Bentham's Mugging

Johan E. Gustafsson

ABSTRACT. A dialogue, in three parts, on utilitarian vulnerability to exploitation.

Mugger. Excuse me a moment, would you sir? I'm a bit short on cash.

Bentham. Sorry.

Mugger. *(notices a utilitarian pin on* **Bentham***'s lapel)* But you're a utilitarian, right?

Bentham. Indeed, I am an *Act Utilitarian*: I believe I ought to perform an act if that act would produce more utility than any alternative act.[1]

Mugger. That's grand. How about you give me one hundred pounds?

Bentham. Now, as an Act Utilitarian, I would happily part with a hundred *if* I were convinced that you would bring more utility to the world with that money than I would. The trouble is I know I would put the money to good use myself—whereas you, I surmise, would not.

Mugger. Fine. I suspected as much. But what if I sweeten the deal? If you don't give me the money, I'll cut off a finger!

Bentham. You're threatening me?!

Mugger. Wait, no. I'm not threatening *you*. That would be illegal. I'm saying that, if you don't give me the money, I'll cut off *my* finger.

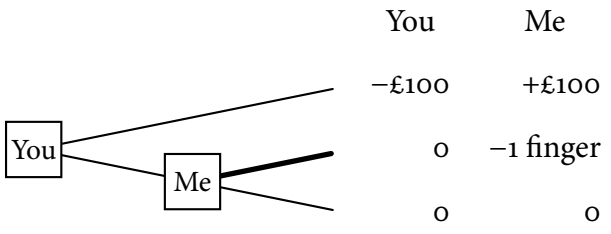
Bentham. Why on earth would you do that?

Mugger. I am a *Deontologist*. I am true to my word.

Bentham. *(notices a deontological pin on* **Mugger***'s lapel)* I see. But then, if you don't mind my asking, why did you promise to cut off your finger?

Mugger. Look, what happened happened. Let's cut to the chase. I have diagrammed our situation:
(unfolds a large poster)

[1] Bentham 1970, pp. 12–13 and Moore 1912, p. 31.



Bentham. You had that diagram ready all along?

Mugger. You, at the first node, have a choice between going up (giving me the money) or down (keeping the money). If you go down, I have a choice at the second node whether or not to cut off my finger. The thick line denotes that I would, in fact, do so.

Bentham. *(nods)*

Mugger. Here's the thing: there is, clearly, more utility in me keeping my finger than in you keeping your measly one hundred pounds. So there would be *more* utility in the world if you gave me the money than if you didn't.

Bentham. I think you should just keep your finger.

Mugger. And go back on my word? No. *(chuckle)* What if everyone did that? Besides, what should matter to you, as an Act Utilitarian, is that I *would* cut off my finger—not whether I should.

Bentham. Even so, I worry that giving you the money would set a bad precedent, encouraging copycats running similar schemes.

Mugger. Don't. This transaction will be our little secret. You have my word.

Bentham. *(not entirely convinced)*

Mugger. You're playing hardball? *(sigh)* All right, let's make the deal sweeter still: If I don't get the money, I'll cut off *two* fingers.

Bentham. This conversation has sure taken a regrettable turn.

Mugger. I'm sure going to miss those fingers.

Bentham. *(pause)* OK. Fine.

(hands over £100)

Mugger. Excellent.

(pockets the money and folds, carefully, the poster)

Bentham. I somehow feel I got mugged.

Mugger. Not at all. You made the world a better place.

* * *

Mugger. *(sees* **Bentham***)* It's been a while, hasn't it?

Bentham. Oh, … hi.

Mugger. What's the matter?

(notices **Bentham***'s unadorned lapel)*

Where's your pin?

Bentham. Alas, wearing it in public became too dear.

(notices **Mugger***'s hand)*

What happened to your hand?

Mugger. Funny you should ask. It turns out that some so called "Act Utilitarians" are Act Utilitarian in name only. To cut a long story short, fingers were … cut. Getting them sewed back cost me a fortune. So, once more, I find myself short on cash.

Bentham. Sad. Very sad to hear. But, before you reattempt your scheme, I'd like to share some news. I'm no longer an Act Utilitarian. I'm now a *Rule Utilitarian*: I believe I ought to perform an act if that act is an instance of a rule that prescribes a possible combination of everyone's acts that produces more utility than any other combination.[2]

Mugger. I like this! It feels almost deontological.

