Is most expected suffering due to worst-case outcomes?

Tobias Baumann

January 2020

For those interested in reducing future suffering, a natural question to ask is: should we focus our efforts on preventing the worst outcomes? Or is (expected) future suffering distributed more broadly over a wide range of plausible futures? This is relevant because it determines the extent to which we accept "Pascalian" wagers about how the future could go very badly.[1] If worst-case outcomes dominate, then it pays to think about specific scenarios and how to prevent them; if not, we should pursue robust strategies that are good over many or all scenarios.

In this post, I will try to define the question more precisely, and give arguments for and against the hypothesis that extreme outcomes should dominate consequentialist calculations. I will refer to this as the *extreme outcome dominance hypothesis* (EODH).

I will mainly discuss future disvalue as that's what I'm most interested in. However, the same question can be asked for the distribution of future value, and the analysis is directly transferable (although including both signs adds some mathematical complications).

Clarifying the question

Let $(\Omega, \mathcal{F}, P)$ be the probability space that describes possible futures and their likelihood. For $\omega \in \Omega$, let $S(\omega)$ be the amount of suffering in future scenario ω. The expected amount of future suffering is

$$E(S) = \int_{\Omega} S(\omega) dP(\omega).$$

For $0 \leq \alpha \leq 1$, let $\Omega_\alpha \subset \Omega$ be the α-quantile of outcomes that contain the most suffering. Then, $S_\alpha = \int_{\Omega_\alpha} S(\omega) dP(\omega)$ is the expected amount of suffering in the worst α-quantile of outcomes.

We can now quantify the EODH in various ways. Choosing some (arbitrary) α - say, $\alpha = 0.01, 0.05, 0.1$ - we can consider the fraction S_α / S_1 of expected suffering that's due to the worst $1\%, 5\%$, or 10% of outcomes. We could now choose arbitrary thresholds to obtain a binary definition of the EODH; e.g. "more than 50% of suffering is due to the worst 1% of outcomes" ($S_{0.01}/S_1 > 0.5$), or "more than 80% of

[1] However, such wagers are often made in the context of specific ways the future could go very badly, which overlooks unknown ways in which the future could go very badly. Thus, one can accept EODH and still prefer robust strategies - and reject wagers favoring particular scenarios - due to great empirical uncertainty. (HT Magnus Vinding for this point.)

suffering is due to the worst 5% of outcomes" ($S_{0.05}/S_1 > 0.8$).

Alternatively, we could consider common income inequality metrics, such as the Gini coefficient. (These are usually applied to income distributions, but of course the math also works for any other distribution.) The Gini coefficient can be calculated as follows:

$$G = \frac{2}{S_1} \int_0^1 S_\alpha d\alpha - 1.$$

The EODH can now be interpreted as "G is close to 1" - say, at least 0.75.[2]
However, considering ex-post future suffering may not be the most action-relevant way to look at this question. In particular, if we are clueless, then the ex-post distribution over future suffering may be very unequal simply due to uncertainty about the size of the universe or multiverse, about whether we live in a simulation, or similar unresolved questions. This is arguably not what we are most interested in.

It might be more fruitful to instead consider a probability distribution over future world states at a specific point.[3] For instance, assuming that civilisation will eventually reach a steady state, the outcome space Ω could describe possible world states at that point. This includes e.g. how powerful different values are, to what degree the relevant actors cooperate, or how technologically capable civilisation is. S would describe how much suffering we would ex ante expect to result from that world state, from our current epistemic position (i.e. being uncertain about many "big picture" questions). We could then analogously consider the Gini coefficient or other metrics.

It is also worth pointing out that it is possible that nothing, or not much, can be done about worst-case scenarios, in which case we should focus on other scenarios anyway, even if EODH is true. (Perhaps the reason why some scenarios contain much suffering is that it is hard to prevent suffering in such scenarios.) Generally speaking, there is as much reason to focus on tractable scenarios as there is to focus on futures that contain a lot of suffering.

Theoretically, we should consider, for any possible intervention, the marginal suffering reduction over all scenarios, weighted by their probability mass. But this is often infeasible. So it still makes sense to consider heuristics of the form "we should focus on worst-case outcomes".

Do extreme outcomes dominate in general?

For purposes of this discussion, I will consider the EODH regarding steady state world descriptions, as outlined above. As a first step, I will examine whether similar claims are true for other bad things in general. This will inform a prior for how plausible EODH is. (I will later discuss if and how future suffering differs from other cases.)

Consider some probability distributions that are frequently used to model a broad variety of phenomena:

[2] For reference, the Gini coefficient of global income is estimated to be between 0.61 and 0.68.

[3] Alternative reference points are: a) the world state at an upcoming pivotal point in history, such as transformative AI, or b) the world state in 100 or 200 years.

