Pascal's Mugger Strikes Again

Dylan Balfour, University of Edinburgh
dylan.balfour@ed.ac.uk

Abstract

In a well-known paper, Nick Bostrom presents a confrontation between a fictionalised Blaise Pascal and a mysterious mugger. The mugger persuades Pascal to hand over his wallet by exploiting Pascal's commitment to expected utility maximisation. He does so by offering Pascal an astronomically high reward such that, despite Pascal's low credence in the mugger's truthfulness, the expected utility of accepting the mugging is higher than rejecting it. In this paper, I present another sort of high value, low credence mugging. This time, the mugger utilises research on existential risk and the long-term potential of humanity to exploit Pascal's expected-utility-maximising descendant. This mugging is more insidious than Bostrom's original as it relies on plausible facts about the long-term future, as well as realistic credences about how our everyday actions could, albeit with infinitesimally low likelihood, affect the future of humanity.

In a quiet pub in the present day…[1]

Mugger: That's a nice-looking pint you have there. You should buy one for me.

Pascal: Excuse me, who are you?

Mugger: Oh, I'm just someone passing by. I recognise you, though. If I'm not mistaken, you are a distant progeny of the great Blaise Pascal?

Pascal: What? How on Earth did you know that?

[1] This paper owes its structure and broad conceit to Bostrom (2009). The term "Pascal's Mugging" was initially coined by Eliezer Yudkowsky (2007) on the *Less Wrong* forums.

Mugger: That's none of your concern. What *should* be your concern is that you're about to condemn humanity to extinction.

Pascal: Come again?

Mugger: If you don't do what I tell you, humanity will perish within a matter of years.

Pascal: I see what's happening. You're trying to do a *Pascal's Mugging* on me, aren't you?

Mugger: I've never heard of such a thing! I'm just trying to warn you of the grave consequences of what you're about to do. I implore you, please heed my advice.

Pascal: You must think me a fool.

Mugger: Quite the contrary, Mr. Pascal. I think you're just as rational as your great, great, great, great, great, great, great, great…

Pascal: Take a breath.

Mugger: …great, great grandfather was. Phew. Anyway, am I correct in saying that you, like your distant ancestor, are an expected utility maximiser?

Pascal: That is correct.

Mugger: And that, like your distant namesake, your utility function is also aggregative in terms of happiness?

Pascal: Again, that's right. And, furthermore, my utility function is unbounded, and I do not subscribe either to risk aversion or temporal discounting. What's more, I'm totally impartial!

Mugger: Perfect. And I see you're an equally engaging conversationalist, too. Tell me, Mr. Pascal, how many years do you think the Earth can sustain human life for?

Pascal: I can't say I've thought about it, to be honest.

Mugger: Not to worry, I've printed out some literature for you. Take a look, if you wouldn't mind.

The Mugger hands Pascal a pile of paper containing the following excerpts: Bostrom (2013: 15-9) Ord (2020: 296-7), and Beckstead (2019: 81). Pascal spends a few minutes leafing through them at the bar.

Pascal: Okay, I've read them. I had a hunch you'd fabricated these yourself, but I've checked them out online, too. It seems that the Earth will remain habitable for around one billion years.

Mugger: Very good. Now assume that, barring an extinction event or civilisational collapse, there would be an average of one billion humans on Earth until it becomes uninhabitable.

Pascal: That seems a reasonable assumption.

Mugger: So if humanity doesn't go extinct, there is a potential for many, many, billions of humans to exist on Earth in the future? You'll see that the literature estimates that around *ten quadrillion* individuals could come to exist. And should we assume that, for the vast majority of these future people, that their lives will be worth living?

Pascal: Again, I will assent. Can you get to the point, please?

Mugger: Well, as I said, *you* are about to prevent all of this future value from coming into existence.

Pascal: That is preposterous.

