The Monist, 2024, **107**, 146–156
https://doi.org/10.1093/monist/onae004
Article

OXFORD

Risk, Non-Identity, and Extinction

Kacper Kowalczyk*, and Nikhil Venkatesh**,

ABSTRACT

This paper examines a recent argument in favour of strong precautionary action—possibly including working to hasten human extinction—on the basis of a decision-theoretic view that accommodates the risk-attitudes of all affected while giving more weight to the more risk-averse attitudes. First, we dispute the need to take into account other people's attitudes towards risk at all. Second, we argue that a version of the non-identity problem undermines the case for doing so in the context of future people. Lastly, we suggest that we should not work to hasten human extinction, even if significant risk aversion is warranted.

Recently, the following argument has entered philosophical discussion: (1) since multiple risk-attitudes are permissible, we should take into account the risk-attitudes of everyone our action will affect, giving more weight to the more risk-averse attitudes in some way; (2) so, when our action affects large numbers of future people who might be risk averse, our risk attitude should be significantly risk averse; (3) so, in these cases, we should support strong precautionary action. Richard Pettigrew goes as far as entertaining the idea that working to hasten human extinction is warranted in order to foreclose serious risks that might materialise in the future.[1]

In this paper, we examine this argument, taking issue with all three steps. First: unless we are elected representatives or legal proxies, there is little reason to take into account other people's risk-attitudes at all—especially if, as we argue, one's risk-attitude cannot plausibly be taken to express what's in one's best interest. Moreover, giving more weight to risk aversion seems undermotivated and at odds with intuition in some cases. Second: taking other people's risk-attitudes into account threatens to become unworkable in cases where our decisions affect future people. This is because of a version of the non-identity problem, whereby our actions can affect not just the identity of future people but also their risk-attitudes. Third: even if significant risk aversion were warranted, it would not follow that we should work to hasten human extinction. It might be preferable to work on directly mitigating risks that humanity will face in the future rather than trying to foreclose them by hastening humanity's demise.

*University College London, UK
**London School of Economics, UK

1. RISK

The argument begins with risk-weighted expected-utility theory, a nonstandard decision theory according to which a rational person should maximise the risk-weighted expectation of a numerical utility function, thus possibly giving greater weight to relatively worse possible outcomes, depending on the person's risk-attitude.[2]

Lara Buchak takes a person's risk-attitude to reflect that person's own judgment about how to pursue their goals under uncertainty. For example, a risk-averse attitude reflects the judgment that it is better to focus on securing a good worst outcome than on securing a good best outcome.[3] When our actions affect someone else—Buchak suggests—we should choose in accordance with that person's risk-attitude if we have sufficient evidence about what that risk-attitude is, but, if we do not, we should default to choosing in accordance with the most risk-averse reasonable risk-attitude. Buchak does not say much about actions which affect multiple other people, but it is clear that she would appeal to some kind of aggregation of risk-attitudes.[4] Pettigrew is explicit that our risk-attitude in that case should be the result of aggregating the risk-attitudes of the people who will be affected, giving more weight to more risk-averse attitudes while giving less weight to attitudes that, given our evidence, are less likely to be instantiated.[5]

But the idea of adopting the aggregate risk-attitude of all affected appears undermotivated, even before we consider situations where our actions affect future people.[6] On the one hand, if a person's preferences about risk-taking do not reflect what's in their best interest but merely an idiosyncratic judgment about how to pursue their goals under uncertainty, then we seem to have little reason to take their risk-attitude into account when making decisions that affect their interests—except, perhaps, in situations where we act as their legal proxies or elected representatives.

