

Longtermism and social risk-taking

H. Orri Stefánsson

Abstract. A social planner who evaluates risky public policies in light of the other risks with which their society will be faced should judge favourably some such policies even though they would deem them too risky when considered in isolation. I suggest that a longtermist would—or at least should—evaluate risky policies in light of their prediction about future risks; hence, longtermism supports social risk-taking. I consider two formal versions of this argument, discuss the conditions needed for the argument to be valid, and briefly compare these conditions to some risky policy options with which actual public decision-makers are faced.

Keywords: Longtermism; risk aversion; social risk-taking; bundling gambles; decision theory.

1. Introduction

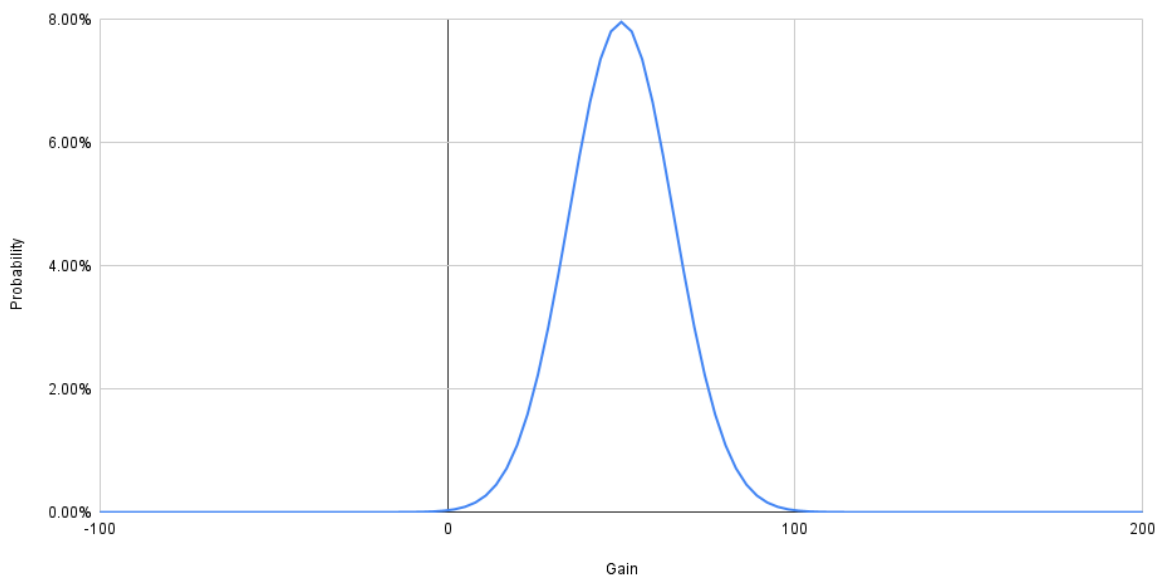
Some risks seem unacceptable when considered on their own, even though they seem acceptable when considered as part of a larger bundle of risks. For instance, while those who are risk averse with respect to money might turn down a 50/50 gamble between losing \$100 and winning \$200, few people would turn down a bundle of 100 such independent¹ gambles (see, e.g., Samuelson 1963, Rabin 2000, Kahneman 2012). After all, such a bundle has a monetary expectation of \$5,000 and has only a 0.04% chance of resulting in monetary loss. As Rabin and Thaler (2001: 223) put it: “A good lawyer could have you declared legally insane for turning down this [bundle].”

Something similar would seem plausible when taking risks that affect others. For instance, suppose that a physician is considering a risky intervention that has a half chance of costing a patient 1 unit of wellbeing and a half chance of benefitting the patient by 2 units of wellbeing. Then if losing 1 unit of wellbeing is significant, the physician might reasonably choose not to make the intervention; in fact, they will not make the intervention if they are moderately risk averse and/or moderately loss averse with respect to the wellbeing of others, in a sense to be made precise in the next section.

¹ Here and elsewhere, gambles are ‘independent’ if the probability distribution associated with any one of the gambles is independent of the outcome of any of the other gambles.

But now suppose instead that the intervention affects 100 patients, where for each patient the intervention is a 50/50 gamble between losing 1 unit of wellbeing and gaining 2 units, and the gambles are probabilistically independent. In that case, it seems that even a moderately risk or loss averse decision-maker should choose the intervention.² After all, the expected total wellbeing effect of the bundle is a gain of 50 units, and the probability that the bundle results in a total wellbeing loss is minuscule (0.0004).³ In addition, the bundle can at most result in each person losing one unit of wellbeing, and each patient’s gamble provides them with positive expected wellbeing. The figure shows the distribution and probability of total wellbeing from the bundle.

Distribution and probability of terminal outcomes



More generally, the more gambles that a risk or loss averse ‘social planner’ considers together, the more prone they should be to accept each gamble, assuming for instance that each gamble has a positive expectation in whatever *objective* quantity the planner takes to be the object of risk-free axiology (for instance, wellbeing or years in full health).

Now, it seems plausible that to truly internalise the long-term perspective (or ‘longtermism’; more on this in a moment) means to judge gambles in light of what one predicts about the (long-term) future. Furthermore, some think that one should treat a sequence of gambles as one would treat a bundle of those gambles. But then the above argument seems to

² It might be worth noting that some would not accept the bundle, for instance, those who accept some version of *leximin*, and thus always choose the gamble where the lowest possible wellbeing level affected by the choice is highest. Such a rule is arguably not *moderately* risk averse. The same is true of some *ex post* contractualists.

³ In fact, the probability of losing more than 50 units in total is merely 0.0000000000013! In contrast, the probability of gaining more than 50 units is 0.46.

show that taking the long-term perspective should often make a social planner more risk prone, as compared with a ‘short-termist’ social planner with the exact same attitudes to risks, gains, and losses. In the next section I present a more formal version of this argument.

But first, let’s consider a rather different and, in some ways, less abstract argument for the same conclusion. Suppose that a social planner is considering a risky social change—an ‘experiment’—such as legalising recreational drugs. The social planner thinks that, in terms of aggregate wellbeing, the experiment has a positive expectation; let’s say they take this to be a 50/50 gamble between the population of interest losing in total 100 units of wellbeing per generation and gaining in total 200 units of wellbeing per generation. Again, if the social planner is risk or loss averse with respect to the population’s total wellbeing, then they might not be willing to take the risk if they focus on the effect on the current and perhaps the next few generations, since they might judge that they cannot justify exposing them to a half chance of losing 100 units of wellbeing for the sake of the same chance of gaining 200 units. However, if they take a longer perspective, then they might reach a different conclusion. As long as a harmful experiment can be stopped, they might reason that the long-term benefit of a gain of 200 units of wellbeing per generation if the experiment is a success, in addition to the knowledge that the experiment is expected to bring—and assuming that that knowledge will be put to good use⁴—would outweigh the risk of harm to the current generation. Indeed, as formally demonstrated in the next section, this is precisely how a moderately risk or loss averse social planner should reason.

