# Dealing with Uncertainty in Ethical Calculations of Existential Risk

Kieran Marray

St Catherine's College, University of Oxford

Acknowledgements

I owe a huge debt of gratitude to Philipp Koralus for this paper, with whom I talked about the main argument on several occasions. This was incredibly helpful, and the ideas within developed significantly thanks to these discussions. I should also give special mention to the participants of the 'Risk and The Culture of Science' workshop hosted by CSER for their robust debate of the conclusions and giving me a chance to present it. This also helped me develop the argument within the paper significantly. Special thanks in this regard is owed to Kai Spiekermann, Elizabeth Baldwin, Martin Glick, Matthew Rendall, Rupert Read, and Hilary Greaves.

## I. Introduction

Quantitative decision procedures, typified by expected utility, are inadequate decision procedures for measuring and ranking existential risks (defined in line with Bostrom[1]). The issue comes from what I shall refer to as the 'anti-knowledge problem', which makes this type of decision procedure untenable. This is because it leads us to attach normative significance to aiming to prevent risks which we cannot know not to be possible because of the infinite disutility that they would cause if they did indeed happen. This in turn leads to an inability to rank existential risks against one another in a satisfactory way, which is a necessary function of a decision procedure. I argue instead for a qualitative decision procedure to be used in certain cases through comparing possible worlds. This is a strong approach as not only does it solve the anti-knowledge problem but it also allows one to assess interactions between agents, which is a key practical feature of decision making as few decisions in the real world are truly independent. It also has some interesting implications for how we treat existential risks, which I shall lay out.

There are some underlying premises in the following argument, hopefully which should be uncontroversial. The first is that the study of the ethics of existential risk should be a practical one. By asking these questions, I assume that we are aiming at guiding discussion and policy around possibilities which could cause the end of our species and not simply engaging in abstract debate. Hence instead of debating some abstract notion of what 'humanity'[2] should do, I shall assess the ethical decision problems faced by agents such as individuals or nation-states. These are the actors who will be making ethical decisions on relevant questions like climate risk, so a practical analysis of the topic must focus on this level of decision making rather than some conceptual aggregate. The second is that when considering existential risks, we should consider the impact upon future generations. This seems to be a relatively intuitive assumption; if humanity as a species is destroyed, then future generations will be affected as they will be unable to come into existence. If an actor is affected by a decision, then it should be considered in any reasonable consequentialist ethical decision procedure. Therefore it seems as if we should take the impact upon future generations into

[1] Bostrom, Nick, *Existential Risks*, Journal of Evolution and Technology, 2002

[2] As many philosophers like Nick Bostrom appear to

account when making decisions to do with existential risk. The third is that there is some potential, however small, for infinite future generations of humanity. One can imagine a future where humanity has escaped the solar system before the expected collapse of the sun in a billion years' time and hence continues to exist in some form ad infinitum. As it is not an impossibility, this should be taken into account when assessing the impact of existential risk on future generations.

## II. The 'Anti-Knowledge Problem'

A common approach in the current study of existential risk is what I shall refer to as the 'quantitative approach' of using expected utility. Briefly put, this is that when ranking the gravity of one thing against another in order to make some form of decision between the two (such as whether to allocate resources to preventing one or the other), one should compare expected utility values. This is the probability of the outcome occurring multiplied by its disutility/utility, and is the average outcome of the event. For the study of existential risk, there is essentially little difference between whether one uses disutility of occurrence or utility of prevention as one's measure; I shall consider disutility of occurrence as it is conceptually clearer but one could easily use utility caused by preventing the risk instead. For example, if one were looking at something with a probability of 0.1 which would cause a disutility of 10,000, the expected utility would be -1000 (or the utility caused by prevention would be 1000). Normally the qualitative approach would lead one to say that when confronted with a set of possibilities the 'worst' possibility is the one with the greatest expected disutility. One should therefore make a decision based on this, usually to aim to prevent that risk over whatever the other possibilities were[3].

The introduction of uncertainty into questions about existential risk is fatal for the quantitative approach to assessing these questions. It raises a problem which I shall call the 'anti-knowledge problem'. There are other issues with uncertainty of probability and outcome, but these can be accommodated and hence I shall not elaborate upon them. However the anti-knowledge problem is certainly fatal for any practical application of the approach, which can be seen when one attempts to make decisions between acting to prevent two existential risks with it.