Mugger: From your perspective, perhaps. But listen to what I have to say. There is a sinister, secretive organisation currently preparing a deadly and highly contagious disease. They're close to completion, and will begin manufacturing the disease in huge quantities very shortly. What's worse is that the disease will be airborne, and will spread so quickly that the surface of the Earth will be rendered uninhabitable in as little as two years. Humanity will not survive, and all those many billions of future people will never exist.

Pascal: What a disturbing fib to tell a stranger!

Mugger: I wish it were a lie. But if you don't prevent yourself from doing what you're about to do, then the fate of humanity will be sealed. Besides, I must be here telling you this for a reason. You ought to assign the idea *some* credence.

Pascal: Fine. I'll give it a *one in a billion* credence. I know that it's technically possible to manufacture deadly diseases, but I'm highly sceptical that there's a malicious organisation with the capability of doing so, nor do I think you'd know about it even if there were. No offence, but strangers in pubs are seldom reliable sources of geopolitical knowledge. However, I am curious about where you're going with this. What is this supposedly calamitous action I was about to perform?

Mugger: You were going to refrain from buying me a pint.

Pascal: What? Wait, so you're seriously trying to tell me that if I buy you a pint, the future of humanity will be saved?

Mugger: Look, I know it sounds far-fetched, but hear me out. You see, the brewery who makes that delicious ale you're currently drinking have another wing to their business. They also develop cutting-edge vaccines. The brewing is just a way to fund the vaccine research. And trust me, they're at the top of their game. No other research organisation is close to developing a vaccine for the coming plague.

Pascal: And so my buying you a pint of their beer will help how, exactly?

Mugger: It's quite embarrassing, really. The vaccine plant is behind on their electricity bills. They need another £2 or else their supplier will shut the factory off, preventing the vaccine from being produced in time to counter the plague. Buying another beer will provide them with enough money to pay for another month of electricity, and the vaccine will be completed. I also happen to know this is currently the only pub which stocks this brewery's beer, so no-one else will be able to buy this vital pint. What do you say?

Pascal: I say that it's still utterly ludicrous.

Mugger: I understand that you're sceptical. But you know the drill. *Precisely* how ridiculous do you think it is?

Pascal: Oh, I don't know. If you're telling the truth about the plague, then it's more likely you'll be correct about this further claim. But I still find it deeply implausible that a brewery could also

function as a vaccine company, and that a measly £2 could be the difference between their vaccine being produced or not. I'll say *one in a million* Besides, if it's so important, why don't you buy the pint?

Mugger: I left my wallet at home. That's why I'm asking you. I know that you want to maximise expected utility just as I do, so I think you'll make the right decision.

Pascal: You're going to have to do more than that to convince me. I don't see how buying this pint could possibly yield a substantial amount of expected utility. You heard the credences I placed in your claims. My credence that they are jointly true is *one in a billion multiplied by one in a million.*

Mugger: I suppose I'm going to have to spell this out. Remember, if this plague is prevented, then humanity could live on Earth for another billion years, sustaining a billion lives at any one time. You saw it yourself in the papers I gave you—there could be *ten quadrillion* average human lives in this future. If this plague is unleashed, then all of this value will be lost.

Pascal: Plague or not, buying you a pint will make no difference.

Mugger: To be sure, from your perspective it almost certainly *won't* make a difference. But, using the credences you gave me, you think there is a precisely a *one in a quadrillion* chance that I am telling the truth about all of this. Now, by your own admission, you appear to have a *one in a quadrillion* chance at saving ten quadrillion lives. Even working by your own credences, you'd be saving an expected ten lives' worth of value. Just by buying me a pint! You'd be a hero!

Pascal: You've mugged me, haven't you.

Mugger: I'm afraid it seems that way.

Pascal: Wait, hold up a second. I think I see a way out. Even if I buy a pint and avert this existential catastrophe, there could still be more existential catastrophes lying in wait. I'm not securing the *entire* future of humanity—we could still go extinct later! If we end up going extinct before the Earth becomes uninhabitable in a billion years, then the only value I'd be securing would consist of the lives which would exist up until then.

Mugger: And what probability do you think there is of humanity going extinct at some point before the billion years is up?