The main objective of this chapter is to carefully formulate two arguments that what I will call ‘longtermism’ should make a risk or loss averse social planner more risk prone; and, in particular, to consider the conditions needed for these arguments. By ‘longtermism’ I shall simply mean the claim that a social planner should take the long-term perspective. This is, of course, admittedly quite vague (and, in fact, I shall discuss two distinct notions of ‘taking the long-term perspective’, corresponding to the aforementioned two arguments). However, as will become evident, the precise meaning of ‘longtermism’ (e.g., *how* long and ‘wide’ perspective the longtermist takes) is unimportant for my argument; a small shift in perspective beyond the present moment is in fact enough for my argument.

It might be worth acknowledging that what I am calling ‘longtermism’ isn’t quite what Greaves and MacAskill (2021) call (axiological strong) ‘longtermism’; informally, the claim that

⁴ For an illuminating discussion of this and related issues, see Barrett and Buchanan (2023), who point out that an argument for what they call “progressive experimentalism” has to address the *uptake problem*, that is, the possibility that the knowledge gained from a social experiment will not be put to good use.

“far-future effects are the most important determinant of the value of our options” (op. cit: 3). However, I think that longtermism as they understand it implies what I am calling ‘longtermism’. If “far-future effects are the most important determinant of the value of our options”, then it would seem we *should* take the (very) long-term perspective. For if we do not, then we will simply ignore the most important determinants of the value of our options.

In the next section I spell out two formal arguments for the claim that (what I call) longtermism supports social risk-taking. An important part of my analysis will consist in examining the conditions required for the success of these arguments. These conditions will also be made precise in the next section. In the third and final section I consider to what extent these conditions hold for real-life risky public policies.

2. The formal arguments

2(a) Framework and definitions

To keep the argument as simple as possible, I will assume throughout this chapter some version of *generalised utilitarianism* and a population of a fixed size. An example of generalised utilitarianism is, of course, standard utilitarianism, according to which the value of a population is determined by the sum total of wellbeing in the population; another example is prioritarianism, according to which the value of a population is determined by the sum of priority-weighted wellbeing in the population. In what follows, I shall use standard utilitarianism for illustrative purposes, but my argument can be easily generalised to, say, prioritarianism.

Let $\mathbf{O} = \{o_1, \dots, o_m\}$ be the set of m (terminal, i.e., risk-free) outcomes. According to generalised utilitarianism, we can take each outcome to be a vector of wellbeing levels, e.g., $o_i = (z_i^1, \dots, z_i^n)$ given a population of n people. Gambles (or ‘lotteries’) are probability⁵ distributions over \mathbf{O} . Let \mathbf{L} be the set of gambles, that is, the set of all vectors $(\lambda_1, o_1; \dots; \lambda_m, o_m)$ such that $\lambda_1, \dots, \lambda_m \in [0,1]$ and $\sum_{k=1}^m \lambda_k = 1$. Let \succsim be a (weak) better than relation (or social preference relation) defined on both \mathbf{O} and \mathbf{L} . The expression ‘ $o_i \succsim o_j$ ’ should be read as saying that outcome o_i is at least as good as outcome o_j . Correspondingly, ‘ $o_i \sim o_j$ ’ means that o_i and o_j are equally good, and ‘ $o_i \succ o_j$ ’ that o_i is strictly better than o_j . Finally, let u be a real valued utility function

⁵ I shall not make any particular assumptions about how or why these probabilities are available to the decision-maker. But it may be worth noting that the arguments I consider can be made for subjective as well as objective probabilities, and they can be extended to imprecise probabilities.

on the set of outcomes (unique up to positive affine transformations⁶) that represents \succsim on \mathbf{O} in the sense that $o_i \succsim o_j \Leftrightarrow u(o_i) \geq u(o_j)$.⁷

By risk aversion, I do *not* mean that the quantities of interest are taken to have decreasing marginal value, that is, I do not mean what economists typically mean by risk aversion. Elsewhere, I have argued that that picture fails to capture important aspects of risk aversion (Stefánsson and Bradley 2019). Informally, by risk aversion, I mean that the value that any of the *better* potential outcomes from a gamble contributes towards the overall value of the gamble is less than the probability-weighted utility of that outcome (Buchak 2013, Stefánsson and Bradley 2015, 2019). The version of this idea that is best-known amongst philosophers is due to Buchak (*op. cit.*). Hers is also a particularly tractable version of this idea, which I will hence use below, for illustrative purposes.

We can now state different theories about how to manage risk and choose between gambles, before defining risk aversion more formally. According to expected utility theory, understood as a theory of rationality, one's preferences between gambles should correspond to how the gambles' expectations of utility compare. A gamble's expectation of utility is found by first weighing the utility of each possible outcome of the gamble by its probability, and then adding together these probability weighted utilities. More formally:⁸

Expected Utility (EU) theory. For any $L_\alpha = (\alpha_1, o_1; \dots; \alpha_m, o_m)$, $L_\beta = (\beta_1, o_1; \dots; \beta_m, o_m) \in \mathbf{L}$ and for any rational \succsim :

$$L_\alpha \succsim L_\beta \Leftrightarrow \sum_{i=1}^m u(o_i) \cdot \alpha_i \geq \sum_{i=1}^m u(o_i) \cdot \beta_i$$

Expected utility theory does not allow for risk aversion in the above sense. EU theory does allow for a particular kind of risk aversion: for instance, if the outcomes are quantities of money, then EU theory can accommodate risk aversion with respect to money in the sense that it allows that a decision-maker strictly prefers any sure sum of money to a lottery whose expectation is that same sum of money. This is achieved by stipulating a *concave* utility function, u . But since the expected utility formula, as stated above, implies that the value that *any*

⁶ Saying that u is unique up to positive affine transformations means that if both u and u' represent \succsim , then there exist real numbers as a and $b > 0$ such that $u' = a + u \cdot b$.

⁷ A utility function is simply a numerical representation of a ranking. Therefore, if the ranking is based on a (morally) 'better-than' relation, then the corresponding utility function will be a (moral) value function. For instance, below we interpret the 'utility' of an outcome as the sum of wellbeing in the outcome.