The anti-knowledge problem is, simply put, the issue brought about by uncertainty over what we are able to rule out as a possible existential risk. The quantitative nature of the approach means that one is forced due to uncertainty over probability to give weight to seemingly absurd risks, and therefore it collapses as a useful heuristic. This is because we can never be fully certain that the probability of such an event is zero. Imagine the following scenario, which illustrates this well.

*You are sitting in the Centre for the Study of Existential Risk minding your own business. In runs a person looking bewildered and panicked. After catching her breath, she explains to you that there is an invisible spaghetti monster on the far side of Mars which is going to destroy the world. It is however undetectable to any known scientific instrument.*

There is no way of you knowing whether the spaghetti monster is in fact there or not, and in fact it is almost absurd to assume that it is. However if it is then all of humanity will be wiped out. You do an expected utility calculation to work out whether you should take it seriously. You know that the probability of this spaghetti monster existing is incredibly tiny, but it cannot be zero. This is because we simply might not have developed the means of detecting it yet, and we cannot know that this isn't the case. We cannot know all of the things that we do not know yet by definition, otherwise we would know them. To know that the risk of the spaghetti monster was in fact zero, we would have to

[3] See Shahar Avin's discussion of catastrophic risk for a typical example of this approach

know that what we do not know (what I shall refer to as our 'anti-knowledge' in line with Taleb[4]) does not include some means of detecting this spaghetti monster. However, the disutility that the spaghetti monster would cause would be infinite. This is because that if it were the case that the spaghetti monster did destroy the world, then it might destroy infinite future generations of humans. Therefore when you do the expected value calculation you realise that the expected disutility of the spaghetti monster is infinite. It therefore must be an incredibly pressing risk! This in a nutshell shows why expected utility is an inadequate heuristic to use in decision making over existential risk. Imagine any risk you want, pluck out your wildest dream of what might wipe out all future generations. By the expected utility heuristic that is an incredibly pressing risk. Even though you just made it up, as we cannot know the limits of our anti-knowledge we cannot be certain that the probability of it being an existential risk is zero. There might be some process that we do not yet know about that makes it a risk. Therefore if performing an expected utility calculation we must assign it some probability greater than zero. This might be incredibly small, but what matters is that it cannot be zero, even though the risk was just made up. If a risk is truly existential, then its disutility should be infinite. This is because it removes the potential for there to be infinite future generations of humans. Due to the nature of infinity, anything multiplied by it is also infinite. Therefore any such expected value will be infinite, and so seem incredibly urgent to deal with. It is intuitively obvious though that one should not give any credence to made-up risks, and so the quantitative approach is rendered absurd.

This intuition is demonstrated even more when one considers how the quantitative approach uses expected utility; that is the implications when it is used as a ranking heuristic for assigning gravity to existential risks. Imagine again the spaghetti monster scenario, except with one critical difference.

*It has been announced that a doomsday device has just been discovered. This device will certainly destroy the world, and you only have a limited time to prevent it. You have a set of limited resources to assign to working out how to prevent existential risks, and which give you the same probability of preventing any existential risk from occurring.*

As an agent, you have an ethical decision to make as to which risk to try and prevent. Do you use your resources to try and prevent the spaghetti monster or the doomsday device? It is clear from our moral intuition that one should try to prevent the doomsday device. That is definitely something which will destroy the world, while the spaghetti monster is almost certainly not real. However, using the qualitative approach as a decision making heuristic, you are equally justified in trying to prevent either the spaghetti monster or the doomsday device. This is because any positive probability multiplied by the infinite disutility will produce an infinite expected disutility. Therefore the expected disutility of both spaghetti monster risk and doomsday device risk is infinite, so if one was to rank them they would have to be equally ranked. The quantitative approach suffers from a "problem of paralysis"[5] across time.

## III. Can 'Anti-Knowledge' and the Quantitative Approach Be Reconciled?

Some might still defend a form of the quantitative approach however as a ranking heuristic, and adapt it in order to do so. An initial approach to doing this might be to consider only the probabilities of risks instead of both the probability and the consequences of each risk. This does help one rationalise the above intuition, that the decision procedure should lead one to choose to prevent the doomsday device rather than the spaghetti monster. It does alleviate the anti-knowledge problem