On the other hand, if a person's preferences about risk-taking do reflect what's in their best interest, then we have to face several uncomfortable choices between plausible ethical principles. To see one example, note that risk-weighted expected-utility theory allows one to prefer

- a 100% chance of a small benefit, rather than
- a 1% chance of a large benefit, a 1% chance of nothing, and the small benefit otherwise,

on the ground that the former is a sure thing while the latter risks getting nothing; while also allowing one to prefer

- a 1% chance of the large benefit, and nothing otherwise, rather than
- a 2% chance of the small benefit and nothing otherwise,

on the ground that both risk getting nothing but at least the former is better in expectation.[7]

Now imagine that we are making a policy that affects two people with these preferences and nobody else. Our choice is between a policy that gives them both the two more preferred prospects or the two less preferred ones, as in Table 1.[8]

Table 1. An Impossibility Result

	1%	**1%**	**98%**	**1%**	**1%**	**98%**
Ann	Small	Small	Small	Large	Nothing	Small
Bob	Large	Nothing	Nothing	Small	Small	Nothing
	First Policy			Second Policy		

In this table, rows specify the possible benefits received by the two people with the probabilities shown in the columns. If a person's preferences about risk-taking reflected what's in their best interest, then it would be in both Ann and Bob's best interest to choose the first policy over the second. Plausibly, then, the first policy would then be better than the second, as it would be in everyone's best interest.

But, impartially speaking, the two policies appear equally good, as they have the same chances of producing the same patterns of benefits. Either way, there is a 1% chance of someone getting a small benefit and someone else getting a large benefit, a 1% chance of someone getting a small benefit and someone else getting nothing, and a 98% chance of someone getting a small benefit and someone else getting nothing. An impartial policy-maker should not care about who gets which benefit. So, regardless of the true state of the world, an impartial policy-maker would find the outcomes of the two policies equally good; and, so, it is plausible that an impartial policy-maker would find the two policies themselves equally good. While this line of argument can reasonably be rejected, it has a good deal of intuitive plausibility. If our policy-maker were to impose some cost on a third party in order to pursue the first policy over the second, it is hard to see how this could be justified to that third party.

This shows that the following principles are incompatible: people's preferences about risk-taking reflect what's in their best interest; these preferences are risk averse in a way allowed by risk-weighted expected-utility theory; one policy is better than another if it is in every affected person's best interest; an impartial policy-maker should find equally good any two policies that have the same chance of producing the same patterns of benefits. It seems to us that the most promising response is to deny that people's preferences about risk-taking reflect what's in their best interest, at least if we grant risk-weighted expected-utility theory as an account of rational preferences. This would not imply that anyone's preferences about which risks to take are rationally impermissible, since a rational person's preferences might diverge from what's in their best interest. But it would undermine one reason for taking other people's risk-attitudes into account when making policy decisions that affect their interests.[9]

The second problem is that it is difficult to motivate the idea of giving more weight to more risk-averse attitudes. This would seem to involve, as Buchak herself concedes, a seemingly unprincipled asymmetry: giving more weight to possible complaints that we have been too risk-seeking than to possible complaints that we have not been risk-seeking enough.[10] It is possible to take other people's risk-attitudes into account in a more even-handed way, for example, by adopting the average of reasonable risk-attitudes—thus giving equal weight to all reasonable complaints—or by only ranking one action above another if it would be so ranked relative to all reasonable risk-attitudes—thus privileging no single risk-attitude.

As Buchak also concedes, the only reason for accepting the unprincipled asymmetry would seem to be that it best explains our case-specific intuitions, as in the following example due to Pettigrew:

> A Hiker's Choice. A lead hiker has to decide whether to continue climbing in bad weather, risking serious harm for the sake of great view from above the clouds. They are unable to consult other hikers roped up to them. But a few of the other hikers are known to be risk-averse enough that they would prefer to descend.[11]

Many will share Pettigrew's intuition that the lead hiker should descend, which he proposes to explain in terms of the lead hiker's duty to adopt the aggregate risk-attitude of all affected,

giving more weight to the more risk-averse attitudes while also giving less weight to attitudes that, given our evidence, are less likely to be instantiated. But it seems to us that a more natural explanation would appeal to the lead hiker's duty to avoid significant risks of unconsented harm.

Besides drawing on familiar ideas from common-sense morality, this alternative explanation can, unlike Pettigrew's own, account for what we suspect to be popular intuitions about other examples. For instance, it can account for an intuition that it is permissible to perform a risky medical procedure that is beneficial in expectation on a patient who agreed to it, even if we do not know their risk-attitude. Pettigrew's suggestion seems to imply that the procedure would be wrong because the patient might well be risk averse. Our alternative explanation, on the other hand, can permit this procedure because of the patient's consent, whilst requiring the lead hiker to descend.