⁸ Note that, formally, the two gambles contain *the same* outcomes. But, for either gamble, many of the outcomes may receive a zero probability and thus not be *possible* outcomes from the gamble.

outcome contributes to a gamble's value equals the outcome's probability weighted utility, expected utility theory cannot accommodate risk aversion with respect to utilities, which is what I mean by 'risk aversion' (for a discussion, see, e.g., Stefánsson and Bradley 2019).

By contrast, rank-dependent utility theory, defended for instance by Buchak (2013) (under the name of 'risk weighted expected utility theory'), allows for what I am calling risk aversion. Informally, rank-dependent utility theory evaluates a gamble by adding to the utility of the gamble's *worst* possible outcome a weighted version of the expected utility that the gamble offers over and above its worst possible outcome, where the weight in question depends on the relevant agent's attitude to risk. If the agent is *maximally* risk averse, then the weight in question turns the expectation into zero, such that a gamble is evaluated by its worst possible outcome. If the agent in question is an expected utility maximiser, however, then the weight leaves the expectation unaltered, meaning that the gamble is evaluated by its expected utility. In most applications, the weight is somewhere between these, that is, the agent is assumed to be risk averse but not maximally so.

To state this theory more formally, now let $\mathbf{O}_{\square} = \{o_{[1]}, \dots, o_{[m]}\}$ be a non-decreasing re-ordering of \mathbf{O} according to the preference relation of interest, meaning that for any i , $o_{[i+1]} \succeq o_{[i]}$. And let r be a real valued and increasing 'risk function' on probabilities, satisfying the constraint that $r(0) = 0$, $r(1) = 1$.

Rank-Dependent Utility (RDU) theory. *For any rational \succeq there is an r such that for any $L_{\alpha} = (\alpha_1, o_1; \dots; \alpha_m, o_m)$, $L_{\beta} = (\beta_1, o_1; \dots; \beta_m, o_m) \in \mathbf{L}$:*

$$\begin{aligned} L_{\alpha} \succeq L_{\beta} &\Leftrightarrow u(o_{[1]}) + \sum_{j=2}^m ([u(o_{[j]}) - u(o_{[j-1]})] r[\sum_{i=j}^m \alpha_i]) \\ &\geq u(o_{[1]}) + \sum_{j=2}^m ([u(o_{[j]}) - u(o_{[j-1]})] r[\sum_{i=j}^m \beta_i]) \end{aligned}$$

In RDU theory, risk aversion is captured by a *convex* risk function, which implies that for each gamble, the better potential outcomes get a lower weight than they do according to EU theory. Following Buchak (2013), I will use $r(x) = x^2$ as a canonical example of a risk function of a risk averse decision-maker. With this risk function, the RDU of a 50/50 gamble between losing one unit (of utility) and gaining two units is:

$$-1 + \left(\frac{1}{2}\right)^2 * 3 = -\frac{1}{4}$$

Next, I define ‘risk aversion’ more generally. The definition makes use of the concept of a ‘mean preserving spread’.⁹ Informally, L_B is a mean-preserving spread of L_A ’s utilities if the two gambles offer the same mean (i.e., expected) utility even though the probability density function (or probability mass function, for discrete outcomes) associated with the former is more spread, that is, assigns higher values to more extreme values.

Risk aversion. *A preference relation, \succeq , is (generally)¹⁰ risk averse if $L_A \succ L_B$ whenever L_B is a mean-preserving spread of L_A ’s utilities.*

To illustrate the above definition, suppose that $u(o_1) = 0$, $u(o_2) = 5$, $u(o_3) = 10$. Then $(\frac{1}{3}, o_1; \frac{1}{3}, o_2; \frac{1}{3}, o_3)$ is a mean-preserving spread of the utilities of $(0, o_1; 1, o_2; 0, o_3)$: they offer the same expectation of utility but the probability mass function associated with the former assigns positive probability to more extreme utilities. So, someone who is generally risk averse would prefer the latter to the former. And, indeed, with $r(x) = x^2$, the risk-dependent utility of the latter is higher than the former: 5 compared to 2.778. To connect this to the informal statement of risk aversion I gave above—that is, the idea that the value that any of the better potential outcomes from a gamble contributes towards the overall value of the gamble is less than the probability-weighted value of that outcome—note that for the above risk-free ‘gamble’ to get a higher value than the risky one, it must be the case that for the risky gamble, at least one of o_2 and o_3 contribute a value to the gamble that is smaller than their probability-weighted utilities.¹¹

Now let’s turn to what I call ‘loss aversion’.¹² Unlike for instance *prospect theory* (Kahneman and Tversky 1979), rank-dependent utility theory does not (by itself) allow for loss aversion. Nor, in fact, can any theory that has been proposed as a *normative* theory for managing risk account for loss aversion. But loss aversion does not seem obviously unreasonable, at least intuitively, when viewed from the perspective of a social planner. For instance, it is not obviously irrational—and some might even think that it is morally right—that a social planner

⁹ For a historically important discussion of the connection between mean-preserving spread and risk aversion, see Rothschild and Stiglitz (1970).

¹⁰ Sometimes we may be interested in preference relations that are not generally risk averse but rather risk averse for say wellbeing levels within some ranges but risk seeking or risk neutral with respect to other ranges. Such complexities can however be set aside for now.

¹¹ This assumes that the value of any ‘gamble’ that results for sure in some outcome is equal to the value (in this case, the ‘utility’) of that outcome, which RDU as formulated above of course implies.

¹² It may be worth emphasising that a decision-maker can be *both* loss averse and risk averse (and, moreover, that some choices can be explained by either risk aversion or loss aversion). To keep things simple, I shall however assume, in what follows, that the social planner of interest is *either* loss averse or risk averse.

is more concerned with avoiding a loss in the present generation's welfare than with securing a comparable increase in the present generation's welfare.

Informally, a social planner who is loss averse in aggregate wellbeing finds that it is worse when their population loses aggregate wellbeing of magnitude x than it is good when their population gains aggregate wellbeing of magnitude x . To make this more precise, let SQ_t be the 'status quo' at time t . Given the assumption of generalised utilitarianism, we can treat this as the wellbeing distribution that holds at time t . We now define a time-relative moral value function V_t —that is, one function for each time t —and we assume that each such function represents the better than relation at the corresponding time, \succsim_t . Then we can define:

Loss aversion. *A preference relation, \succsim , is (generally)¹³ loss averse with respect to aggregate wellbeing if for any time t , for any triple of wellbeing vectors $(z_i^1, \dots, z_i^m) = o_i$, $(z_j^1, \dots, z_j^m) = o_j$ and $(z_k^1, \dots, z_k^m) = o_k$ such that*

$$\left(\sum_{l=1}^m z_i^l - \sum_{l=1}^m z_j^l \right) = \left(\sum_{l=1}^m z_j^l - \sum_{l=1}^m z_k^l \right) > 0,$$

if $(z_j^1, \dots, z_j^m) = SQ_t$ then any V_t that represents \succsim_t satisfies:

$$V_t(o_i) - V_t(o_j) < V_t(o_j) - V_t(o_k)$$

Less formally, at any time t and given any magnitude of wellbeing x , it is worse, according to a loss averse relation, that society's aggregate wellbeing decreases by x than it is good that society's wellbeing increases by x .