Pascal: I'm almost certain it will happen. I'd give it a probability of 99%.

Mugger: So the expected value of buying the pint is now only *one hundredth* of what it was.

Pascal: I guess it is.

Mugger: So, buying me a pint still has an expected value of zero-point-one lives.

Pascal: That hardly makes me a hero.

Mugger: I'm not so sure about that. There's what, twenty thousand or so days in the average life? Buying a pint would still be worth two-hundred days. It still seems like a *highly* laudable act of generosity.

Pascal: Fine. You've got the better of me. Enjoy your drink.

Pascal orders a pint for the Mugger. He turns to leave, but the Mugger taps him on the shoulder.

Mugger: Not so fast! I'm not done with you yet. I can't let you go without a further grave warning, I'm afraid.

Pascal: You can't be serious. What is it?

Mugger: Once you leave this pub, you're going to be mugged every waking moment for the rest of your life.

Pascal: Good grief, that's quite the threat.

Mugger: It is no threat, Mr Pascal, it is fact!

Pascal: So, you're going to follow me everywhere I go? I don't know how you think that will work. You know about restraining orders, right? Besides, those sinister robes of yours don't look as though they permit much leg movement. I could escape you with a gentle jog.

Mugger: Quite right, Mr Pascal, it would be foolish of me to pursue you myself. I will not be personally exploiting your decision-theoretic commitments any longer. In fact, no one will *per se*. You are going to be mugged, as it were, by the future of humanity itself!

Pascal: Blimey. You're going to have to explain this quickly, my evening church service is starting soon.

Mugger: So be it. You bought me this pint because it carried an infinitesimal chance at preventing an existential catastrophe, correct?

Pascal: Correct.

Mugger: So, if you have some evidence which tells you an action has some tiny chance at preventing an existential catastrophe, that action's expected utility is going to be high. Higher, probably, than your other available actions. Correct?

Pascal: I suppose so.

Mugger: Well unfortunately, Mr Pascal, this evidence is everywhere. Every action carries with it *some* risk of destroying humanity. Every step you take, every sentence you utter, even every mote of dust you disturb could spell the end for our species. As an expected utility maximiser, you must perform the action which seems least likely to condemn us to extinction. Considerations of existential risk are going to almost outweigh the proximate effects of your actions.

Pascal: That's just absurd.

Mugger: It is! But, unfortunately for you, absurdity is the currency of muggings.

Pascal: Where does this evidence even come from? I can't see how any of my actions might feasibly destroy humanity. The only reason I bought you a pint is because you explicitly *told* me that refraining from doing so would be disastrous.

Mugger: How myopic. Use your imagination, Mr Pascal. Catastrophic risks abound. You just need to know how to look for them.

Pascal: How does one *look* for a catastrophic risk?

Mugger: Let's walk through an example. Imagine you're deciding between two shops in which to purchase your weekly groceries. Each shop, I assume, requires you to drive a different route. Well, Mr Pascal, you might want to think about which route takes you closer to any important diplomatic buildings. You see, if you take the road closest to an important international institution, there's a slightly higher chance that you'll be travelling along the same route as an important envoy. Now, this envoy might be on the verge of brokering a crucial peacekeeping treaty which significantly reduces the threat of nuclear war. Humanity's future would be put at risk were you to collide with this envoy and prevent the treaty from coming to fruition. This will not do. You'll have to drive to the other shop, Mr Pascal. As an avowed expected utility maximiser, I'm sure you wouldn't take the risk.

Pascal: But that's just a far-flung story you made up.

Mugger: Yes, it is. But it is also in the realm of possibility. You ought to give it some credence, and this credence will no doubt be sufficient to contribute higher expected utility to shopping further away from the important diplomatic building.

Pascal: But surely, we can tell a similarly plausible story for how going to the *other* shop carries just as much existential risk. The expected utilities conferred by existential risk will just balance out, leaving me free to choose the shop which is best for me.

Mugger: Go on, then. Tell me a story about the other shop.