Similarly, it can easily account for an intuition that it is permissible to gift someone a raffle ticket instead of the equivalent in cash, even if we do not know their risk-attitude. After all, the raffle is assumed to be as good as the cash and can result in no harm. On the other hand, Pettigrew's suggestion seems to imply that gifting the ticket would be wrong because the recipient might well be risk averse. In order to generate an intuitive verdict in this case, Pettigrew has to restrict his account to cases not involving harms. We think that it is more natural to appeal to harm-avoidance directly. But we do not claim that there really is a duty to avoid significant risks of unconsented harm. We merely claim that it is a more natural account of popular intuitions, undermining Pettigrew's claim that popular intuitions support giving more weight to the more risk-averse attitudes.[12]

2. NON-IDENTITY

The idea of adopting the aggregate risk-attitude of all affected seems not only undermotivated, but also unworkable when our choices impact future people. Parfit observed that many policies are population policies: they affect who will exist, not merely how well off they will be.[13] They can also affect which risk-attitudes will be instantiated in the population. To see this, consider the following simple example:

> Earlier or Later? You can have a child in your turbulent twenties. The child is equally likely to have a very happy life or a very unhappy life. But they would grow to seek stability and thus become risk-averse. You can alternatively have a child in your stable thirties. This would be a different child who is certain to have a merely satisfied life. But they would grow to despise stability and thus become risk-seeking. All else is equal.

This case is illustrated in Table 2, where a long dash "—" denotes nonexistence:

Table 2. Earlier or Later?

	50%	**50%**	**50%**	**50%**
Timid Child	Happy	Unhappy	—	—
Bold Child	—	—	Satisfied	Satisfied
	Child Earlier		Child Later	

Let's assume that, relative to the risk-averse attitude, a 50/50 gamble between a very happy life and a very unhappy life is less valuable than the certainty of a merely satisfied life, but, relative to the risk-seeking attitude, that gamble is more valuable than the certainty of a merely satisfied life. So, if you have a child in your twenties, the aggregate risk attitude of everyone who will be affected—the first child—will be risk averse and, so, adopting that aggregate risk-attitude would favour having the child in your thirties. On the other hand, if you have a child in your thirties, the aggregate risk-attitude of everyone who will be affected—the second child—will be risk-seeking and, so, adopting that aggregate risk-attitude would favour having the child in your twenties.

So, whatever you choose, it will be the case that you should have chosen something else. This is puzzling for several reasons.[14] First, it is puzzling that which option is right depends on which option you will choose; this results in a view that could not be action-guiding, even in highly idealised circumstances, since choosing what to do arguably presupposes ignorance about what will be chosen. Second, it is puzzling that, whichever option you will choose, you will choose wrongly, thus implying that a certain sort of moral dilemma is unavoidable. Lastly, it is puzzling that it is impermissible to have a child who will certainly have a satisfied life, simply because they will happen to be risk-seeking.

This puzzle is analogous to familiar puzzles of preference change. To see this, consider the following example that Pettigrew himself discusses elsewhere:

> A Writer's Dilemma. You want to write the great American novel. If you write it, you will be satisfied but you will develop high literary standards and you will wish that you had never written it. If you do not write it, you will be unsatisfied but you will keep your current literary standards and you will wish you had written it.[15]

A standard solution to this puzzle is snappily summarised by Krister Bykvist: "No cross-world intervention is allowed: a life should be judged only by the desires one would develop if one were to lead that life."[16] So, in this case, becoming a great writer should be evaluated in terms of the high literary standards you would have if you became one, while the alternative should be evaluated in terms of your current literary standards. So you should write the novel because you will be more satisfied if you write it than if you do not.

