eidos
Fecha de recepción: 1 de junio de 2023
Fecha de aceptación: 18 de septiembre de 2023
DOI: https://dx.doi.org/10.14482/eidos.41.407.111

Review of the book: Tobias Baumann, Avoiding the Worst (2022, 105 pages)

Mat Rozas
ORCID ID 0000-0001-8161-188X
University of Santiago de Compostela
mat.rozas@usc.es

In this book, Baumann examines key questions about how to reduce risks of future suffering (s-risks). He defines these as risks that the future contains astronomical amounts of total suffering on an unprecedented scale. Given that these risks are not negligible, and that our present acts could increase or decrease their likelihood, Baumann argues that focusing on s-risk prevention is a sound priority. To support this, he appeals to two other quite plausible, but rarely considered, assumptions. First, temporal impartiality: the view that whether we exist in some concrete timeline is as arbitrary and morally irrelevant as where we live or the species to which we belong. Second, the expected value of the long-term future: in the long-term future, there will be many more individuals than in the present and the short-term future (if only because the long-term future will last many orders of magnitude longer).

To clarify the problem at stake, Baumann distinguishes s-risks from other risks and undesired future scenarios, such as existential risks (x-risks) and dystopias. Dystopias are different from s-risks because not every dystopian future entails astronomical amounts of suffering.[1] For their part, x-risks are risks that human beings (or their descendants) do not develop their full potential. This may happen because humanity goes extinct or for other reasons.

[1] Huxley, A. (2022), *Brave New World*, Penguin Books.

Although there is some overlap between certain s-risks and x-risks, they differ.

Baumann distinguishes in great detail between different types of s-risks, including agential, natural, and incidental s-risks. Agential s-risks are those caused intentionally by an agent (e.g., a sadistic dictator who enjoys causing astronomical amounts of pain). Natural s-risks are caused by the processes of the universe without external intervention (e.g., wild animal suffering). Lastly, incidental s-risks are the unwanted result of some process that is highly beneficial for some agents but very harmful for certain sentient beings (e.g., animal exploitation). Next, Baumann differentiates between three additional categories of s-risks: known and unknown s-risks, influenceable and non-influenceable s-risks, and s-risks that affect humans, nonhuman animals, and artificial sentient entities. Known s-risks are those s-risks that we can think about today (e.g., expanding wild animal suffering). In contrast, we cannot imagine how unknown s-risks may come about since we cannot currently conceive of these risks (e.g., people in ancient Greece could not conceive of the risks of nuclear warfare). Influenceable s-risks are those s-risks that we can tackle. Baumann thinks that we have to focus entirely on this kind of s-risks because we cannot do anything about non-influenceable s-risks (e.g., we cannot try to tackle s-risks that may happen in unreachable parts of the universe).

Baumann argues that s-risks may be astronomical in at least two ways. First, future technology may allow us to colonize space. Consequently, in the future, there could be sentient beings throughout the galaxy. Hence, the number of sentient beings populating the galaxy could be truly astronomical. Second, future technology may also make creating artificial sentient beings feasible. Given that creating large amounts of artificial sentient beings could be very cheap and profitable (e.g., experimenting with astronomical amounts of artificial sentient beings could be the cheapest and most reliable way to obtain the most accurate

EIDOS N.º 41 (2023) PÁGS. 275-280
ISSN 2011-7477

results in some scientific areas), the scale of the problem would be very high. Thus, if technology develops faster than our ethical concerns, the consequences of this differential progress could be catastrophic on an unprecedented scale.

Baumann argues that those accepting some form of suffering-focused ethics, according to which preventing negative things such as suffering from occurring has priority over promoting the occurrence of positive things, will lead us to be particularly concerned about s-risks. However, he also points out that those who endorse other views (such as total utilitarians, for instance) would have reasons to care about them too since s-risks are non-negligible, and any plausible view would need to be concerned about reducing suffering. Moreover, Baumann develops a further argument in favor of focusing on s-risk reduction drawn from expected value theory. According to the simplest version of this view, the expected value of a prospect can be calculated by multiplying the assigned probability and the assigned net value of such a prospect. According to Baumann, if we do not assign a very low probability to s-risks happening in the future (he assigns a probability no lower than 0.001 to this possibility), given the dimension of the possible catastrophes that could come about if s-risks materialize, any plausible expected value theory will entail that preventing the worst possible outcomes is one of our most important priorities. However, Baumann warns us that this could not be so simple, given that the future is highly uncertain. Therefore, it is very difficult to predict the long-term effects of our actions. Furthermore, we have to take into account that, since there will be many agents trying to shape the far future, our efforts to reduce s-risks could be washed away because of this. Despite these problems, Baumann thinks that current efforts to reduce s-risks are very valuable because they are pioneering efforts in a highly neglected area and, thus, are very likely to be effective in preventing s-risks.

