

Act-Utilitarian Truthfulness and Hodgson's Problem

Johan E. Gustafsson*

Draft: April 19, 2026 at 12:15 a.m.

ABSTRACT. In 1967, D. H. Hodgson put forward an ingenious problem for act utilitarianism: Given their indifference to truth-telling, communication between rational act utilitarians is futile. This dialogue argues that, despite many attempts, Hodgson's problem remains unsolved.

*Two act utilitarians, both rational, are locked up in separate cells. In one cell, there are two buttons: one red, one green. One button unlocks the cells; the other button — or failing to press either button within ten minutes — releases a fatal gas. The prisoner in the buttonless cell (**Knower**) knows which button unlocks, while the prisoner in the cell with buttons (**Pusher**) does not. And they both know all this. And they both know they both know that. And they both know they both know they both know that. And so on.¹*

Knower. It's the green button! Push it, and let's get out of here. Drinks later?

Pusher. Hold on. I'm feeling good about the red button.

Knower. But...you know I know the right button. And I am telling you it's the *green* button.

Pusher. I hear you. You're saying it's the green button. But why would your saying it's the green button be evidence that it's the green button? Like me, you're an *act utilitarian*: we believe acts are right if and only if they maximize the expected sum of happiness.² I guess, technically, this is *subjective* act utilitarianism — the expectational form of act utilitarianism. Anyway, my point is: we're indifferent to mere truth-telling.

* I would be grateful for any thoughts or comments on this paper. They can be sent to me at johan.eric.gustafsson@gmail.com.

¹ This set-up is a variation of a case in Lewis 1972, p. 17 — which, in turn, is a variation of a case in Hodgson 1967, pp. 38–41.

² Bentham 1789, pp. 2–4; 1970, pp. 12–13.

Knower. Look, you know I want us to get out of here alive. We both want that. You *know* we both want that. We're on the same team. Why would I lie?

Pusher. Why would you tell the truth?

Knower. To get you to push the right button. We have the same values. So I want you to have true beliefs so you can achieve our shared goals.³ Like, say, getting us out alive.

Pusher. I don't doubt your wanting me to have true beliefs. I doubt my having a reason to believe you. It would only be in everyone's interest for me to believe you if you were telling the truth. But you have no reason to do so, because you have no reason to believe I would believe you. So I have no reason to believe you.

Knower. (*stares incredulously*)⁴

Pusher. You see, knowing you're rational, I only have a reason to believe you if I have a reason to believe you've got a reason to tell the truth. But, as an act utilitarian, you don't care about truth-telling as such. The only reason you could have for telling me the truth is that you have a reason to believe that I would believe you and so push the right button. And, knowing I'm rational, you have no reason to believe I would believe you unless you have a reason to believe I have a reason to believe you. So you have no reason to believe I have a reason to believe you unless I have a reason to believe you.

Knower. And so, having a reason to believe me, you believe me. Good. Push green, and let's leave. (*sighs in relief*) For a moment there, you had me worried!

Pusher. Note that what I said is compatible with my having no reason to believe you. A reason to believe you would be baseless.⁵

Knower. All right, go on. Why would it be baseless?

Pusher. Okay. (*takes a deep breath*) My having a reason to believe you depends on my having a reason to believe you've got a reason to tell the truth. My having a reason to believe you've got a reason to tell the truth depends on your having a reason to tell the truth. Your having a reason to tell the truth depends on your having a reason to believe I would believe you. Your having a reason to believe I would believe you depends on your having a reason to believe I have a reason to

³ Narveson 1971, pp. 215–16 and Hardin 1988, p. 64.

⁴ Lewis 1973, p. 86; 1986, pp. 133–5.

⁵ Hodgson 1967, p. 44.

believe you. Your having a reason to believe I have a reason to believe you depends on my having a reason to believe you. (*catches breath*) Putting this together, we find that my having a reason to believe you depends on my having a reason to believe you. So, having come full circle, that alleged reason to believe you would be based on nothing.

Knower. So there's no reason to believe me?

Pusher. (*sotto voce*) There's no reason to believe you're even trying to communicate.⁶

Knower. Now, one reason for telling the truth is to gain an instrumentally valuable reputation as someone who cares about truth-telling.⁷

Pusher. Sure. If other people can be duped into thinking you care, then *great*. As for me, I know that you're an act utilitarian and thus that you don't care one bit.