In our case, the analogous solution would be to demand that an option should be evaluated in terms of the risk-attitudes that would be instantiated if that option were taken. So the risk-attitude of one possible child cannot favour taking or eschewing risks that could only affect the other possible child. You should have the child that will have the better prospect, according to their own risk-attitude; this could be either child, depending on further details of the case.

This solution, however, is not generally plausible in cases where our actions will affect people's risk-attitudes. To see this, consider the following simple example:

> City or Suburb? You will always live a turbulent life. If you have a child, the child is equally likely to have a very happy life or a very unhappy life, independently of where you happen to live. But, before conceiving, you have a choice whether to settle in a quiet suburb or in a hectic city centre. If you settle in the suburb, the child you have will grow up to be risk-averse, but if you settle in the city centre the child you have will grow up to be risk-seeking. The quiet suburb is a slightly more pleasant environment for a child to grow up in, compared with the hectic city centre. All else is equal.

This case is illustrated in Table 3:

Table 3. City or Suburb?

	50%	50%	50%	50%
Timid Child	Happy^{+}	Unhappy^{+}	—	—
Bold Child	—	—	Happy	Unhappy
	Suburb		City Centre	

This time there are risks for whoever is created. But a 50/50 gamble between a very happy life and a very unhappy life is more valuable when evaluated in terms of the risk-seeking attitude than when evaluated in terms of the risk-averse attitude. So, evaluated in terms of the characteristically metropolitan risk-seeking attitudes, moving to the city centre is preferable to settling in the suburb, as long as that is evaluated in terms of the characteristically suburban risk-averse attitudes; this is so even if the city centre is a slightly less pleasant environment for a child to grow up in.

So the solution under discussion implies that you should settle in the hectic city centre. This is puzzling for at least two reasons. First, it is puzzling that you have to create a child who would have the same chances of the same life outcomes as another possible child, except that the first child's outcomes would be slightly worse. Second, it is puzzling that your risk-attitude should change depending on the option considered. This is irrational according to risk-weighted expected-utility theory which requires a coherent risk-attitude across different options.

These puzzles arise because our actions might affect which risk-attitudes are instantiated in the population. In these cases, adopting the aggregate risk-attitude of all affected threatens to become unworkable. So, even if we had some reason to take other people's risk-attitudes into account elsewhere, cases in which our actions affect future people would likely be an exception.[17]

3. EXTINCTION

If we do not adopt the aggregate risk-attitude of all affected, we might still happen to be risk averse: we might think that defaulting to the most risk-averse reasonable risk-attitude is required in cases where our decisions affect future people, or we might think that many risk-attitudes—including risk aversion—are permissible in these cases. What then ought we to do?

Pettigrew argues that if we are risk averse but accept a broadly utilitarian axiology—treating actual present people and possible future people equally—we should work to increase the chance of human extinction.[18] By contrast, many philosophers have recently argued that if we are risk neutral and accept a broadly utilitarian axiology, we should work to decrease the chance of human extinction.[19]

Pettigrew points out that humanity's continued existence on Earth can be seen as a gamble between several possible futures: a long happy future if moral and material progress continues; a future that, because of either its brevity or mediocrity, is middling in value; and a long future so miserable that, in the words of Winston Churchill, "a fortunate collision with some wandering star, reducing the earth to incandescent gas, might be a merciful deliverance."[20] Shifting probability mass away from extinction increases the risk of a long miserable future, while shifting probability mass towards extinction decreases that risk. To see this, consider the four possible futures in Table 4:

Table 4. Four Futures

Long Happy Middling Long Miserable	Extinction

In terms of this table, Pettigrew's point is that, given sufficient risk aversion, it might be preferable to shift probability mass from left to right—towards extinction—rather than from right to left—away from extinction.

We want to point out two issues with Pettigrew's argument. First, working to hasten human extinction is not necessarily the most effective way to reduce the risk of a long miserable future. It might be more effective to work directly on mitigating threats that could make the long future miserable, such as the danger of permanent totalitarianism or the persistent moral atrocities of factory farming, war, and poverty.[21] In terms of our table, even if avoiding a long miserable future is the overriding priority, it might be preferable to shift probability mass from bottom to top, towards nonmiserable possibilities, rather than from left to right, towards extinction. In fact, for a given reduction in the probability of a long miserable future, it would be better—no matter one's risk-attitude—to increase the probability of middling and long happy futures than to increase the probability of extinction. This direct approach also promises to be more palatable than pro-extinction interventions, as it would appear to be both more feasible and—in view of its potential collateral benefits to present people—more robustly good.