Now, someone might object to the terminology, *loss aversion*, for the following reason.¹⁴ A decision-maker who is loss averse, as I have defined it above, isn't strictly speaking particularly concerned with *a person losing* wellbeing; for instance, if one person loses wellbeing of magnitude x while another person gains wellbeing of magnitude y , then that may well count as an overall improvement, according to such a decision-maker, even if y is only marginally greater than x .¹⁵ What they are averse to is *loss in total wellbeing* compared to some status quo. Moreover, some might think that loss aversion is only plausible if it concerns losses to persons. I disagree. Loss aversion understood as aversion to specific persons losing wellbeing (as opposed

¹³ We could similarly define loss aversion with respect to some wellbeing levels (or differences), but general loss aversion suffices for our purposes.

¹⁴ I thank Andreas Schmidt for making this point.

¹⁵ If the planner is a loss averse prioritarian, then we add the qualification that the person losing x is better off than the person gaining y .

to aversion to the population losing aggregate wellbeing) would have extremely anti-egalitarian implications: it would sometimes imply that we should not make an intervention that would both increase aggregate wellbeing and increase inequality, for instance, by redistributing resources from those who are better off to those who are worse off. So, I think that the loss aversion, as I have defined it, is more normatively appealing.¹⁶

So, let's now consider what a loss averse outcome axiology might look like. We can start by zero-normalising V_t around SQ_t , that is, $V_t(SQ_t) = 0$ for any time t (which of course requires re-normalising when the distribution changes). Next define the function:

$$\emptyset(x) = \begin{cases} x & \text{if } x \geq 0 \\ 3x & \text{if } x < 0 \end{cases}$$

Finally, we can state:

Loss averse utilitarianism (LAU). For any (z_i^1, \dots, z_i^m) and any time t , if $SQ_t = (z_j^1, \dots, z_j^m)$ then:

$$V_t(z_i^1, \dots, z_i^m) = \emptyset \left(\sum_{l=1}^m z_i^l - \sum_{l=1}^m z_j^l \right)$$

Less formally, the value of any population at time t , according to loss averse utilitarianism, is found by comparing its aggregate wellbeing to the aggregate wellbeing of the 'status quo population' at time t (that is, the actual population at time t), in a way that inflates the difference just in case the status quo population offers higher aggregate wellbeing. This implies that loss averse utilitarianism will, at any time, agree with ordinary utilitarianism—which we can understand as the special case of loss averse utilitarianism when $\emptyset(x) = x$ for any x —on the ranking of all *risk-free* outcomes. But as we shall see in the next subsection, loss averse utilitarianism may disagree with utilitarianism about how to rank *gambles*, even if, say, expected utility theory is assumed as the theory for managing risk.

2(b) *First argument: Bundling gambles*

¹⁶ Why not conclude from the above that loss aversion isn't appealing at all, rather than settling for loss aversion with respect to aggregate wellbeing, as I am suggesting? Don't we also need a *positive* argument in favour of loss aversion with respect to aggregate wellbeing, in addition to the argument against loss aversion at the level of individual wellbeing? (I thank a referrer for raising these questions.) Offering anything like a conclusive positive argument of this sort is beyond the scope of this chapter. Instead, the following brief suggestion will have to suffice. In general, it may seem plausible that one should be especially *cautious* when taking risks on other people's behalf, in particular, when one cannot consult each person (for a related discussion, see Buchak 2017 and Thoma 2023). And we can understand a social planner as making choices on behalf of the generation that selects them to do so (at least in a democratic society). One way to be cautious when taking risks on behalf of *the generation*, interpreted as a unified agent, is to be more concerned with avoiding losses in *its* wellbeing than with corresponding gains.

Let's now turn to the first argument that longtermism supports social risk-taking. The argument is based on the observation that sometimes a sufficiently large bundle of independent gambles of a particular type is acceptable even when an individual gamble of that type is not.

I will continue to focus on 50/50 gambles between someone losing 1 unit of wellbeing and gaining 2 units. For illustrative purposes, I will assume that the value of the status quo (at the time of decision) is 0, both when applying the risk averse (RDU) and loss averse (LAU) theory. This allows us to determine whether a gamble is worth taking, according to each theory, by checking whether the gamble's value is positive, according to that theory.

Let's start by considering a social planner who applies rank-dependent utility theory to manage risk. Such a planner can evaluate risk-free outcomes various ways, but again I will work with a utilitarian outcome axiology to keep things simple. Given the stipulation that $r(x) = x^2$, a utilitarian who uses rank-dependent utility theory to manage risk will turn down the single gamble, since, as shown above, its value is $-\frac{1}{4}$. Such a decision-maker will also turn down a package of two such gambles, but they will accept a package of three such gambles, since its value is:¹⁷

$$-3 + 3 * \left(\frac{7}{8}\right)^2 + 3 * \left(\frac{4}{8}\right)^2 + 3 * \left(\frac{1}{8}\right)^2 = 0.09375$$

Similarly, a loss averse utilitarian who uses expected utility theory to manage risk, will, given the above choice of \emptyset , turn down the single bet, whose value to them is $-\frac{1}{2}$; they will however be indifferent between accepting and rejecting a package of two such bets; and they will accept a package of three such bets, whose value to them is:

$$-\frac{9}{8} + \frac{9}{8} + \frac{6}{8} = \frac{6}{8}$$

To appreciate the importance of the assumption of probabilistic independence—for instance, the assumption that the probability that any gamble turns out unfavourably is independent of how any other gamble turns out—consider first the extreme case where the gambles are all perfectly positively correlated; that is, either they all turn out well or all turn out

¹⁷ In fact, in the limit, with an *infinite* number of gambles, a risk averse maximiser of rank-dependent utility behaves exactly like an expected utility maximiser (see, e.g., Buchak 2013: 217-218).

badly. In that case, the bundle of three gambles is a 50/50 gamble between losing 3 units of wellbeing and gaining 6 units, and a rank dependent utilitarian and loss averse utilitarian will respectively evaluate the bundle as follows:

$$-3 + 0.25(9) = -0.75$$

$$-9 + 0.5(6) = -6$$

In other words, both the loss averse utilitarian and the utilitarian who uses rank-dependent utility theory to manage risk will turn down the bundle of three gambles when the gambles are perfectly correlated. Now, the same is not true of all instances of *imperfect* positive correlation; that is, cases where some gambles are probabilistically dependent without being perfectly positively correlated. How much the gambles can be positively correlated for the argument to go through depends on details about both the gambles and the degree of risk or loss aversion of the decision-maker. But the general point is that the argument that risk and loss averse decision-makers become more prone to accepting a risky gamble when it is part of a larger bundle of gambles does not hold if the gambles are *too* positively correlated.