[4] Taleb, Nicholas Nassim, *The Black Swan*, Penguin, 2008

[5] Bostrom, Infinite Ethics, *Analysis and Metaphysics*, Vol. 10, 2010

by removing what makes it a problem. What makes it a problem for a quantitative approach is the fact that when multiplied by an infinite disutility, anything with any positive probability will be ranked the same. Therefore removing the infinite disutility from the calculation seems to remove the problem as well. However it itself brings an even greater problem with cross comparison between catastrophic and existential risks. For agents to practically make decisions between alternatives it must be possible to evaluate alternatives between existential risk prevention and prevention of other types of risk as these are common decisions to be made in risk prevention. A decision procedure cannot be a useful one to have if it cannot do this, as in these cases it would not enable an agent to make these choices. Imagine you are confronted with two risks, and have to make a decision about which to try and stop. One has a probability of 0.99 and will wipe out all of humanity, while the other has a probability of 1 but will kill much fewer people, say a few thousand. Intuitively one would say that one should decide to stop the first risk. However the probabilities approach would advocate trying to stop the second instead, as the probability of it occurring is greater than the first. This issue arises because of the fact that the outcome is an important consideration when making decisions about types of risk, and this ignores it. The other potential qualification one might try to make is ranking based upon only outcome therefore, but this again falls foul of the anti-knowledge problem. One in this case would again have no way of discriminating between the spaghetti monster and the doomsday device, as both would cause an infinite disutility and that is all a purely outcomes based quantitative approach would consider. Therefore it appears as if the quantitative approach cannot be rescued, and another approach is needed when deciding how actors should make decisions around existential risk.

I would argue that instead a more qualitative approach is better when considering existential risks. Though I do not sketch out a full system, I show how this might be carried out when considering human developed existential risk. Broadly speaking, it is possible for us to delineate two different types of what I shall refer to as 'negative existential possibilities'. I use this terminology instead of the term 'existential risk' because these are scenarios which might cause such a risk instead of necessarily risks. There are those which occur if our current situation continues on into the future and those that might occur in the future due to some change in conditions by the actions of some agent. The first shall be referred to 'type o' and the second as 'type p'. In simplistic terms, an example of the second type might be pursuing AI research or exploiting some new fuel resource which might exacerbate climate change, while an example of the first might be a continuation of our current energy policies causing catastrophic climate change.

## IV. A Potential Alternative Approach

To create a decision making heuristic for dealing with these possibilities, it seems useful to look at the possible trade-offs involved in a decision concerning them. To do this, and to analyse these trade-offs, one can adapt Parfit's argument in *Reasons and Persons*[6], and as a personal preference[7] continue to use the metric of utility as a normative criteria. Imagine four possible states of the world, which I shall refer to as A, A', B and C. B is the state of the world at our current temporal locus. A is a state of the world at some future point which is significantly more advanced than B due to some agent-made decision. A' by contrast is the state of the world where the agent had not made that decision at that same point in the future, so we can assume more advanced than B but not as advanced as A. C is a situation where the whole of humanity has been wiped out, including all

[6] Parfit, Derek, *Reasons and Persons*, Oxford University Press, 1984

[7] Though it seems as if one could use any consequentialist criterion, utility simply is easiest as it appears to be the most prevalent in the field.

possible future generations. In terms of utility, we can assume that A > A' > B > C due to this advancement (though this is for simplicity, the following would also hold under the situation A > B > A' > C). As Parfit argues, the ethical difference between a situation in which all of humanity is killed and a small proportion survives is much greater than the difference between a situation in which all survive and a tiny proportion survive, due to the fact that a situation where all of humanity is destroyed removes the potential for all future generations whereas a situation where part of it is destroyed does not. Analogously, we can maintain that the negative difference in utility between B and C is greater than the positive difference between B and A, and that the difference between B and A' is a less than the positive difference between B and A. The decision that an agent has to make when making a decision whether or not to pursue an action which causes a 'type p' negative existential possibility is one of comparing trade -offs based around these possible worlds, and hence one can compare these moves between worlds as a decision procedure which does not suffer from the problems of the quantitative approach. This can also be applied to multi-agent interactions, which is an important strength as agents such as nation states do not make decisions in isolation. In these cases, one can rank these qualitative moves, and given that the move of one agent will affect the other find dominant strategies and Nash Equilibria.