Since threats that could result in a long miserable future are difficult to eliminate, especially in view of their potential resurgence in the long run, some might reply that working to hasten human extinction is nonetheless a more effective intervention. But the second issue with Pettigrew's argument is that doing so could plausibly exacerbate these threats, increasing the risk of a long miserable future. In terms of our table, it might be impossible to shift probability mass towards imminent extinction without also shifting some towards a long miserable future.

This is because, given a prevalent desire for having children and for heterosexual intercourse, it seems unlikely that human extinction could be achieved by voluntary means. This point is conceded by Les Knight, the founder of the Voluntary Human Extinction Movement, who is resigned to failure in view of the fact that his movement would need one hundred percent compliance and no accidents to achieve its goal.[22] Plausibly, hastening human extinction could only be achieved through less peaceful and more painful means: rapid exploitation of fossil fuels in the hope of triggering catastrophic climate change, the creation of powerful nuclear and biological weaponry to be used in a future great-power war, or expedited progress towards advanced artificial intelligence threatening to cause human extinction.[23] The problem is that all of these interventions, even though they would increase the chance of human extinction, would also increase the risk of a long miserable future: human life would be worse in a warmer world without fossil fuel reserves, wars cause suffering and reduce the potential for international cooperation needed to make the world a better place, and advanced artificial intelligence might disempower humanity even if it does not extinguish it.

So, even if a broadly utilitarian axiology combined with a sufficiently risk-averse decision theory implies that reducing the risk of a long miserable future is an overriding priority, it is not clear to us that working to hasten human extinction would be an effective way of reducing that risk.[24]

Pettigrew's underlying decision-theoretic view does nonetheless undermine plausible instrumental considerations in favour of delaying human extinction, namely, that doing so would give humanity both the time to learn more about the value its future would have, as well as the ability

to prevent unacceptable risks from unfolding elsewhere in the cosmos.[25] It is already widely known that risk-weighted expected-utility theory can undermine instrumental reasons to seek free information; so we focus on explaining how it might also undermine instrumental reasons to acquire useful abilities.[26] To see this, consider the following unrealistic example:

> Doom Twice or Once? We can decide the fate of two planets: ours and a nearby one. There is a 30% probability that all civilizations descend into dystopia and a 70% probability that they all evolve into utopia. According to our risk-averse decision theory, this gamble is not worth taking on any planet. Humanity can become extinct now or wait to see if future technology might allow humanity to reach the other planet and prevent the emergence of civilization there. This technology is as likely as not to be feasible.

In this example, if we are risk averse, we will choose to extinguish civilisation on this planet sooner or later. If we decide to do it sooner, there are two possible outcomes:

- This planet extinct, the other planet a dystopia (probability 30%),
- This planet extinct, the other planet a utopia (probability 70%).

If we decide to do it later, there are three possible outcomes:

- This planet extinct, the other unreachable and a dystopia (probability 15%),
- This planet extinct, the other reachable and made extinct (probability 50%),
- This planet extinct, the other unreachable and a utopia (probability 35%).

With a moderately risk-averse attitude, going extinct sooner is preferable to going extinct later.[27] This is puzzling, since delaying our extinction could prevent what is supposed to be an unacceptably risky gamble from unfolding on another planet. This implication is harder to explain away than information aversion because waiting for the new option to become available carries no risk of being misleading. The feature of risk-weighted expected-utility theory responsible for this implication is that the weight given to an outcome does not depend solely on its probability but also on its position relative to other potential outcomes. Introducing a mid-range outcome—both planets becoming extinct, an option humanity would select if the other planet were accessible—could reduce the weight given to the best possible outcome more than it reduces the weight given to the worst.

