Valuing the impact of civil emergencies

Toby Ord & Owen Cotton-Barratt*

The 2013 National Risk Register outlines the estimations of the impacts of civil emergencies as follows. Each of the following dimensions is scored from 0 to 5, then the mean of these is used.

- fatalities
- illness or injury
- social disruption
- economic harm
- psychological impact

We have three main concerns with the current method:

1) The scale might be too restricted.

2) The scales might not match up with each other.

3) The mean is not the right way to average these.

Fixing these should be fairly easy and provide a significant improvement in the accuracy of the impact scores and thus the priority setting.

1. The scale is restricted

The scale for each dimension and for the overall impact runs from 0 to 5. Examples used in the report suggest that this is meant to be a roughly logarithmic scale, with each point being several times the impact of the one before. This would be a good way of doing it. Among other things, it would mean that the contour lines of equal expected badness on Figures 1 and 2 in the NRR would be straight lines. If each impact level were defined to be 10 times the one before, then this would match the probability categories and the contour lines would have a 45° slope, which would be especially intuitive (see diagram on next page).

However, even if each level is ten times the stakes of the one before, it doesn't seem like 5 levels would be enough to fit all the relevant risks. This might be true even with conventional risks. For instance, we assume 1,000 deaths would be at least a 1 on the fatalities scale, but then a 5 would be 10,000,000 deaths. In this case, Spanish

* University of Oxford.

flu at 20–40 million deaths might really be a 6 (or more if 1,000 deaths would be score higher than 1).

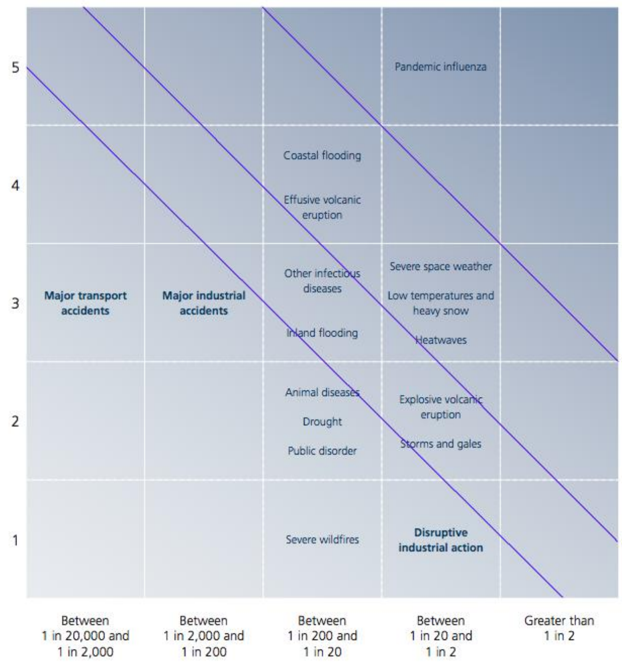

Figure 2 reproduced with some contour lines of equal expected value added (known as indifference curves in economics). Their slope here assumes a 5 is ten times worse than a 4 and so on.

This becomes more acute if dealing with unprecedented risks from emerging technologies. It is possible that some deliberate or accidental uses of synthetic biology could kill hundreds of millions (or billions) of people, meaning it should perhaps be scored as a 7, 8, or 9. The likelihood of this happening in the next five years is extremely low (so it might not be relevant to the next NRR), but it appears more plausible over a 20 year time frame, so it could be relevant to the NSRA. Even then, the likelihoods will be very low. However the stakes may well be high enough to justify them being major risks of interest even with low likelihoods. If we cap the stakes at level 5, it is difficult to see this. Indeed an event like this could be one of the most important under consideration even if its likelihood in the next five years were below the lowest current category of 1/20,000 to 1/2,000. A cap of 5 creates an artificial penalty for high stakes low probability risks. While people do not all agree that we should simply look at expected values for low-probability high-stakes events, when there is disagreement it is usually because we should be risk-averse. This makes these high stakes events even more important to track.