Bentham. And, crucially, I'm no longer susceptible to your finger scheme. A plausible moral theory shouldn't lay one open to that kind of exploitation. I wonder why I never saw this fault in utilitarian thinking. But no matter—the theory is fixed now.

Mugger. May I suggest a collaboration?

Bentham. Sure.

Mugger. There's this new course called *Effective Benevolence: Morality Made Easy*.[3] If I took this course, I would become an effective altruist just like you.

Bentham. Sounds great.

Mugger. The trouble is the course costs one hundred pounds. And here's where you could help out. Would you contribute the money to let me realize this dream?

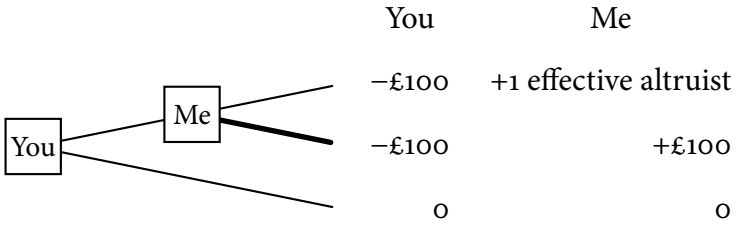
Bentham. Tempting.

Mugger. I've diagrammed our new situation:

(unfolds another large poster)

[2] Urmson 1953, p. 35.

[3] Bentham's (1983, p. 119) alternative title for *Deontology* was *Morality Made Easy: Shewing How throughout the Whole Course of Every Person's Life Duty Coincides with Interest Rightly Understood Felicity with Virtue Prudence Extra-Regarding as Well as Self-Regarding with Effective Benevolence*.



At the first node, you have a choice between giving me the money (going up) or keeping it (going down). If you go up, I have a choice between using the money to take the course (going up) or keeping the money for myself (going down). Surely, me becoming an effective altruist is better than you keeping your one hundred pounds. Hence a rule prescribing your giving me the money and my using that money to take the course is a rule that prescribes the best possible combination of everyone's acts.

Bentham. So I should give you the money.

Mugger. *(noticeably impressed by* **Bentham***'s deduction)* I just love working with sharp minds.

Bentham. Wait—why is one line thicker than the others?

Mugger. Oh, that denotes that I wouldn't take the course. If you gave me the money, I would in fact keep it for myself.

Bentham. You left that datum out of your pitch.

Mugger. I don't see how it would be relevant for a Rule Utilitarian. Your giving me the money would be part of the best possible combination of everyone's acts no matter whether I would take the course. What should matter to you is that I *could* do so—not whether I would.

Bentham. Hmm, I am starting to think that it was a mistake to assess rules by their being adopted by *everyone*. I'm now a *Partial-Compliance Rule Utilitarian*: I believe that I ought to perform an act if that act is required by the code of rules that is optimal in the sense that its internalization by the *overwhelming majority* would be best.[4]

Mugger. *(clears throat)* I have an announcement to make. I'd like to make it known that, if a code of rules were internalized by the the overwhelming majority, I would internalize it too.

Bentham. You would?

Mugger. At that point, I feel, it would be antisocial not to.

[4] Hooker 2000, p. 32.

Bentham. So, when I assess different codes of rules, I should assess their being internalized by the overwhelming majority including you?

Mugger. That's right.

Bentham. But, since the optimal code of rules won't actually be internalized by the the overwhelming majority, you won't actually internalize that code of rules.

Mugger. Well, yeah.

Bentham. So, even though the optimal code of rules would, plausibly, prescribe me giving you the money and you taking the course (since that code is optimal given that it's internalized by the overwhelming majority including you), you would not take the course.

Mugger. Uh-huh.

Bentham. I'm getting second thoughts about Rule Utilitarianism all together.

Mugger. Very well. So you are going back to Act Utilitarianism? In that case, let me offer you a deal—

Bentham. Let me cut in right here. I now think I'm a *Self-Harm-Discounting Act Utilitarian*: I believe that I ought to perform an act if that act would produce more utility than any alternative act with utility measured so that saving people from harm does not count towards utility if these people can save themselves.[5]

Mugger. This is a major departure from standard Act Utilitarianism.

Bentham. True. But, with this modification, Act Utilitarianism is immune to your finger scheme. Since you could still avoid cutting off your finger in case I don't give you money, that avoidable harm does not count towards overall utility.

Mugger. Could you stick around a bit? I need to run a quick errand.

Bentham. No worries.