So, surprisingly, a proponent of extinction motivated by risk aversion might have to disagree with Bernard Williams's suggestion that a commitment to impartiality would lead us to try to spread extinction across the cosmos:

> What would it be like to take on every piece of suffering that at a given moment any creature is undergoing? . . . if for a moment we got anything like an adequate idea of what that is, and we really guided our actions by it, then surely we would annihilate the planet, if we could; and if other planets containing conscious creatures are similar to ours in the suffering they contain, we would annihilate them as well.[28]

While a concern to avoid great suffering gives rise to instrumental reasons to delay human extinction—as that might help us prevent great suffering elsewhere—a concern to avoid great risks undermines these kinds of instrumental reasons. We leave it open whether this implication is a weighty reason to abandon risk-weighted expected-utility theory. But it does point to important differences between the concern to avoid great suffering and the concern to avoid great risks.

4. CONCLUSION

Recent philosophical discussion has featured an argument for working to hasten human extinction, drawing on a decision-theoretic view that accommodates the risk-attitudes of all affected while giving more weight to the more risk-averse attitudes. The underlying view is, however, undermotivated, unworkable when applied to future people, and it does not necessarily support working to hasten human extinction, even if it does undermine plausible instrumental considerations in favour of delaying it. The idea that human extinction should be hastened is not only counterintuitive, but might also be dangerous. Realistic attempts to realise it would likely involve violence in the present and risk suffering in the future. It is therefore important to emphasise that the reports of a philosophical case for humanity's demise have been greatly exaggerated.[29] [30]

NOTES

1. See Pettigrew 2022; 2024. Lara Buchak uses a similar argument to support aggressive action on climate change; see Buchak (2019, building on Buchak [2017]). See also Mogensen (2022).
2. See Quiggin 1982 and Buchak 2013. More formally, according to the orthodox expected-utility theory, a rational person should maximise the expectation of some numerical utility function, that is:

   $$\sum_x p\,(x)\,u\,(x).$$

   In this formula, the utility function u is a numerical index of value, while p is a function assigning probabilities to different possible outcomes, which will be assumed to be finite in number. On the other hand, according to risk-weighted expected-utility theory, a rational person should instead maximise the following quantity:

   $$\sum_x [r\,(p\,(\geq x)) - r\,(p\,(> x))]\,u\,(x).$$

   In this formula, the risk function r is a function that continuously increases between 0 and 1, "$p(\geq x)$" denotes the probability of getting an outcome at least as preferred as x, and "$p(>x)$" denotes the probability of getting an outcome more preferred than x; for the purposes of this formula, indifferent outcomes are not distinguished. Thus, according to expected-utility theory, the weight of each outcome is given by its probability, while, according to risk-weighted expected-utility theory, it might also depend on its position relative to other possible outcomes. A convex risk function, like $r(x) = x^2$, gives more weight to relatively worse outcomes; a concave risk function, like $r(x) = \sqrt{x}$, gives more weight to relatively better outcomes; with a linear risk function, we recover expected-utility theory.
3. See Buchak 2013, 53–56; 2014; 2017; 2019.
4. See Buchak 2017; 2019.
5. See Pettigrew 2022; 2024.
6. Compare Thoma 2023.
7. This is a version of the preference pattern involved in the Allais Paradox; see Allais (1953).
8. A similar example appears in Nebel (2020) and Bradley (2022), but see also McCarthy et al. (2020).
9. This is not the only possible response to this case. For example, one might instead deny the principle, which we might call "outcome impartiality," that an impartial policy-maker should find equally good any two policies that have the same chances of producing the same patterns of individual outcomes. It might instead be suggested that the correct form of impartiality is "prospect impartiality," which only requires that an impartial policy-maker should find equally good any two policies which produce the same pattern of individual prospects. We think that both principles are intuitively plausible. Moreover, as there is no logical incompatibility between them, assuming one of them does not beg the question against those attracted to the other. We thank an anonymous referee for pressing us to clarify this point.
10. See Buchak 2019, 73.
11. See Pettigrew 2022, 13–14.
12. In restricting his account to harms, Pettigrew departs from Buchak; see Pettigrew (2022, 16). What's more, Buchak (2019) herself switches back and forth between talking in terms of consent and deference to risk-attitudes, despite important differences between the two kinds of accounts.

