, pages 129–146. Oxford University Press.

Schoenfield, M. (2014). Permission to believe: Why permissivism is true and what it tells us about irrelevant influences on belief. *Noûs*, 48(2):193–218.

Schroeder, M. (2004). The scope of instrumental reason. *Philosophical Perspectives*, 18(1):337–364.

Stalnaker, R. (1968). A Theory of Conditionals. In Harper, W. L., Stalnaker, R., and Pearce, G., editors, *Studies in Logical Theory, American Philosophical Quarterly*, pages 98–112. Blackwell.

Way, J. (2013). Value and reasons to favour. *Oxford Studies in Metaethics*, 8.

White, R. (2005). Epistemic permissiveness. *Philosophical Perspectives*, 19(1):445–459.

Yablo, S. (1992). Mental causation. *Philosophical Review*, 101(2):245–280.