This leads to some interesting implications about classification of existential risk. If the agent chooses to pursue the possibility, then the agent is trading off a chance of a move from B to A with a chance instead of a move from B to C. They might cause an increase in utility, but it is also likely that they might cause a decrease in utility which is many times more significant instead. Indeed it seems as if the agent likely does not know the true chance of such a decrease due to the problem of anti-knowledge as discussed earlier. If the agent does not choose to pursue the possibility, then the agent instead makes a move from B to A'. They will bring about a lesser increase in utility, but there is no potential that through their action there will be a significant decrease in utility. Rationally, it seems as if therefore the agent should not pursue the action which brings about the 'type p' possibility. This is because the move from B to C is so much more significant than the move from B to A due to the fact that it removes the potential for all future generations, and so the trade-off is a bad one to make compared to simply the move from B to A'. If it is the case that the agent should not make the decision to pursue the action which brings about the 'type p' possibility, then it seems as if the results of 'type p' possibilities should not be considered existential risks of the same significance as the results of 'type o' possibilities. This is because the 'existential risks' resulting from 'type p' possibilities should simply not occur and so not be a risk, whereas the existential risks resulting from 'type o' possibilities likely will occur and so be risks. For example, imagine that one is approaching this decision from the perspective of a nation-state, which can be generalised as the highest level of agent in practical terms. One has two choices, to allow the pursuit of a 'type p' possibility or to ban it. Allowing it would give the potential of a move from B to A and consequently the potential instead of a move from B to C, while banning it gives one a move from B to A'. Therefore, following the reasoning above, the state should simply ban the negative existential possibility and so it is not an existential risk. This is also the dominant-strategy Nash Equilibrium when considering multi-agent interaction, which suggests that such a ban should be in affect worldwide[8].

A defender of the quantitative approach might argue that this decision procedure is in fact flawed because it does not take into account the probabilities of the potential moves. Imagine a case where we believe that we know with full certainty that there is a 99.999% chance that pursuing a 'type p' existential possibility will lead to a move from B to A, and a correspondingly small chance of 0.001%

[8] I have not included the decision matrix here as in its current form it is quite messy and indecipherable, though it is available on request

that it will lead to a move from B to C. In this case it might instead initially appear rational to pursue the move from B to A, if the utility differential between A and A' seemed large enough. However, this ignores the issue of anti-knowledge touched upon earlier. For one to know that the probability of a move from B to C is small enough, one must have certainty about one's predictive ability. To have such certainty, one must know that there will be no future advancement which will throw that predictive ability into question, i.e one must know what one does not know. One cannot know all of what one does not know, and so one cannot have reasonable certainty that the move from B to C is small enough to warrant the potential of it occurring.

This decision procedure does in fact have to be qualified however. There are some 'type p' possibilities where the agent's decision to pursue the possibility leads in turn to the existential risk of such possibility being reduced. This is because there are some cases where possession of whatever it might be which causes the risk in turn makes the risk from it to be lower than it previously might have been, from what we can reasonably judge. An example of this might be if one were to create a super-pathogen in a laboratory that one might expect to soon evolve, in order to create a vaccine against it. The fact that the super-pathogen now exists is a serious existential risk; if the pathogen was released into the environment through some accident then it might kill all of humanity. However, the fact that it has been created and a vaccine has been made off it means that we can now consider the existential risk from it to be much lower than before. For if it were released into the environment or develop in any other way, one would now be able to vaccinate against it. Therefore the attempted move from B to A might indeed be a rational one in this case for the agent to pursue, as by doing so the agent prevents any potential move from B to C occurring, not just by their own actions but by any equivalent actions of other agents in the future. Hence this caveat must be added when comparing the possible worlds, and comparing multi-agent interactions. This means that when weighing up possible worlds, one must take into account whether the negative existential possibility will reduce the overall risk of a move to C from what one can ascertain. If it can, then the decision procedure suggests that one should pursue it, while if it does not then it does not suggest this. In the case of multi-agent interactions, this in fact leads to a dominant-strategy Nash Equilibrium of both agents deciding to pursue the possibility.

The defender of the qualitative approach might question this by applying the previously stated criticism of using probabilities. If probabilities are irrelevant as argued above, then one cannot maintain this qualification. By adding it they might argue that one is simply arguing that it is rational to pursue the negative existential possibility when the probability of the existential risk occurring is low. This is explicitly what is argued against above. However, this is a misinterpretation as there is a significant and heuristically relevant distinction to be made. In cases like the super-pathogen, there is an already present potential of a move from B to C. Due to the casual link, these are cases where one can be reasonably certain (regardless of anti-knowledge) that creating this risk of a move from B to C by making this decision will in fact reduce the overall risk of a move from B to C. Therefore, the fact that there is a knowable reduction (even if it cannot be quantified) in the chance of the risk in these cases is what is significant, not the level of the risk.