13. See Parfit 1984, 351–79.
14. Compare Bykvist 2006. See also Carlson (1995, 100–102).
15. See Pettigrew 2019, 227–28. See also Bricker (1980).
16. Bykvist 2006, 279. Instead of taking this standard route, we might appeal to present, actual, or noncontingent risk-attitudes, by analogy with the different possibilities available in response to the puzzles of preference change; see Bykvist (2006).
17. Pettigrew (2022, 16–17) anticipates that it might be difficult to aggregate risk-attitudes of future people, saying that we should adopt a risk-averse attitude at least if every possible population will be, in aggregate, risk averse. This condition is, however, unlikely to be satisfied; and even if different possible populations are risk averse to different degrees, the puzzles of this section will reappear.
18. See Pettigrew 2022; 2024.
19. See Bostrom 2003; Beckstead 2019; and MacAskill and Greaves 2021.
20. Churchill 1931, 67.
21. On some of these threats, see Ord (2020, 153–58).
22. See Buckley 2022. It is often thought that, eventually, a high-fertility subpopulation will emerge, reversing any voluntary depopulation; see, for example, MacAskill (2022, 157). But see Arenberg et al. (2022) for an argument to the contrary.
23. On these extinction risks, see Ord (2020, 89–162).
24. For all we said, extinction sooner might be better than extinction later, if these were the options. Our point is that they are not our options.
25. See Ord 2020, 56–57, 311–12.
26. See Wakker 1988. Buchak (2013, 191–97) defends information avoidance by arguing that free information might carry some risk of being misleading; see also Ahmed and Salow (2019).
27. Calculation: Let 100 be the utility of a planetary utopia, −100 the utility of a planetary dystopia, and 0 the utility of an extinct planet. Let's assume that the utility of two planets is equal to the sum of their utilities. Then the risk-weighted expected-utility of hastening our extinction is:

$$[r(1) - r(0.7)](-100) + [r(0.7) - r(0)](100),$$

while the risk-weighted expected utility of delaying our extinction (and then possibly extinguishing both planets) is:

$$[r(1) - r(0.85)](-100) + [r(0.85) - r(0.35)](0) + [r(0.35) - r(0)](100).$$

For a moderately risk-averse risk-function $r(x) = x^2$, the first evaluates to −2, while the second evaluates to −15.5. The risk-weighted expected-utility of extinguishing both planets would be 0 and that of preventing extinction on both planets would be −4.
28. Williams 2006, 146–47.
29. This sentence adapts Mark Twain's famous comment; see Twain (1906, 460).
30. We would like to thank Richard Pettigrew and an anonymous referee for their comments, other participants of the UCL Global Priorities Reading Group for their feedback, and Longview Philanthropy for financial support. Kacper Kowalczyk received funding from Longview Philanthropy during the writing of this paper.

REFERENCES

Arif Ahmed and Bernhard Salow 2019. "Don't Look Now," *British Journal for the Philosophy of Science* 70(2): 327–50. doi: 10.1093/bjps/axx047.

Allais, Maurice 1953. "Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'ecole americaine," *Econometrica* 21(4): 503–46. doi: 10.2307/1907921.

Arenberg, Samuel, Kevin Kuruc, Nathan Franz, Sangita Vyas, Nicholas Lawson, Melissa LoPalo, Mark Budolfson, Michael Geruso, and Dean Spears 2022. "Intergenerational Transmission Is Not Sufficient for Positive Long-Term Population Growth," *Demography* 59(6): 2003–12. doi: 10.1215/00703370-10290429.

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

Bostrom, Nick 2003. "Astronomical Waste: The Opportunity Cost of Delayed Technological Development," *Utilitas* 15(3): 308–14. doi: 10.1017/s0953820800004076.

Bradley, Richard 2022. "Impartial Evaluation Under Ambiguity," *Ethics* 132(3): 541–69. doi: 10.1086/718081.