Finally, Baumann examines several ways in which we can get involved in reducing s-risks. He considers that capacity and move-

EIDOS N.º 41 (2023) PÁGS. 275-280
ISSN 2011-7477

ment building are sound possibilities to increase future capacity for action. However, he also considers other courses of action, such as improving political institutions. Although improving political institutions may be difficult to achieve, establishing better political systems (e.g., progressively replacing presidential systems with parliamentary systems) might make conflicts and instability less likely. Therefore, better institutions may make s-risks less probable. Baumann nonetheless pays more attention to moral advocacy, as he claims that one of the factors that will shape the far future the most will be the set of values that future individuals will endorse. As such, we can work today to ensure that people care about reducing s-risks in the future. One way in which this could be done would be by working on moral circle expansion since s-risks are less likely to happen if the relevant agents in the future fully take into consideration the interests of every sentient being that may be affected by their actions. Nevertheless, Baumann believes that moral circle expansion could backfire if it encouraged conflict or if it made future agents endorse the wrong values. However, this claim seems controversial, given that the scenarios where this could happen seem significantly less plausible than those where moral circle expansion would have a very positive impact.

Baumann also presents some guidelines for dealing with s-risks related to the development of technologies that can significantly increase the number of sentient individuals that there may be in the future. He argues that since we cannot realistically halt technological progress, the best we can do is to ensure that there is no differential progress in these two areas. Consequently, developing AI safety appears to be one of the best ways to tackle this problem. Moreover, Baumann argues that the development of these technologies must not cause political instability. This is because such instability could generate dynamics that increase the likelihood of certain s-risks materializing due to issues such as arms races, malevolent agents having access to these technologies to cause harm, etc.

Overall, Baumann's examination of the topic is very thorough and insightful, and I find myself agreeing with him most of the time. Nevertheless, I believe that several claims in Baumann's book should be qualified. I will focus here on only three of them.

The first one has to do with the definition of s-risks as risks that the future contains *astronomical* amounts of *total* suffering on an *unprecedented scale*. This claim is problematic for two reasons. First, why should we consider only future scenarios that could bring about astronomical suffering on an unprecedented scale? It is difficult to determine the exact point at which one possible future scenario becomes an s-risk. Some cases are clear, but others are not. For example, the fact that tomorrow I might hurt my finger when I wake up does not constitute an s-risk, whereas the fact that in the future suffering may spread throughout the galaxy constitutes an s-risk. However, did the development of industrialized animal exploitation constitute the materialization of an s-risk in the past? According to Baumann's definition, it did not, since the total suffering caused by animal exploitation is not astronomical on an unprecedented scale. I find this definition too restrictive. Thus, I prefer to define s-risks as possible future scenarios that may bring about substantial amounts of suffering.[2]

Second, focusing on total suffering seems to have some unsound implications. For instance, imagine that we colonize outer space in the future. We expand throughout the universe, such that the number of sentient beings in the universe becomes exponentially high. Moreover, imagine that the lives of all these individuals are amazing, except for a very mild unpleasant sensation that they feel for some minutes. According to Baumann's definition of s-risks, this possible future is a very serious s-risk. We have a few options at our disposal to avoid this implication. One would be to focus instead on average well-being. Since average views are

[2] Another possible definition that is more neutral concerning pluralist axiologies is that of possible future scenarios that are highly disvaluable in expectation.

sensitive to population size, they avoid this implication. Another option would be to focus on reducing the number of net-negative lives in the future, that is, lives in which the bad features outweigh the good features. All of these options have unpalatable implications.[3] However, it is far from obvious that the implications of the total view are the ones that are the least unsound.

Third, there is a problem with expected value theory relevant to the points Baumann examines that he nonetheless overlooks throughout the book, known as the problem of fanaticism. The problem can be roughly portrayed in this context as the implication of expected value theory that, no matter how small the probability of an s-risk happening, as long as such a probability is not zero, and provided that the assigned net value to that possible scenario is low enough, avoiding that such a prospect materializes should become our foremost priority. Baumann endorses this view, which most people find very hard to accept. Unfortunately, we do not yet have a good non-fanatic alternative to the expected value theory. Nevertheless, even those who find fanaticism unacceptable can endorse s-risk reduction. Provided that we do not understand s-risks as restrictively as Baumann does, we can focus on reducing s-risks that are more likely to materialize, even if their magnitude is less severe.[4]

Despite these minor points, Baumann's book is an excellent essay that tackles many relevant questions in a very important and neglected area. His book is groundbreaking in a field in which much work remains to be done.

[3] Arrhenius, G. (2000a). "An impossibility theorem for welfarist axiologies". *Economics and Philosophy*, 16, 247–266.

[4] Buchak, L. (2013). *Risk and Rationality*. Oxford: Oxford University Press.

EIDOS N.º 41 (2023) PÁGS. 275-280
ISSN 2011-7477