To conclude, I have shown that the quantitative approach to decision making around existential risk, that is calculating expected utilities and using those values to make decisions between multiple possibilities, is flawed. This is because pursuing it leads to absurd conclusions when anti-knowledge and the potential for infinite future generations are taken into account. As the potential qualifications of this also lead to absurdities, I advocate instead a qualitative approach using possible worlds. The most interesting implication of this approach, which survives potential criticisms, is that certain negative existential possibilities such as AI risk should not be considered on the same level as

risks caused by a continuation of current action such as catastrophic climate change. This is because it appears as if in all but a minority of cases where they might be helpful, such risks should simply be not allowed to occur rather than time and resources used to regulate them.

Several critiques and objections have been raised to this conclusion in conversation. I have tried to lay them out and then deal with them as best as I can below.

## V. Some Objections - A Note On Utilities and Infinity

Elizabeth Baldwin raised an objection to this conclusion from the on the basis of utility, from the perspective of an economist. I shall lay out this objection, and ways in which this can in fact be overcome. These shall be in order of strength: firstly I shall argue that the conclusions can in fact be accommodated with what Broome calls "axiomatic utility theory"[9], and then secondly I shall show why that is not the idea of utility that is appropriate for answering questions in the ethics of existential risk.

The objection, broadly put, was that the conclusion does not hold as one cannot have a utility function (of the form $\sum^{\infty}_{r=1} \beta \, . \, U(r)$ where $\beta < 1$) which tends to infinity. If this is the case, then it does not seem as if one can have an infinite disutility from an existential risk. This is because if it in fact tends to a single number, the disutility from a risk which removes all potential future happiness and so on would be a finite number (the number which it would tend to). This would on the face of it seem to be a strong objection. If the disutility of an event tends to a fixed number rather than infinity, then an existential risk cannot cause infinite disutility. If it cannot cause infinite disutility, then it does not seem as if one would be forced to assign high expected disutility to absurd risks such as spaghetti monster risk. This is because the probability would be low enough to offset the large disutility value in the calculation. If one isn't forced to assign such risks a high expected disutility value, then one isn't forced to give them a high credence. Hence one does not seem forced to give such risks credence if the objection holds. Therefore the conclusion that expected utility is a flawed decision procedure for dealing with existential risks does not seem to be valid; one can use it to discriminate between risks still because the absurd risks will be given a much lower credence by the calculation than the probable ones. This means that it fits with our intuitions that one should normatively assign them a much higher credence in terms of resource distribution and so on.

The use of a utility function however implies the use of what shall be referred to as the 'economic conception of utility', what Broome calls "axiomatic utility theory", under which infinite utility can still be accommodated. Imagine the following case, adapted from the case originally posited by Nozick[10].

*There comes into existence a person whose brain is cognitively wired in a special way; they have insatiable preferences. For every preference of theirs which is satisfied in some time period, they develop two others which are caused by it being satisfied. Reading a book might lead to them developing an interest in the analysis of certain pages, the satisfaction of which leads them onto a set of divergent interests in certain schools of analysis and so on. Consequently, the satisfaction of a set of preferences of theirs in a time period leads to the development of more preferences to be satisfied. Imagine a case where these are all satisfied; in every time period they are able to satisfy every single preference that they develop. Also imagine that they develop equal satisfaction from the*

[9] Broome, John, *Utility*, *Economics and Philosophy* Vol. 7, 1991
[10] Nozick, Robert, *Anarchy, State and Utopia*, Basic Books, 2013

*fulfilment of each preference. It is not as if the preferences resulting from the satisfaction of the first are subsets of that preferences but must be considered fully fledged preferences in their own right. They therefore have an exponentially increasing satisfaction and preference set over time.*

An agent's utility on the 'economic conception' corresponds to preference satisfaction, and so utility in this case would be exponentially increasing over time. An appropriate utility function for this 'utility monster' would be:

$\sum^{\infty}_{t=1} \beta \, . \, U(t)$ where $\beta < 1$ and $U(t) = \beta \, . \, 2^t$

As $t \rightarrow \infty$ $\beta \, . \, U(t) \rightarrow \infty$

Therefore it is possible to have a utility function where there is an infinite total utility, when one considers economic utility accrued within an agent's life and over time. If this is the case, then it is at least possible that the total utility is infinite. Then therefore it is possible that even in the 'economic sense', an existential risk might cause infinite disutility.