What if gambles are *negatively* correlated. In the extreme case, where two gambles of the kind we have been considering are perfectly negatively correlated, the gambles together offer a sure gain of 1 unit of wellbeing, since one gamble will result in a gain of 2 units while the other gamble results in a loss of 1 unit. Therefore, both loss averse and risk averse social planners will accept the bundle of two such gambles. More generally, since negative correlations reduce the possible spread in outcomes, both loss and risk averse decision-makers will typically welcome such correlation.

But returning now to (sufficiently) independent gambles, the upshot of the above is that whether the social planner is risk averse or loss averse, they will want to turn down a gamble like the one we are concerned with here if offered only one such gamble, but they will accept a bundle of three such gambles.¹⁸ And, of course, the same is true of larger bundles, as long as they do not risk a catastrophe or ‘extinction’ (more on which below).¹⁹

Now, since the aim of this chapter is to explore what impact a *long-term perspective* should have on social risk taking—compared to a short-term perspective—the interesting question

¹⁸ As Samuelson (1963) observed, the same is not true however of a ‘standard’ expected utility maximiser, for instance, someone who thinks of monetary outcomes in terms of their own terminal wealth and whose preferences satisfies the axioms of expected utility theory. Such agents will turn down the single gamble just in case they turn down the bundle.

¹⁹ For a fascinating application of this logic to the question of how to design ethical AI, see Thoma (2022).

however is not what bundles of gambles a decision-makers should accept when offered at the same time, but rather when offered in sequence over time. Most social planners presumably expect that over their time in office they will be faced with a number of independent gambles that each have a positive objective expectation but which they would nevertheless be tempted to turn down when viewed in isolation. The presumption in favour of longtermism might then seem to imply that such planners should judge each gamble in light of the gambles that they expect they—and perhaps even subsequent social planners—will be faced with in the long-run.

However, some subtle philosophical issues now arise, concerning rational commitment and planning. Recall that we are assuming that the planner prefers to turn down each gamble when viewed in isolation. But then given that at each point in time, a decision-maker can arguably at most choose to accept or reject the gamble on offer *at that time*, we can view the decision-maker in question as being faced with a *non-cooperative game* with different *time-slices* of herself. But then the only Nash-equilibrium—that is, the only outcome where no time-slice can do better for itself in light of what others do—is one where each time-slice turns down the gamble with which it is faced. So, we reach an outcome analogous to what Hardin (1968) famously called the ‘tragedy of the commons’ where all gambles are rejected. (For a more detailed version of this argument, see Stefánsson 2023.)

Here is a different way to arrive at the above conclusion. Suppose that the decision-maker in question predicts that they will know when the *last* gamble on offer arrives. At the start of their time in office, they may predict that they will turn down that last gamble on offer; perhaps because they don’t trust their successor to accept similar gambles. Knowing this, they predict that they will turn down the penultimate gamble, and so on, all the way to the first gamble. So, by *backward-induction*, they reason themselves into rejecting all the gambles (cf. Samuelson 1963). Note however that this conclusion does not follow if the decision-maker predicts that they will never believe that a gamble on offer is the last on they will face.

It might also be possible to avoid the conclusions of the last two paragraphs by assuming that the decision-maker has fully internalised longtermism. In both paragraphs, the assumption was that the decision-maker sees themselves not only as having preferences about the outcome of the sequence, but also about the outcome at each point in time; moreover, they see themselves not as making one decision for the long-term, but several decisions that together have a long-term effect. Maybe that is not what it means to internalise longtermism. Perhaps we can instead assume that a longtermist social planner would decide, at the start of their tenure, to accept gambles like those under consideration—due to their prediction about many similar gambles being offered later—and would stick to that decision at each point in time.

Now, I am not sure that this assumption is plausibly implied by longtermism. But this assumption (or something like it) is in any case needed for the argument under consideration to establish that longtermism makes a loss or risk averse social planner more risk-prone.²⁰

If the above is what it means to internalise the longtermist perspective, then a longtermist who is either loss or risk averse (in the sense defined above) is what is called a *resolute chooser* (see, e.g., McKlennen 1990).²¹ A resolute chooser sometimes resolves to follow a plan and does so even if that means choosing counter-preferentially at some points in time. Some find such counter-preferential choice to be irrational (e.g., Stefansson 2023). However, it is far from irrational to set up institutions and structures or to pass laws that bind a social planner, in the sense that it removes options that would otherwise be tempting. Nor is it unusual. For instance, the decision by lawmakers in many countries to pass laws that make it illegal for them to meddle with the central bank's interest rate could be seen as an example of a self-binding law, while constitutions can be seen as examples of a social structure that binds lawmakers.

2(c) *Second argument: Taking other risks into account*

Let's now turn to the second argument that longtermism leads to social risk-taking. This argument is based on the observation that in terms of total risk exposure, an individual gamble makes a greater difference when evaluated from the perspective of a risk-free status quo than when evaluated in light of all the risks with which one is and will be faced.²² So, in this argument, the importance of taking the long-term perspective does not consist in the fact that the risky *options* with which one will be faced in the future may affect the value of the option with which one is faced today. Rather, this time the long-term perspective is important due to the risks with which one predicts one will inevitably—whatever choices one makes—be faced. Therefore, the subtle philosophical issues about rational commitment and planning that I discussed above do not afflict the second argument.

I will continue to focus on 50/50 gambles between losing 1 unit of wellbeing and gaining 2 units, but now instead of assuming that the social planner expects to be faced with multiple gambles of that kind, suppose that the social planner evaluates the gamble in light of all the risks with which they predict society will be faced (in whatever time frame they consider

²⁰ This raises the interesting question of whether there is any limit to how many gambles the longtermist should bundle; and, more generally, whether there is rule for deciding how many and which gambles to bundle. I thank David Thorstad for raising this question, which sadly I do not have a good answer to.

²¹ See also Thoma (2019) for a useful discussion of resoluteness in the type of decision-problems under consideration. For a recent response to Thoma (2019), see Wilkinson (ms.).