Pascal: Okay. Let's say that the shop closer to the important diplomatic building uses a light blue colour for its branding. Its logo and all its own-brand product packaging is covered in this delicate, calming hue. Let's say that the other shop uses a bright shade of orange for its branding, and all of *its* packaging is covered in this fierce, stimulating colour. Now, it is clear to common sense that colours can affect one's mood and behaviour. By spending one's time surrounded by bright orange, one will be slightly likelier to lash out later that day. But by spending one's time surrounded by light blue, one will be slightly less likely to do so. Now, by shopping at the orange shop, I marginally increase its long-term financial viability, and likewise with the blue shop. Thus,

by going to the blue shop, I make it more probable that the blue shop will continue to serve customers into the future than if I go to the orange shop. Furthermore, I assume that important political decision makers often go shopping as the rest of us do. By giving my custom to the blue shop, I make it more likely that some future influential figure will shop there instead of the orange shop, and thereby refrain from making some impulsive, rash decision they would otherwise have made. This decision may have led to global instability, increasing the risk of a host of existential threats. Although incredibly unlikely, this contributes fairly substantial expected utility to shopping *closer* to the important diplomatic building, despite the risk of colliding with an envoy. How about that for an absurd story?

Mugger: That was impressive, Mr Pascal! And, indeed, I agree with you. We can tell a huge number of similar stories about the potentially catastrophic consequences of all our actions. But do you really think this means the existential risk associated with all our actions is *precisely* the same?

Pascal: Well, perhaps not exactly the same, but it might as well be. Given how many tiny factors are at play, it will be impossible to ascertain just how much existential risk each of our actions carry. We should just treat them as if they are equal, so that their expected utilities are determined by their upfront effects.

Mugger: To do that would be an error, Mr Pascal. The fact that each of our actions may lead to existential catastrophe does not entail that each action is *equally likely* to bring about disaster. Even the smallest difference in probability is enough to prefer one action to another. And just because we can tell a catastrophic story about each of our actions does not entail that these stories are equally plausible. I'm afraid you can't get off the hook that easily.

Pascal: Good grief. So, what does this mean for me?

Mugger: As I said before, you'll need to maintain constant vigilance, thinking constantly about which of your actions is least likely to destroy humanity. This action will almost always end up having the highest expected utility. Actions which seem good at first might, on further consideration, be among the very worst of your available actions. Likewise, actions which seem

intuitively distasteful may carry the most expected utility. You might even have to commit murder if that's how the expected utilities fall. Are you prepared to kill, Mr Pascal?

Pascal: Stop it! Is there really no way I can escape this? Please help me here. There must be some argument I can make recourse to.

Mugger: Well, you could, of course, drop your commitment to maximising expected utility.

Pascal: How dare you even suggest that! I come from a long and esteemed lineage of devoted maxi—

Mugger: I was joking.

Pascal: It wasn't funny. You've ruined my life. Now, if you'll excuse me, I'm leaving. I'm off to maximise expected utility the old-fashioned way: by expressing my devotion to a God I don't really believe in.

Mugger: Okay, but be careful on your way out. The future of humanity is at stake.

Bibliography

Beckstead, Nick. 2019. A Brief Argument for the Overwhelming Importance of Shaping the Far Future'. in Greaves, Hilary and Pummer, Theron (eds). *Effective Altruism: Philosophical Issues*. (Oxford: Oxford University Press, 2019), pp. 80-98

Bostrom, Nick. 2009. Pascal's Mugging. *Analysis* 69(3): 443-5

Bostrom, Nick. 2013. Existential Risk Prevention as Global Priority. *Global Policy* 4(1): 15-31

Ord, Toby. 2020. *The Precipice: Existential Risk and the Future of Humanity*. (London: Bloomsbury)

Eliezer Yudkowsky. 2007. Pascal's Mugging: Tiny Probabilities of Vast Utilities. *Less Wrong*. Available at: https://www.lesswrong.com/posts/a5JAiTdytou3Jg749/pascal-s-mugging-tiny-probabilities-of-vast-utilities [accessed 20th August 2020]