Infinite utility can also be accommodated with a more conventional utility function, such as a one which is a form of a natural log, due to aggregation. Utility functions can only be seen to apply to an individual rather than a society. This is because they reflect the preferences and tastes of each individual within society, rather than society as a whole. With a more conventional utility function, it will be the case that for each individual their utility tends to a fixed value. Call this δ. Imagine a case where across time the number of individuals tends to infinity. In this case, the aggregate total utility across time will be:

$\delta + \delta + \delta + \ldots..$

i.e $n\delta$ where $n \rightarrow \infty$, which also tends to infinity

The consequences of an existential risk which occurred would be to remove the potential for all future generations of humanity. This means that it would in this case remove a number of individuals which tend to infinity. Hence the disutility from its occurrence would tend to infinity too, and so the conclusion holds.

However it is better to instead reject the use of the 'economic conception' for decision making to do with existential risks, as it violates a commonly held premise of the study[11]. This is the premise P ex

P ex) An agent should be concerned about existential risks which might occur outside of the agent's own lifetime or the lifetime of anyone that the agent is connected with

P1) What one prefers is not what is necessarily 'good'.

P2) The 'economic conception' of utility is only about what one prefers, not what is 'good'[12].

P3) If we hold P ex), then necessarily what is 'good' for one is not simply what one prefers. One's own preference set in the sense of 'preference' implied by the 'economic conception' is unaffected whether or not the world ends after one's lifetime or the lifetime of anyone one is connected with.

[11] I have presented the argument for this in the premise-conclusion form for ease of reading

[12] One's utility might be enhanced for example by inflicting serious pain upon oneself, if that is what one prefers. One does not have to necessarily have a preference for something in the strong or weak sense because it generates more pleasure.

Therefore - One cannot hold both an 'economic conception' of utility and that existential risk outside of one's lifetime is an issue that one should care about.

One would like to maintain that one should care about existential risks outside of one's lifetime if one is studying existential risks. Indeed it seems a necessary premise for someone studying existential risk to hold, as most hypothetical existential risks that the study is concerned with (like the risk from climate change or AI) are ones which would not necessarily occur within the lifetime of someone currently studying existential risk. Hence, it seems as if one must reject the 'economic conception' of utility as a relevant conception for ethical decision making in the study of existential risk.

It was suggested to me by Matthew Rendall[13] that an existential risk might be able to cause an infinite disutility using this conception of utility instead due to its effect within a single time period, though I feel as if this should be rejected for practical reasons. This could be argued due to the potential for there to be an infinity of possible alien civilisations. If this is the case then by an analogous argument to that which applies over time an existential risk which destroyed all of these civilisations would have an infinite disutility. Therefore if it were considered in this way, such a risk could have an infinite disutility even if utility did not tend to infinity over time. This is because such an effect would occur within a single time period, and would affect an infinity of 'people'. Therefore regardless of how a utility function tended over time, such an existential risk would have an infinite disutility. The issue with using this as an argument for the infinite disutility of an existential risk is that it leads one to exclude a lot of risks one would like to describe as existential risks. If the justification for a risk being one which would cause disutility which tended to infinity was that it would wipe out all possible civilisations, then it would have to be a risk which affected all such civilisations. Hence it would have to be a risk the effects of which extended infinitely across space in order to do this. The majority of risks we consider to be existential risks do not extend infinitely across space, but instead only affect life on the earth. Man-made climate change, the dangers from artificial intelligence and so on would all only affect life on earth. Therefore if one were to consider existential risks to have infinite disutility because of their effects across space rather than their effects across time, then one could not consider these to be existential risks of this type. As an agent, the risks that a decision procedure would be required to make decisions about which would be existential for humanity will often not be those which would have this effect across space. Therefore a decision procedure which considered risks to be existential due to their effects across space in this way could not be a practical one, it could not be applied to these prominent risks. As assumed, to be a good decision procedure it must be practical; the study of existential risks is a practical one. Therefore as this makes the study impractical, it is in turn not a good justification to use.

## VI. A Note on Negative Existential Possibilities and Consequentialism

A criticism was given by Hilary Greaves to the distinction between what I have called type p and type o negative existential possibilities. In this paper, I seem to both hold consequentialism and a distinction between these types of negative existential possibilities. The criticism is that these are incompatible; one cannot hold both consequentialism and such a distinction as one contradicts the other. This can be interpreted in two ways, which shall be dealt with in turn. The first is that there cannot be a morally relevant difference between the two types if one is a consequentialist. For a consequentialist, all that can be relevant for the goodness or badness of an act is the outcome.