- The normal distribution is most common, as it arises from the sum of many factors with independent variation. There appears to be no closed form expression for its Gini coefficient - see here for the Lorenz curve - but it's clear that the normal distribution isn't dominated by extreme outcomes. (Note also that the normal distribution can be negative, and may therefore not be a suitable model when considering risks of bad outcomes.)
- The log-normal distribution arises from the product of many factors with independent variation. Its Gini coefficient is $G = \text{erf}(\sigma/2)$ (Source), where erf is the error function. For instance, with $\sigma = 1$, the Gini coefficient is 0.52, with $\sigma = 2$ it's 0.84.
- The exponential distribution has a Gini coefficient of 1/2. (Source)
- A Pareto distribution with Pareto index α has Gini coefficient $\frac{1}{2\alpha - 1}$, for $\alpha \geq 1$. (Source).

Overall, it seems that distributions with high Gini coefficients (i.e. distributions that would support EODH) are not particularly common, but also not unheard of. The most plausible examples are the log-normal distribution with high variance (e.g. $\sigma = 2$ or higher), or Pareto distributions with Pareto index close to 1. (In addition, there are some heavy-tailed and fat-tailed distributions with infinite or undefined expected value or variance.)

Next, consider some real-world examples, asking whether worst-case outcomes are generally most worrisome:

- War: The literature on warfare suggests that the badness of wars, measured by the number of casualties, follows a Pareto distribution (1, 2). See here for some estimates of the parameter value. So there is fairly strong evidence for the idea that war casualties are heavy-tailed, and it seems at least plausible (but not completely obvious) to focus on preventing the worst wars.
- Natural disasters: Damage or casualties from natural disasters is also often modeled using heavy-tailed distributions (see e.g. here). It seems right to focus on severe disasters, though it is worth pointing out that people seem to not focus on conceivable *very* extreme disasters (e.g. an earthquake that kills 50% of the population).
- Crime: Some crimes are much worse than others (e.g. homicide vs. theft). In terms of how much to personally worry about more vs. less severe crimes, it seems non-obvious either way whether the worst crimes cause the most disvalue. Police forces tend spend significant fractions, but not all, of their resources on the worst crimes.

In conclusion, it seems that focusing on worst-cases is moderately supported by "common sense", over a broad class of domains. However, this is usually not taken to extremes. People do not constantly freak out over very bad but unlikely scenarios, and in many cases, both worst-cases and less severe outcomes merit attention.

Is the situation different for future suffering?

Of course, the case of future suffering is plausibly very different from these examples. What reasons are there to expect that EODH holds to a larger, or smaller, degree for future suffering, compared to the prior

of focusing on worst-cases only to a moderate degree?

One could argue that most futures contain near-zero suffering relative to the physical maximum. Therefore, it's particularly plausible that the worst outcomes are many orders of magnitude worse, making it more likely (relative to the prior) that EODH holds true.

However, I'm not sure if the span of outcomes is actually that much larger for future suffering compared to other cases. The possible range of casualties in a war also spans many orders of magnitude, and murder is many orders of magnitude worse than petty theft. Also, when looking at expected suffering given a world state (see above), I think it's not clear whether this spans more than, say, 6 orders of magnitude (considering unknown s-risks etc.).

Another possible line of argument is that we have found specific mechanisms for how tail scenarios with extreme amounts of suffering could come about - namely, threats and other agential risks. But it seems very unlikely that large-scale threats will materialise in the future. It is unclear whether the degree of plausibility is higher than in other cases (where one can also construct narratives of supercatastrophes or World War III).

Another difference is that there is much more model uncertainty when it comes to future suffering. That could leave room for Pascalian wagers, although one can also argue that uncertainty smooths out the expected value (at least in the ex ante formalisation). The overall upshot seems unclear.

Finally, it is instructive to consider contemporary (human-caused) sources of suffering, as that is closer to future suffering than the other examples.[4] I think this does not really support EODH. It has been suggested that factory farming is by far the largest source of suffering; but this is actually not very clear when considering e.g. industrial fishing, entomophagy and other forms of insect farming, or animals that are harmed incidentally through human activity.

Conclusion

I've argued that in general, the idea of focusing on extreme outcomes is only moderately supported by common sense and real-world examples. There are some arguments for why EODH holds for future suffering in particular, but they don't seem that strong - so I wouldn't deviate very much from the prior. (This is for the ex ante distribution over world state descriptions; it does seem more likely that the ex post distribution is dominated by extreme outcomes.)

This suggests a balanced approach of dedicating a (significant) fraction of the available resources to worst cases, and the rest to other outcomes.

In terms of a concrete mathematical model, it seems plausible to model the amount of future suffering as a product of many independent factors (e.g. degree of cooperation/conflict, distribution of values), resulting

[4] This is somewhat different matter though: it's about whether sources of suffering are heavy-tailed, while this post is about whether scenarios are heavy-tailed. However, this is plausibly highly correlated - contemporary sources of suffering can be thought of as the realisation of certain scenarios from the perspective of past effective altruists.

in a lognormal distribution. The question, then, is how large the σ-parameter is.