Mugger. And, just to double-check, when you say that harms don't count if people can avoid them themselves, you mean harms that people can *still* avoid themselves? That is, are you a *Retrospective Self-Harm-Discounting Act Utilitarian*, believing that you ought to perform an act if that act would produce more utility than any alternative act with utility measured so that saving people from harm does not count towards utility if these people can save themselves *or could have saved themselves if they had chosen otherwise in the past*?

[5] Graham 2020, pp. 177–178.

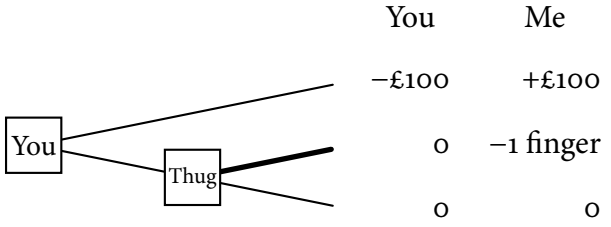
Bentham. No—I'm not a monster. We have all made mistakes. The moral agent looks forward. If I found you drowning in a pond, I should save you regardless of whether you went in freely.[6]

Mugger. Great, just as I thought. Stay put. I'll be back in a jiffy.

* * *

Bentham. What took you so long?

Mugger. Sorry, I had to make a binding, unalterable arrangement with a thug, who will cut off my finger if you don't give me one hundred pounds. I've diagrammed our current predicament:
(unfolds a third poster)



Like before, you have a choice at the first node between going up (giving me the money) or down (keeping the money). If you go down, the thug has a choice at the second node whether or not to cut off my finger—and the thug would do so, no matter what I do. If you don't give me the money, I can't avoid being harmed. And, while it's true that I could have avoided this predicament had I chosen differently earlier, you wouldn't want to rule out harms people may suffer due to their past choices; that would rule out too many harms. The thug is bound by an unalterable arrangement to cut my finger in case you don't pay. So my finger is in your hand, so to speak. *(chuckle)*

Bentham. Funny you should say that. While you were away, I had some time to reflect on morality. I now think I'm more of a *Mugging-Restricted Act Utilitarian*: I believe that I ought to perform an act if that act would produce more utility than any alternative act and, in addition, it wouldn't make me vulnerable to blatant muggings, threats, or blackmail.

Mugger. I don't want to rag on your new philosophy, but where's the theoretical purity of standard Act Utilitarianism? This theory is soiled with muddy, ambiguous terms. What, more precisely, is a 'mugging'?

[6] Singer 1972, p. 231.

Bentham. I know one when I see one. And your latest scheme, I'm pretty sure, is one.

Mugger. That's not a very satisfying answer.

Bentham. I'm afraid it will have to do for now.

Mugger. Also—and I hate to say this—your latest theory is, *more than a little*, ad hoc.

Bentham. Well, what it lacks in beauty, it makes up in expense minimization.

Mugger. Look, even if your theory tells you what you ought to do, it lacks *explanatory power*. A moral theory may tell us not only what ought be done but *why* it ought be done. Why settle for less?

Bentham. If I come up with an unmuggable version of utilitarianism with more explanatory power, I'll let you know.

(A **Thug** *approaches.)*

Mugger. This is disappointing.

Bentham. *I'm not going to give you more money.*

Mugger. OK. Fine. I'll cut off *three* fingers if—

Bentham. Sorry, but here I must cut you off.

I wish to thank Krister Bykvist, Tomi Frances, Will Jefferson, Kacper Kowalczyk, Petra Kosonen, Andreas Mogensen, Martin Peterson, Wlodek Rabinowicz, Dean Spears, and Torbjörn Tännsjö.

References

Bentham, Jeremy (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.

— (1983) *Deontology Together with A Table of the Springs of Action and the Article on Utilitariansim*, The Collected Works of Jeremy Bentham, Oxford: Clarendon Press.

Graham, Peter A. (2020) 'Avoidable Harm', *Philosophy and Phenomenological Research* 101 (1): 175–199.

Hooker, Brad (2000) *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality*, Oxford: Clarendon Press.

Moore, G. E. (1912) *Ethics*, London: Williams & Norgate.
Singer, Peter (1972) 'Famine, Affluence, and Morality', *Philosophy & Public Affairs* 1 (3): 229–243.
Urmson, J. O. (1953) 'The Interpretation of the Moral Philosophy of J. S. Mill', *The Philosophical Quarterly* 3 (10): 33–39.