[13] Based off of Bostrom, *Infinite Ethics*, *Analysis and Metaphysics*, Vol. 10, 2010

Assuming away any differences in outcome specific to one type of existential risk over another (the entire of humanity being tortured to death would for a utilitarian be worse than them dying painlessly for example) the outcomes of fulfilled type p and o negative existential possibilities would be the same, the removal of all current generations and the potential for future ones. If there is no difference in outcome, for a consequentialist there can be no morally relevant difference. Therefore it seems as if there cannot be a morally relevant difference between the types as a whole for a consequentialist. The normative framework of comparing possible worlds presented appears to depend upon both there being a morally relevant difference between the types and consequentialism. Therefore it might be flawed due to being based upon an inconsistency, at least at first glance.

However in fact it does not require the distinction to be a morally relevant one, and so it is compatible with consequentialism. The following example illustrates this.

*Jenny has a decision to make about whether to open a grate on the side of her house or not. If it is sunny, she will open it by going outside of the house and pulling it open. If it is cold though, she prefers to do it by pushing it open from the inside. The way in which she does it produces no further consequences itself, each are consequence neutral in their relative situations. The outcome is also the same in either situation.*

One can hold that there is a distinction between the cases, it is clear that there must be. However they produce the same overall consequences, there is no morally relevant distinction. This shows that one can hold two cases to be distinct even if the outcome is the same while maintaining consequentialism. One does not have to see the difference between them to be morally relevant to hold that there is a difference. This is the case with the distinction between type o and p negative existential possibilities; the distinction that is required is merely 'heuristically relevant' rather than morally relevant. What is meant by this is that it is a useful distinction to hold for making moral decisions rather than being relevant for the morality of each decision itself. In both cases, one ought to try to remove the potential for a negative existential event to occur. However in one case one can do this by preventing there from ever being a possibility of an existential risk, while in the other one can do this by preventing the possibility from occurring. The outcomes of these is equivalent, but that does not stop the distinction from being useful for working out how one should act in these situations.

The second interpretation of this criticism would be the fact that the distinction appears to refer to a single agent making a moral decision means that it is incompatible with consequentialism. However it does not matter who that agent is, no intrinsic quality of the agent is relevant to the act that should be carried out in this case. The agent could be myself, or the Pope, or Theresa May; the agent would still be faced with the same moral decision regardless. Hence this cannot make it incompatible with consequentialism. If it were then a consequentialist could never argue that any agent should make any moral decision in any situation. A utilitarian could not say that one should kill one to save five in the famous 'trolley problem' for example. This is clearly not the case.

## VII. A Defence of Anti-Knowledge

Rupert Read instead presented a criticism of the use of anti-knowledge as a concept that one should take seriously. The argument was presented as follows (though it is presented here in an abbreviated and simplified form). If one is to take the idea of anti-knowledge seriously, then one would have to accept that the letter E might be a source of existential risk. This is because one is

unable to rule out that it is not, there might be some unknown unknown that leads it to be so. It seems absurd to assign the letter E a non-zero probability as an existential risk. Therefore it is implied by this that one might not want to take anti-knowledge to its logical conclusion as the argument requires that one should for flying spaghetti monsters and so on to be assigned a non-zero probability of being an existential risk.

However the initial formulisation of this argument suffers a flaw which renders it much weaker than it could be, as Rupert himself agreed. Assigning the letter E a non-zero probability of being an existential risk implies that there must be a probability of it being non-conceptual. This is because concepts are defined a priori as things which cannot interact with the physical world, and hence cannot be an existential risk. There can be no room for any anti-knowledge suggesting otherwise in such cases as they are true by definition; there cannot be any unknown unknowns concerning it. For example, there cannot be any non-zero probability of a bachelor not being an unmarried man as what we refer to when we are using the word 'bachelor' is an unmarried man. However the letter E is by definition a concept, when we refer to 'E' rather than a specific case of its occurrence then we seem to be referring to the concept of E. Therefore it is necessarily true that it cannot be an existential risk, as it is necessarily true that concepts cannot be.