Knower. Listen, you know me. You know I've been truthful in the past. And, if you have a reason to believe I'm following a truth-telling convention based on my having done so before, then you have a reason to believe me — and I, knowing this, have a reason to tell the truth. All this is compatible with act utilitarianism. So act utilitarianism can consistently require you to follow my advice.⁸

Pusher. Being act utilitarians, we don't care about the following of a convention just because it's been followed before.⁹ We're indifferent to mere convention. Though your streak of truth-telling may be remarkable, I know that you don't mind breaking it. And you know I know *that* and so on. And so we have no reason to care about your past truth-telling.

Knower. Let me see if I've got this. Supposing we had been in this situation a zillion times and we pulled through each time by me telling the truth and you following my advice, that *still* wouldn't move you towards doing the same this time?¹⁰ You wouldn't — in any way — see this as setting up a convention for our communications?

Pusher. Exactly. It's not puzzling that signalling conventions can arise and be kept in place if people with shared interests care about follow-

⁶ Regan 1980, p. 35.

⁷ Hodgson 1967, p. 46 and Singer 1972, p. 101.

⁸ Sartorius 1972, pp. 216–17; 1975, p. 72, Harris 1972, p. 347, and Gibbard 1978, pp. 95–102. See also Lewis 2020, pp. 248–50.

⁹ Piper 1978, pp. 198–9.

¹⁰ Hoerster 1973, p. 414.

ing conventions.¹¹ While we do have the same interests, we don't care about conventions.

Knower. We don't care about the old ways.

Pusher. Right. So I can consistently believe that you've always been truthful in the past but that you're untruthful now. And so act utilitarianism can consistently allow me to flout your advice.

Knower. Yet you agree that, *if* you believed me and I believed that you believed me and so on, then I would have a reason to tell the truth and you would have a reason to believe me. And then act utilitarianism would require me to tell the truth and you to heed my advice. So, by beginning to believe each other, we *create* the reasons for believing each other.¹² Isn't that beautiful? Why don't we just do that? I'll do my part.

Pusher. I love the initiative. I really do. One hurdle, though: unless I already believe that you believe I believe you, I have no reason to believe you. And, unless you already believe I believe that you believe I believe you, you have no reason to believe I believe you. So, acting individually, we have no reason to start believing each other.

Knower. But, if our not believing each other predictably leads to worse outcomes, wouldn't act utilitarianism prescribe that we believe each other?¹³

Pusher. Alas, act utilitarianism only prescribes voluntary acts. And it's, at best, unclear whether we can voluntarily choose what to believe.¹⁴

Knower. I choose to believe we can.

Pusher. Cute. *However*, we would still need our beliefs to align with the beliefs of the other. And we can't choose what the other believes. If I don't believe you and you don't believe I believe you, then — holding fixed what the other believes — neither of us can make it so I believe you and you believe I do. And then act utilitarianism wouldn't require us to believe each other.

Knower. Fine. But even so, given a suitable pattern of beliefs, act utilitarians *can* consistently believe each other. I suggest we simply believe the *Principle of Truthfulness* — the principle that other act utilitarians are truthful whenever it's best that one have true beliefs about those

¹¹ Skyrms 2010, pp. 9–12, 39–40. Cases without pure common interest are harder; see Crawford and Sobel 1982.

¹² James 1896, pp. 342–3; 1897, pp. 24–5; 1979, pp. 28–9.

¹³ Sumner 1969, p. 641.

¹⁴ James 1896, p. 330; 1979, pp. 15–16.

matters they know about. This principle is compatible with act utilitarianism.¹⁵ And, if we both believe it and both believe we both believe it and so on, we can communicate.

Pusher. That principle, though, is more or less the very thing I'm doubting. What would help is a reason to believe it.¹⁶

Knower. It's common sense.¹⁷

Pusher. As an act utilitarian, I tend to doubt common sense. Frankly, I find it biased towards deontology. Is it common sense to kill babies for the greater good?¹⁸ (*shakes head*) It is not.

Knower. Touché.

Pusher. And what kind of error am I making if I fail to believe your principle? Belief in that principle is hardly prescribed by rationality or by act utilitarianism. Consider the *Principle of Untruthfulness* — the principle that other act utilitarians are *untruthful* whenever it's best that one have true beliefs about those matters they know about. This principle is also compatible with act utilitarianism. And, if we both believe it and we both believe we both believe it and so on, we can likewise communicate. So why believe your principle rather than my alternative? As the choice between the principles is arbitrary, we have no reason to believe either.