²² The argument in this subsection is inspired by Thoma and Weisberg (2017). (They however only discuss risk aversion, not loss aversion.)

relevant). Suppose that the social planner *expects* that the total wellbeing over the time-period with which they are concerned is 100 units. The precise number is of course more or less arbitrary; it just needs to be relatively high compared to the potential loss from the evaluated gamble. Suppose that the planner’s expectation is based on a normal distribution around the mean. This assumption of a normal distribution is not essential but simplifies the argument.

For illustrative purposes, suppose that the lowest aggregated wellbeing that the planner considers possible over the relevant time-period is -5 and the highest aggregated wellbeing that the planner considers possible over the relevant time-period is 205. However, the argument would hold even if we increased the numbers in both directions. Perhaps most importantly, the argument would also work if we assumed that the planner thinks that things *could go much* more badly. Finally, to keep the calculations both tractable and illustrative, let’s just work with two deviations in each direction from the mean when evaluating the status quo, which corresponds to five deviations in either direction when evaluating the 50/50 gamble. (But the argument can of course be made by assuming a continuous rather than a discrete distribution.) Table 1 then represents the status quo and the gamble.

| Probability | Status quo | Accept gamble | |
|-------------|------------|---------------|------|
| | | Win | Lose |
| 0.025 | -5 | -3 | |
| 0.025 | -5 | | -6 |
| 0.1 | 0 | 2 | |
| 0.1 | 0 | | -1 |
| 0.25 | 100 | 102 | |
| 0.25 | 100 | | 99 |
| 0.1 | 200 | 202 | |
| 0.1 | 200 | | 199 |
| 0.025 | 205 | 207 | |
| 0.025 | 205 | | 204 |

Table 1: Distribution of possible outcomes given on the one hand the expected status quo and on the other hand the expected status quo plus accepting the gamble.

Recall that both the loss averse utilitarian and the utilitarian who uses rank-dependent utility theory to manage risk will turn down a 50/50 gamble between losing 1 unit of wellbeing and gaining 2 units of wellbeing when the gamble is evaluated in isolation. However, as I shall

now demonstrate, both types of decision-makers will accept the gamble when evaluated in light of the assumed expectation of a favourable but risky future, as represented by Table 1.

Let us this time start by considering loss aversion. How a loss averse utilitarian *now* evaluates this future, with or without the gamble, depends on the *current* sum of wellbeing. To keep things simple, let us set the current sum at 0; for the purpose of the argument, the only thing that matters is that the current sum is low compared to the most likely future sums of wellbeing. A loss averse utilitarian (as formalised in subsection 2.1) who uses expected utility theory to manage risk will then judge the predicted future without the gamble as follows:

$$205(0.05) + 200(0.2) + 100(0.5) + 0(0.2) - (5 * 3)(0.05) = 99.5$$

But they will evaluate the future with the gamble as follows:

$$207(0.025) + 204(0.025) + 202(0.1) + 199(0.1) + 102(0.25) + 99(0.25) + 2(0.1) \\ - (1 * 3)(0.1) - (3 * 3)(0.025) - (6 * 3)(0.025) = 99.85$$

So, they will accept the gamble when evaluated in light of their belief that the future will most likely be better than today.

Let's then consider a standard (i.e., not loss averse) utilitarian who uses rank-dependent utility theory to manage risk. They will evaluate the future without the gamble as follows:

$$-5 + 5 * 0.95^2 + 100 * 0.75^2 + 100 * 0.25^2 + 5 * 0.05^2 = 62.025$$

But they will evaluate the future with the gamble as follows:

$$-6 + 3 * 0.975^2 + 2 * 0.95^2 + 3 * 0.85^2 + 97 * 0.75^2 + 3 * 0.5^2 + 97 * 0.25^2 + 3 * 0.15^2 \\ + 2 * 0.05^2 + 3 * 0.025^2 = 62.27375$$

So, the risk averse too will accept the gamble when evaluated in light of their prediction about the risky but, in expectation, favourable future.

In other words, both the loss averse and the risk averse will accept the gamble when evaluated in light of their belief that the future will most likely be better than today but could be worse. And surely that is the belief that a reasonable social planner who takes the long-term perspective would have. We have very good reasons to believe that there will be continued

economic growth for the foreseeable future, despite climate change and other potential catastrophes; but we should surely be open to the possibility that things will not go so well.

It may however be worth noting that even without allowing for such negative possibilities, a loss-averse utilitarian who takes the long-term perspective and who predicts social and economic development to continue will accept the gamble in question when viewed in light of their predicted future (while rejecting it when viewed in isolation). For instance, if they are *sure* that the total wellbeing over the relevant time-period is 100, then they will evaluate the gamble by its expected value. The same is not true of a REU utilitarian: if they are *sure* that the aggregate wellbeing over the relevant time-period is 100, say, then they will turn down the 50/50 gamble, since then the REU of the future without the gamble is 100, while the REU of the future with the gamble is $99 + 3 * 0.5^2 = 99.75$.

The picture may look differently if the social planner expects the future to be *worse* than the present. One complicating factor, however, is that those who are risk averse when they are gambling at or above what they see as their relevant reference point (which often corresponds to their status quo), often become risk *seeking* when they find themselves below that reference point; this is in fact one of the psychological phenomena that Kahneman and Tversky's (1979) prospect theory was designed to capture. So, a social planner who sees their *current* status quo as the relevant reference point, but expects the future to be worse—for instance, in the extreme case, sees human extinction on the relevant horizon—may become risk seeking. And similarly for those who are loss averse when gambling at, above, or below what they take to be the relevant reference points. In contrast, if either a risk or loss averse social planner of the kind we have been considering expects the future to be *precisely as good* as what they take as the relevant reference point, then they will turn down the single gamble, even when evaluated in light of their prediction about the future, for the same reason as why they turn down the gamble when considered in isolation.

Let me summarise the findings of this whole section before turning (in the next section) to discussing potential policy implications. We have seen that a policy maker who is risk or loss averse, and therefore turns down some gamble that has a positive expectation of the quantity of interest (in this case, wellbeing) if the gamble is viewed in isolation, will not turn down the very same gamble when it is viewed either in light of a favourable (albeit risky) future or in light of sufficiently many additional gambles of the same kind. Moreover, I have proposed that what it means to internalise 'longtermism' is to view each gamble in light of what one expects of the long-term future. But then we seem to have an argument that a longtermist social planner

should be more risk prone than a planner who evaluates risks, gains, and losses exactly as the longtermist planner does but nevertheless takes a more short-term perspective.

We have also seen that the arguments in question don't hold for *all* gambles with a positive objective expectation. In the next section I briefly discuss the conditions needed for the arguments in question, and compare them to some real-life risky public policies.