One can strengthen the argument as follows though. If one is to take the idea of anti-knowledge seriously, then one would have to accept that witches might exist. This is because one is unable to rule out that they do not, there might be some unknown unknown that leads it to be so. It seems absurd to assign a non-zero probability to the fact that witches exist. Therefore it is implied by this that one might not want to take anti-knowledge to its logical conclusion, as to do so one must accept the premise that there is a non-zero probability that witches exist. However when presented in this form it can be seen that it is not in fact a problem. It seems reasonable in fact to assign witches a non-zero probability of existing. One can hardly definitively prove that they do not. What seems unreasonable is assigning them a high probability of existing, which is not implied by taking anti-knowledge seriously.

## Appendix A – Discounting in Ethical Decision Procedures

In discussion of this paper, the issue of discounting over time was frequently raised. Consequently I have included the following argument as an appendix despite its tangential relevance to the overall argument of the paper itself.

Discounting over time is commonly used in economics when considering utilities. This is the idea that future utility is weighted relative to current utility, generally as being of lower value. This is usually done through considering some mathematical values of said utilities, and then using a discounting constant. Due to well documented effects such as instant gratification, this seems applicable when considering preferences within an individual's life. However though it might be applicable for these types of decisions, it is not good to use in ethical decision procedures over time. The following case illustrates this well.

*Beth is presented with a button, which she has a choice whether to push or not. If she pushes the button, then it will cause the painful deaths of a million people with lives worth living in a billion years' time. It will also however give her a mild pleasurable sensation now.*

If discounting over time is used for assessing the consequences of the action, and both outcomes (the lives of the billion people, and the mild pleasurable sensation) are assigned positive utility

values, it implies there exists some discounting constant β such that Beth would be ethically indifferent between pushing the button or not. Consequently, it is possible to imagine some discounting constant α such that $\alpha < \beta$ and therefore the act that Beth should pursue would be to push the button. However it is obviously not the case that pushing the button produces the best overall consequences; the disutility caused by the painful deaths of a million people obviously outweigh the mild increase in utility caused by pushing the button. It does not therefore seem as if a consequentialist decision procedure that is not purely egoistic should advocate pushing the button; it does not produce the best consequences in this individual case and it is hard to imagine that if pursued over an agent's lifetime it would produce the best consequences overall. This gets to the heart of the issue, for discounting to be applicable it implies a totally egoistic consequentialist theory. It might matter less for Beth if a million people die in a billion years' time; they may not care at all what happens to future generations. When considering overall consequences however, it seems counterintuitive to argue that the current consequences should be more valuable than future ones simply in virtue of their temporal locus. There is no reason why they should be if one is trying to consider what is best overall. Therefore, it seems that discounting across time is incompatible with a consequentialist ethical decision procedure that is not simply agent-centred. A theory which is simply agent centred would lead us to reject Px); there is no reason why existential risks which occur outside of the lifetime of any of the people that one cares about should matter when making decisions if the decision procedure used is simply agent centred. It would have no effect upon oneself or anyone one cared about. As argued earlier, it seems as if one does not wish to reject Px) if one is concerned with existential risks. Therefore one must reject discounting, and its implied egoistic decision procedure, instead.

Bibliography

Askell, Aamanda, *Look, Leap or Retreat?*, EAGx Oxford, 2016

Armstrong, Stuart, *Extreme Risks and Extreme Opportunities*, Tedx Athens, 2015

Avin, Shahaar, *Classification on Catastrophic Risks*, EAGx Oxford, 2016

Bostrom, Nick, *A Philosophical Quest For Our Biggest Problems*, Ted, 2005

Bostrom, Nick, *Existential Risks*, Journal of Evolution and Technology, 2002

Bostrom, Nick, *Existential Risk Prevention As Global Priority*, *Global Policy* Vol. 4 Issue 1, 2013

Bostrom, Nick, *Infinite Ethics*, *Analysis and Metaphysics*, Vol. 10, 2010

Broome, John, *Utility*, *Economics and Philosophy* Vol. 7, 1991

Knight, Frank, *Risk, Uncertainty and Profit*, Hart, Schaffener and Marx; Houghton Mifflin Company, 1921

Kripke, Saul, *Naming and Necessity*, Harvard University Press, 1980

Macaskill, William, *Doing Good Better*, Avery, 2015

Nassim Nicholas Taleb, *The Black Swan*, Penguin, 2008

Nozick, Robert, *Anarchy, State and Utopia*, Basic Books, 2013

Parfit, Derek, *Reasons and Persons*, Oxford University Press, 1984