Knower. The Principle of Untruthfulness conveys less information. If, at 5 p.m., you ask for the time and I lie, saying it's 3 p.m., how are you supposed to work out what time it is?¹⁹

Pusher. It *can* convey as much information — if you lie with specificity. For instance, 'It's not 5 p.m.'

Knower. That's longer than 'It's 5 p.m.' Seems inefficient.

Pusher. In that case, sure. But suppose you wish to convey a negated message. When you wish to convey that it's not 5 p.m., you need only say 'It's 5 p.m.'

Knower. 'That's a good example.'

Pusher. Thanks.

Knower. There's also the sheer inconvenience of the Principle of Untruthfulness. Truth-telling is far easier than lying. Lying involves so

¹⁵ Lewis 1972, pp. 18–19.

¹⁶ Österberg 2011, p. 187; 2019, p. 231.

¹⁷ Lewis 1972, p. 18.

¹⁸ Dostoevsky 1912, p. 258.

¹⁹ Singer 1972, pp. 98–9.

much invention.²⁰

Pusher. Just negate what you take to be true. That's not all that inventive. But, more to the point, we're indifferent to mere inconvenience. So we still lack a reason to favour your principle over mine.

Knower. That is inconvenient.

Pusher. Some truths are.

Knower. Can't we break this tie with the help of salience? Truthfulness, I think it's fair to say, is more salient than untruthfulness.²¹

Pusher. So what? We're indifferent to mere salience. We might as well communicate by coordinating on the non-salient principle.²²

Knower. Granted, that would be non-salient. (*pause*) Anyway, suppose that, in this case, there's no reason to tell the truth. All the same, there's no reason *not* to tell the truth. In this kind of deadlock, isn't there a natural human tendency in favour of telling the truth — making it at least slightly more likely that I do so? And then, given this slightly greater likelihood, there would be a reason for you to believe me. And then there would be a reason for me to tell the truth after all.

Pusher. Are you appealing to our common humanity? Remember, we are rational act utilitarians who act in light of reason — not by animal instinct.

Knower. I'm appealing to the part of us that is still human.

Pusher. Look, if we were just to let ourselves get carried away by human instinct whenever there's an exact tie in act-utilitarian reasons, it would be hard not to let that instinct get the best of us when lying would be a tiny bit better. So a human tendency in favour of truth-telling would lead to wrongdoing. That is, this kind of tiebreaker would break more than ties.

Knower. Okay, let's try a new tack. Either you believe me or you don't. If you believe me, it's better to tell you the truth. If you don't believe me, it doesn't matter what I tell you. But there's at least *some* chance you do believe me. So, to maximize expected value, I should tell you the truth. And you, by going through the same reasoning, should believe me.

Pusher. There's also the possibility that I believe you're lying. As I see it, I have no more reason to believe you than to believe you're lying.

²⁰ Bentham 1827, pp. 202–3 and Wilde 1889, p. 36; 2007, p. 74.

²¹ Gauthier 1975, p. 216. See also Schelling 1960, pp. 54–8 and Lewis 1969, p. 35.

²² Provis 1977, pp. 508–9.

So you have no reason to have more credence in my believing you're telling the truth than in my believing you're lying. And then the expected value of your telling the truth shouldn't exceed that of not doing so.

Knower. Is it crucial here that you have only two options? That is, is it crucial that, if I'm lying when I say it's the green button, then it has to be the red button? What if there had been a third button? Suppose that, in addition to the red and green buttons, there's a blue button. In this three-button case, when I say 'It's the green button,' the Principle of Truthfulness is more informative than the Principle of Untruthfulness. Note that I didn't say 'It's not the green button.' So I only conveyed something specific if I told the truth. And so, knowing I'm rational, you should conclude that I told the truth.

Pusher. In the three-button case, we would need to consider some further possibilities for what you may be up to. For example, when you say 'It's the green button,' you may be, first, plainly telling the truth; second, using a code where 'green' means 'red'; third, using a code where 'green' means 'blue'; and so on. Avoiding any dubious appeals to common sense, I have no more reason to believe the first possibility than the second or the third. So, in my choice between the three buttons, I'd still have no reason to believe you. But, thankfully, given our lack of trust, we're only dealing with two.

Knower. (*sighs*) All right. Enough is enough. I'm *asserting* that it's the green button.

Pusher. You're asserting it?

Knower. I am. And, according to a widespread norm, one may only assert what one knows.²³

Pusher. That norm conflicts with your act utilitarianism. I know you don't observe it.