One reason to be sceptical of a need for extending the scale would be if a 5 on each of the different dimensions exhausted the amount of damage that could be done to the UK. For instance if a 5 on fatalities was the entire UK population or a 5 on economic harm was the entire economy. We doubt that 5s currently go so high, but even if they do, we think we need to worry about international damages too. We don't think we need to worry about all international damages, as they are often not relevant for the purposes these reports are used for. However if the international damages are *caused* by an event in the UK (such as some dangerous research), then we need to factor them into an analysis when thinking about whether or how to regulate such research.

2. The scales might not match up with each other

Averaging the scales together makes an implicit assumption that the scales line up properly with each other. Technically it means that the difference in impact of emergencies at two different points on one dimension are roughly equal to the difference in impact between those same points on each other dimension. If the scales are being used intuitively rather than according to certain quantitative rules, then this is very unlikely to be true. For instance if using the rule of thumb that a 5 on each dimension is 'as bad on that dimension as we can realistically conceive of', then we are probably in trouble as the worst cases we can conceive of for fatalities might be a lot worse than for injuries or other things (e.g. there is probably no number of injuries in the UK that would be as bad as 10 million deaths).

If a 4 on one dimension is as bad as a 5 on another dimension, then that means the first dimension is being weighted too little in the overall average or that the second dimension is being weighted too much. Sorting this out could make a big difference in the relative impact estimates of different emergencies for relatively little effort.

We recommend that the scales be matched up with each other using the standard values the government uses in cost-benefit analysis (e.g. the Green Book), though this might be difficult for some categories. Even careful non-formal assessment might help a lot. For instance if 4 on fatalities is a range of 1,000,000 to 10,000,000 deaths, then before assigning an emergency a 4 on social disruption it would be important to consider whether the social disruption is roughly as bad as 1,000,000 to 10,000,000 deaths. It is plausible that for some of the dimensions 4s and 5s just shouldn't be assigned.

3. The mean is not the right way to average the different scores

Averaging using the mean implicitly assumes that an emergency with a 3 and a 3 on two dimensions is as bad as if it had a 1 and a 5 on those dimensions. However, if the scores are logarithmic estimates of the damages on each dimension (as we think they currently are and that they should be) then a 1 and a 5 would be much worse than two 3s. For example, if a 1 is 1,000 fatalities a 3 is 100,000 and a 5 is 10,000,000, then the difference between a 5 and a 3 greatly exceeds the difference between a 3 and a 1. Suppose that the injury scale is set up to be 10 times the number of people as the fatalities scale. In this case a 5 in fatalities and a 1 in injuries would mean 10,000,000 deaths and 10,000 injuries, which is much worse than a 3 in each (100,000 deaths and 1,000,000 injuries).

The mathematical solution is to convert the scores to a linear scale, average those, then convert back. For instance, if the scores are chosen so that each number is ten times worse than the one before, then to average x and y, we average 10^x and 10^y then take the log base 10 of the result. The average of 3 and 3 would be 3, while the average of 1 and 5 would be 4.7. Even the average of a 5 with four 0s would be 4.3. So if the above method is deemed too complex, you would get a decent approximation just by using the maximum score over all five dimensions as the 'average'. This would work much better than taking the mean.

The only hesitation we have in suggesting this is if the estimates are already very unreliable and outlier scores for a dimension are mainly due to bad estimates rather than real differences. If so, then it is possible that the benefits of taking the mean in dampening down any outliers exceeds the costs of the fact that this is mathematically the wrong kind of average. We doubt this is so, but mention it here just in case.

Summary

We think there are several precise and important problems in how the overall impact scores are currently calculated. Fixing these would greatly increase the accuracy of the estimates and hence the priority setting using NRR or NSRA. These could be quite easily fixed by:

1) Allowing the scale to be open ended.
2) Making sure that a 3 on one is as bad as a 3 on the others and so forth. This would ideally be done with some simple cost-benefit analysis, but even some intuitive methods could avoid the biggest problems here.
3) Taking the right kind of averages (either via exponentiating the scores before averaging them, or via just taking the maximum of all the dimensions).