3. From theory to practice

Let's begin by considering the probabilistic independence assumption (which was required for the first formal argument) in relation to real-life policy decisions. This assumption will clearly not be true of all 'gambles' with which a social planner is faced. For instance, suppose that a social planner is considering lifting restrictions due to Covid-19 off preschools. We can think of this as a gamble and we can suppose that the social planner sees it as a 50/50 gamble between gaining a benefit of magnitude x and losing a benefit of magnitude $0.5x$; so, the gamble offers an expected benefit. Another similar gamble might be to lift Covid-19 restrictions of high schools. Again, we can suppose that the social planner sees it as a gamble with the same structure as the gamble to lift restrictions off preschools. But the probabilistic independence assumption will presumably not be satisfied with respect to these two gambles. The conditional probability that the gamble to lift restrictions off preschools turns out unfavourably given that the gamble to lift restrictions off high schools turns out unfavourably is presumably greater than the unconditional probability that the gamble to lift restrictions off preschools turns out unfavourably—unless, of course, we are already certain how the policy change will turn out. So, despite the previous argument, we haven't seen a reason why bundling these and similar gambles should make the social planner more risk prone than if they considered each gamble in isolation.

The probabilistic independence assumption will however plausibly be satisfied between many other policies. For instance, suppose that the social planner is considering raising the top-tax rate, and let's imagine that they see it as a risky gamble with a positive expectation (say, in terms of aggregate wellbeing). Let's moreover suppose that the social planner is considering full legalisation of recreational drugs, and also considers this to be a risky gamble with a positive expectation (in terms of aggregate wellbeing). In this case there will presumably be sufficient probabilistic independence between the two gambles.²³ For instance, the conditional probability that raising the top tax rate turns out unfavourably given that full legalisation turns

²³ If the reader disagrees, then consider instead on the one hand offering free cancer screening to everyone over fifty and on the other hand switching from left- to right-hand driving.

out badly is presumably (at least roughly) the same as the unconditional probability that raising the top tax rate turns out unfavourably.

Now, it is worth noting that there is a tension between, on the one hand, the probabilistic independence condition and, on the other hand, one of the potential benefits of taking social gambles, or doing ‘social experimentation’, namely, the information that is gained from both successful and unsuccessful experiments (Barrett and Buchanan 2023).²⁴ For instance, if (contrary to what was assumed above) the social planner were to *predict* (correctly or incorrectly) that, say, if legalising drugs turns out badly then that teaches them something very important about how people react to incentives, then there might not be sufficient probabilistic independence between the gamble to legalise drugs and the gamble to raise the top tax rate.

Nevertheless, in practice, there certainly seem to be realistic ‘social gambles’ between which there is sufficient probabilistic independence such that a risk or loss averse social planner would find each gamble to be ‘too risky’ if considered in isolation, but should nevertheless find the bundle of gambles to be acceptable when considered together. Therefore, if internalising longtermism—or simply taking the long-term perspective—means that the social planner evaluates such gambles together rather than individually, then it follows that there are some real-life examples of risky policies that a longtermist social planner would implement, even though the very same policies would not be implemented by another social planner with the exact same attitudes towards risks, losses, and risk-free outcome, but who takes the short-term perspective (in the sense of evaluating each gamble in isolation).

The example of full legalisation of recreational drugs is useful for exploring some of the other conditions needed for the arguments that longtermism leads to more risk-taking. In the introduction, I gave an informal argument that a social planner might find such an experiment to be unacceptably risky when only considering the impact on the current generation, but would think it acceptable when viewed with a long-term perspective, both because of the potential benefits in the long-run if the experiment turns out well, but also because of the knowledge gained even if the experiment turns out badly. Now, although it might be tempting to think of such an experiment as a bundle of gambles, one gamble for each generation, the argument that bundling should make a risk or loss averse decision-maker more prone to accepting the risk clearly doesn’t apply to this case, since, for instance, the probability that the experiment is harmful to one generation is not independent of how the experiment turns out for another generation.

²⁴ I am grateful to Jacob Barrett for a very useful discussion of this issue.

The condition that I however want to use this example to illustrate is the qualification that the experiment can be stopped if harmful, and that it is moreover predicted that it *will* be stopped if harmful. If that is not the case then it is unclear that the gamble should be considered less risky given a long-term perspective, since then it is a gamble with a potentially huge loss over the long-term.

How does this last condition map onto the formal arguments? The reason the condition is satisfied in the formal arguments is that it was assumed that the potential downside from each gamble was rather small in comparison to the total expected outcome (and also in comparison to the most likely outcomes); where, in the first formal argument the total expected outcome is from the bundle of gambles, whereas in the second formal argument the total expected outcome is from the predicted future (with or without the gamble). There are, however, examples of real-life public policy choices where this condition would not seem to be satisfied.

Let's take a historical example. In the 90s Sweden saw a radical social experiment where the education sector was completely opened to private actors: whereas the state would continue to fund education, anyone who fulfilled some minimal criteria could open a school and provide state-funded education. At the same time, parents were given (in principle) complete freedom to choose schools. The hope, of course, was that competition between schools—both private and public—combined with parent's preference for getting the best education for their children would result in a situation where only those education providers who could offer the best education for the budget given by the state would survive. (A similar experiment was made in various other areas, for instance, in the health-care sector.)

Today most experts seem to agree that the experiment has been overall harmful. The average quality of primary and secondary education seems to have diminished—as judged by students' performances in international tests—while the spread in education outcomes has increased (Molander 2017). Moreover, the public seems to be quite unhappy with the current system; for instance, a majority of Swedes are in favour of a cap on how high dividends private education providers can pay their owners (see, again, Molander 2017). But despite this popular and expert opinion, the proportion of private providers has only increased (op. cit.) and political attempts at capping dividends have so far proven mostly unsuccessful. One perhaps obvious reason for this is all the lobbying done on behalf of private education providers (see, e.g.,

Tylström and Svallfors 2017). The experiment created a financially very strong interest group that of course does whatever it can to maintain the current status quo.²⁵

Coming back to the formal arguments from the last section, the observations from the last paragraph may show that the experiment to open up the education sector to private providers should not be seen as a gamble where the potential downside is not really significant in comparison to the expectation about the long-term. If the harmful effects that seem to have resulted from this experiment get reproduced for each subsequent generation of school children for a very long time, then the actual long-term downside is in fact very significant.