Knower. The norm could be construed as a linguistic norm. And, as such, it would be compatible with act utilitarianism.²⁴

Pusher. Sure. But, if it's just a linguistic norm, then, by your act-utilitarian lights, there's no reason to follow it. We're indifferent to mere linguistics.

Knower. Couldn't the norm be motivated instrumentally?

Pusher. Then the question is whether — in this case — following the

²³ Unger 1975, p. 262 and Williamson 2000, p. 243.

²⁴ Mackie 1973, pp. 297–8 and Williamson 2000, p. 240.

norm would be better in expectation than not following it. That is, whether, in expectation, it would be better for you to be truthful. And then we're back where we started.

Knower. (*pause*) I'm starting to feel we're making this harder than it needs to be. Most people trust each other and communicate just fine.

Pusher. Most people aren't act utilitarians.

Knower. Yet, somehow, I manage to communicate with my other act-utilitarian friends.

Pusher. I suspect your other friends aren't really act utilitarians. I suspect that, in everyday situations, they simply follow common sense rather than their professed act utilitarianism. They haven't quite thought things through. The trouble is, we have.

Knower. Still, save for the button issue, we've been communicating quite well for close to ten minutes now.

Pusher. What was the other issue?

Knower. Look, time is running out. You need to push that green button.

Pusher. All this reminds me of the old problem of act utilitarianism not performing well in coordination games:²⁵ (*starts scribbling on the wall*)

		You	
		Tell the truth	Lie
Me	Follow advice	<i>We live</i>	<i>We die</i>
	Flout advice	<i>We die</i>	<i>We live</i>

Ignorant of what the other player will do, we have no reason to choose one option over the other. Do you know about Nash equilibria?

Knower. Yes. But, again: (*points to watch*) Push the button.

Pusher. Great. But, in case you need a refresher, a *Nash equilibrium* in a game is a combination of strategies for each player where no player can do better by changing their strategy, holding fixed the strategies of the other players.²⁶ In our case, there are three Nash equilibria: one where you tell the truth and I follow the advice, one where you lie and I flout the advice, and one where we both randomize evenly between our options. Act-utilitarian reasons alone do not single out the truth-telling equilibrium. Usually, these coordination problems are thought to disappear if the participants are able to communicate. But, as we've

²⁵ Schelling 1960, p. 294, Gibbard 1965, p. 218, and Parfit 1986, p. 867.

²⁶ Nash 1950, p. 49.

seen, proper act utilitarians — knowing they're both rational act utilitarians — can't reliably communicate with each other.

Knower. *Please, time's almost up.*

Pusher. You know what? This does seem like a problem for act utilitarianism. I guess the reason it's not more apparent is that most professed act utilitarians don't really follow the theory. But the fact that people don't follow the theory is hardly a defence of it. Or what do you think?

Knower. PUSH GREEN NOW.

Pusher. Nevertheless, I have to say that act utilitarianism, despite its problems, still has a lot going for it. It's a beautiful theory. For the record, I'd like to assert that my conviction is, on balance, unshaken.

Knower. PUSH THE GREEN BUTTON OR WE'RE GOING TO DIE!

Pusher pushes the red button. The cells unlock, and the happy pair are off to the pub.

I wish to thank Krister Bykvist, Ray Buchanan, Daniel Drucker, Petra Kosonen, Harvey Lederman, Wlodek Rabinowicz, Dean Spears, Christian Tarsney, and Torbjörn Tännsjö for valuable comments.

References

- Bentham, Jeremy (1789) *An Introduction to the Principles of Morals and Legislation*, London: T. Payne.
- (1827) *Rationale of Judicial Evidence, Specially Applied to English Practice*, vol. I, London: Hunt and Clarke.
- (1970) *An Introduction to the Principles of Morals and Legislation*, eds. J. H. Burns and H. L. A. Hart, *The Collected Works of Jeremy Bentham*, London: Athlone.
- Crawford, Vincent P. and Joel Sobel (1982) 'Strategic Information Transmission', *Econometrica* 50 (6): 1431–1451.
- Dostoevsky, Fyodor (1912) *The Brothers Karamazov*, ed. Constance Garnett, London: Heinemann.
- Gauthier, David (1975) 'Coordination', *Dialogue* 14 (2): 195–221.
- Gibbard, Allan (1965) 'Rule-Utilitarianism: Merely an Illusory Alternative?', *Australasian Journal of Philosophy* 43 (2): 211–220.
- (1978) 'Act-Utilitarian Agreements', in Alvin I. Goldman and Jaegwon Kim, eds., *Values and Morals: Essays in Honor of William Frankena, Charles Stevenson, and Richard Brandt*, pp. 91–119, Dordrecht: Reidel.