Bricker, Phillip 1980. "Prudence," *Journal of Philosophy* 77(7): 381–401.
Buchak, Lara 2013. *Risk and Rationality*, Oxford: Oxford University Press.
———. 2014. "Risk and Tradeoffs," *Erkenntnis* 79(S6):1091–117. doi: 10.1007/ s10670-013-9542-4.
———. 2017. "Taking Risks Behind the Veil of Ignorance," *Ethics* 127(3): 610–44. doi: 10.1086/690070.
———. 2019. "Weighing the Risks of Climate Change," *The Monist* 102(1): 66–83. doi: 10.1093/monist/ony022.
Buckley, Cara 2022. "Earth Now Has 8 Billion Humans. This Man Wishes There Were None," *The New York Times*, November 23.
Bykvist, Krister 2006. "Prudence for Changing Selves," *Utilitas* 18(3): 264–83. doi: 10.1017/S0953820806002032.
Carlson, Erik 1995. *Consequentialism Reconsidered*, Dordrecht: Kluwer Academic Publishers.
Churchill, Winston 1931. "Fifty Years Hence," *Maclean's (Toronto)* 44(22): 7, 66–67.
MacAskill, William 2022. *What We Owe the Future: A Million-Year View*, New York: Basic Books.
MacAskill, William and Hilary Greaves 2021. "The Case for Strong Longtermism," GPI Working Paper No. 5-2021. https://globalprioritiesinstitute.org/wp-content/uploads/The-Case-for-Strong-Longtermism-GPI-Working-Paper-June-2021-2-2.pdf. (Accessed 10 January 2024.)
McCarthy, David, Kalle Mikkola, and Teruji Thomas 2020. "Utilitarianism with and without Expected Utility," *Journal of Mathematical Economics* 87: 77–113. doi: 10.1016/j.jmateco.2020.01.001.
Mogensen, Andreas 2022. "Respect for Others' Risk Attitudes and the Long-Run Future," Global Priorities Institute, GPI Working Paper No. 20-2022. https://globalprioritiesinstitute.org/wp-content/uploads/Andreas-Mogensen-Respect-for-others-risk-attitudes-and-the-long-run-future.pdf. (Accessed 10 January 2024.)
Nebel, Jacob M. 2020. "Rank-Weighted Utilitarianism and the Veil of Ignorance," *Ethics* 131(1): 87–106. doi: 10.1086/709140.
Ord, Toby 2020. *The Precipice: Existential Risk and the Future of Humanity*, London: Bloomsbury Publishing.
Parfit, Derek 1984. *Reasons and Persons*, Oxford: Oxford University Press.
Pettigrew, Richard 2019. *Choosing for Changing Selves*, Oxford: Oxford University Press.
———. 2022. "Effective Altruism, Risk, and Human Extinction," Global Priorities Institute, GPI Working Paper No. 2-2022, Version 1.0. https://globalprioritiesinstitute.org/wp-content/uploads/Richard-Pettigrew-Effective-altruism-risk-and-human-extinction-January-2022.pdf. (Accessed 10 January 2024.)
———. 2024. "Should Longtermists Recommend Hastening Extinction Rather Than Delaying It?" *The Monist* 107(2): 130–45.
Quiggin, John 1982. "A Theory of Anticipated Utility," *Journal of Economic Behavior and Organization* 3(4): 323–43. doi: 10.1016/0167-2681(82)90008-7.
Thoma, Johanna 2023. "Taking Risks on Behalf of Another," *Philosophy Compass* 18(3). doi: https://doi.org/10.1111/phc3.12898.
Twain, Mark 1906. "Chapters from My Autobiography: II," *North American Review* 183(599): 449–60.
Wakker, Peter 1988. "Nonexpected Utility as Aversion of Information," *Journal of Behavioral Decision Making* 1(3): 169–75. doi: 10.1002/bdm.3960010305.
Williams, Bernard 2006. "The Human Prejudice," in A.W. Moore, ed., *Philosophy as a Humanistic Discipline*, Princeton, NJ: Princeton University Press, 135–52.