A general lesson that emerges from the above social experiment is that the argument from longtermism to social risk-taking may not work in cases where a gamble is likely to create influential groups who have an interest in maintaining the resulting status quo irrespective of the social consequences. This would seem relevant when reasoning about, for instance, experimenting with legalising recreational drugs. Such legalisation would presumably prove quite profitable for some legal companies (e.g., today's pharmaceutical companies) that would then have an interest in maintaining the resulting the status quo—unless, of course, private actors would not be allowed to profit from the production or sale of recreational drugs. More generally, when assessing whether taking the long-term perspective supports accepting a particular social gamble, one must ask whether the potential downside of the gamble is likely to be reproduced for the coming generations.

Finally, and to conclude, I will say a few words about the condition that the gambles in question have no chance of causing a *catastrophic* outcome (in particular, extinction). For instance, in the first formal argument, about bundling gambles, if losing some gamble(s) in the bundle means that the decision-maker will not be in a position to accept further gambles, then the argument in question would not be applicable. The same is true if losing one gamble means that one won't enjoy the fruits of winning a future gamble. Similarly, in the second formal argument, about the expectation of a favourable future, if the gamble under evaluation is sufficiently large-scale such that losing it means that the future is unlikely to be favourable, then that argument for risk-taking does not apply. This means that the arguments in question do not apply to gambles with potentially catastrophic outcomes. In other words, the formal arguments that show that, under some conditions, longtermism supports risk-taking, do not show that a longtermist should be less averse to catastrophic risk than a shorttermist.

²⁵ I speculate that another reason why it has proven hard to discontinue this radical experiment is that it is generally much harder to implement policies that are seen as removing a freedom to choose than it is to implement policies are seen as increasing the freedom to choose.

In fact, under some conditions, being a longtermist should make one *more* averse to catastrophic—in particular, extinction—risks (a point made e.g. by Greaves and MacAskill 2021). As noted above, how a risk or loss averse social planner behaves if they expect human extinction whatever they do, depends on details about how gambling below the relevant reference point affects one’s attitudes to risks and losses. But it is straightforward to demonstrate that a risk or loss averse social planner who predicts the future will be favourable, *unless their society goes extinct*, will be less prone to accept gambles that risk extinction than their shorttermist counterpart. For a simple demonstration, let’s consider a social planner who predicts they will be offered a sequence of the type of 50/50 gambles that has so far been the focus of this chapter. Further, suppose that they in addition currently face a 50/50 gamble between on the one hand gaining 2 units of wellbeing and on the other hand ‘extinction’, interpreted as them not being in the position to accept further gambles. Finally, suppose that they don’t expect any potential for gains or losses except for, first, those from the aforementioned sequence and, second, from the gamble currently on offer. Whether risk or loss averse (in the sense defined above), the social planner may then accept this gamble if they are so shorttermist that their planning horizon only includes, say, the next two gambles in the sequence. For then they may reason that the future is so bleak that they might as well take the gamble. In contrast, if their planning horizon includes the next hundred gambles, say, then they will reject the extinction-risk-gamble, since it threatens preventing them from accepting over time an incredibly favourable bundle of gambles. So, a longtermist who expects the future to be favourable will be more averse to extinction risk than their shorttermist counterpart.²⁶

Bibliography

- Barrett, Jacob and Buchanan, Allen (2023), “Social Experimentation in an Unjust World”, in David W. Sobel & Steven Wall (eds.), *Oxford Studies in Political Philosophy Volume 9* (Oxford University Press), 127-152.
- Buchak, Lara (2013), *Risk and rationality* (Oxford University Press).
- Buchak, Lara (2017), “Taking Risks Behind the Veil of Ignorance”, in *Ethics* 127/3: 610-644.
- Greaves, Hilary, and MacAskill, William (2021), “The Case for Strong Longtermism”, GPI Working paper 5-2021.
- Hardin, Garreth (1968), “The Tragedy of the Commons”, in *Science*, 162/3859: 1243–1248.

²⁶ This chapter has been presented at the 9th Oxford Workshop on Global Priorities Research and at the PPE seminar at the Institute for Future Studies in Stockholm (both in March 2022). I am very grateful for the comments and suggestions I received at these two occasions. Johanna Thoma read a draft of the chapter and provided very helpful written feedback. Finally, thanks to the editors of this volume and a referee, for useful comments and suggestions.

- Kahneman, Daniel and Tversky, Amos (1979), “Prospect Theory: An Analysis of Decision Under Risk”, in *Econometrica* 47/2: 263-291.
- McClennen, Edward F. (1990), *Rationality and dynamic choice: Foundational explorations* (Cambridge University Press).
- Molander, Per (2017). *Dags för Omprövning*. Expertgruppen för Studier i Offentlig Ekonomi 2017:1 (Regeringskansliet).
- Rabin, Mathew & Thaler, Richard H. (2001), “Anomalies: Risk Aversion”, in *The Journal of Economic Perspectives* 15/1: 219-232.
- Rabin, Matthew. (2000), “Diminishing Marginal Utility of Wealth Cannot Explain Risk Aversion”, in Daniel Kahneman & Amos Tversky (eds.), *Choices, values and frames* (Cambridge University Press), 202–208.
- Rothschild, Michael and Stiglitz, Joseph (1970), “Increasing risk: I. A definition”, in *Journal of Economic Theory* 2/3: 225-243,
- Samuelson, Paul (1963), “Risk and Uncertainty: A Fallacy of Large Numbers”, in *Scientia* 98: 108–113.
- Stefánsson, H. Orri & Bradley, Richard (2015), “How Valuable Are Chances?”, in *Philosophy of Science* 82/4: 602-625.
- Stefánsson, H. Orri & Bradley, Richard (2019), “What Is Risk Aversion?”, in *British Journal for the Philosophy of Science* 70/1: 77-102.
- Stefánsson, H. Orri (2023), “The Tragedy of the Risk Averse”, in *Erkenntnis* 88/1: 351-364.
- Svallfors, Stefan, and Tyllström, Anna (2017), “Lobbying for Profits: Private Companies and the Privatization of the Wellbeing State in Sweden”, Institute for Futures Studies Working Paper 2017:1.
- Thoma, Johanna & Weisberg, Jonathan (2017), “Risk Writ Large”, in *Philosophical Studies* 174/9: 2369-2384.
- Thoma, Johanna (2019), “Risk Aversion and the Long Run”, in *Ethics* 129/2: 230-253.
- Thoma, Johanna (2022), “Risk Imposition by Artificial Agents: The Moral Proxy Problem”, in Silja Vöneky, Philipp Kellmeyer, Oliver Müller & Wolfram Burgard (eds.), *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives* (Cambridge University Press), 50-66.
- Thoma, Johanna (2023), “Taking Risks on Behalf of Others”, in *Philosophy Compass* 18/3: 1-13.
- Wilkinson, Hayden (ms.). “Risk Aversion and the Not-so-long Run”, unpublished manuscript.