- Hardin, Russell (1988) *Morality within the Limits of Reason*, Chicago: University of Chicago Press.
- Harris, N. G. E. (1972) 'Nondeliberative Utilitarianism', *Ethics* 82 (4): 344–348.
- Hodgson, D. H. (1967) *Consequences of Utilitarianism: A Study in Normative Ethics and Legal Theory*, Oxford: Clarendon Press.
- Hoerster, Norbert (1973) 'Is Act-Utilitarian Truth-Telling Self-Defeating?', *Mind* 82 (327): 413–416.
- James, William (1896) 'The Will to Believe', *The New World* 5 (18): 327–347.
- (1897) *The Will to Believe and Other Essays in Popular Philosophy*, New York: Longmans Green and Co.
- (1979) *The Works of William James: The Will to Believe and Other Essays in Popular Philosophy*, eds. Frederick H. Burkhardt, Fredson Bowers, and Ignas K. Skrupskelis, Cambridge, MA: Harvard University Press.
- Lewis, David (1969) *Convention: A Philosophical Study*, Cambridge, MA: Harvard University Press.
- (1972) 'Utilitarianism and Truthfulness', *Australasian Journal of Philosophy* 50 (1): 17–19.
- (1973) *Counterfactuals*, Cambridge, MA: Harvard University Press.
- (1986) *On the Plurality of Worlds*, Oxford: Blackwell.
- (2020) *Philosophical Letters of David K. Lewis Volume 2: Mind, Language, Epistemology*, Oxford: Oxford University Press.
- Mackie, J. L. (1973) 'The Disutility of Act-Utilitarianism', *The Philosophical Quarterly* 23 (93): 289–300.
- Narveson, Jan (1971) 'Promising, Expecting, and Utility', *Canadian Journal of Philosophy* 1 (2): 207–233.
- Nash, John F. Jr. (1950) 'Equilibrium Points in n -Person Games', *Proceedings of the National Academy of Sciences of the United States of America* 36 (1): 48–49.
- Österberg, Jan (2011) 'Is Utilitarianism Self-Defeating?: Hodgson and His Critics', in Rysiek Sliwinski and Frans Svensson, eds., *Neither/Nor: Philosophical Papers Dedicated to Erik Carlson on the Occasion of His Fiftieth Birthday*, pp. 167–207, Uppsala: Uppsala University.
- (2019) *Towards Reunion in Ethics*, Berlin: Springer.
- Parfit, Derek (1986) 'Comments', *Ethics* 96 (4): 832–872.
- Piper, Adrian M. S. (1978) 'Utility, Publicity, and Manipulation', *Ethics* 88 (3): 189–206.
- Provis, C. (1977) 'Gauthier on Coordination', *Dialogue* 16 (3): 507–9.
- Regan, Donald (1980) *Utilitarianism and Co-Operation*, Oxford: Clarendon Press.

- don Press.
- Sartorius, Rolf E. (1972) 'Individual Conduct and Social Norms: A Utilitarian Account', *Ethics* 82 (3): 200–218.
- (1975) *Individual Conduct and Social Norms: A Utilitarian Account of Social Union and the Rule of Law*, Encino, CA: Dickenson.
- Schelling, Thomas S. (1960) *The Strategy of Conflict*, Cambridge, MA: Harvard University Press.
- Singer, Peter (1972) 'Is Act-Utilitarianism Self-Defeating?', *The Philosophical Review* 81 (1): 94–104.
- Skyrms, Brian (2010) *Signals: Evolution, Learning, & Information*, Oxford: Oxford University Press.
- Sumner, L. W. (1969) 'Consequences of Utilitarianism', *Dialogue* 7 (4): 639–642.
- Unger, Peter (1975) *Ignorance: A Case for Scepticism*, Oxford: Clarendon Press.
- Wilde, Oscar (1889) 'The Decay of Lying: A Dialogue', *The Nineteenth Century* 25 (143): 35–56.
- (2007) *The Complete Works of Oscar Wilde Volume 4: Criticism: Historical Criticism, Intentions, The Soul of Man*, ed. Josephine M. Guy, Oxford: Oxford University Press.
- Williamson, Timothy (2000) *Knowledge and Its Limits*, Oxford: Oxford University Press.