Thinking with Whitehead about Existential Risk

James Burton, Goldsmiths, University of London

1. Thinking Existential Risk

Nick Bostrom defines an existential risk as a threatened destructive event that would be global in scope and terminal in intensity, such that it "would either annihilate Earth-originating life or permanently and drastically curtail its potential" (2002: 1.2).[1] He and other researchers associated with the Future of Humanity Institute (founded in 2005) argue that the topic has received scandalously little attention. They attribute this neglect to a number of factors, including the multidisciplinary nature of the problem (Bostrom 2013: 26), observation selection effects and other forms of cognitive bias such as "scope neglect" (Ćirković 2008; Yudkowksy 2008), the relative newness of many of the types of existential risk they identify, and, more generally, "an aversion against thinking seriously about a depressing topic" (Bostrom 2002: 2). Perhaps the strongest overall factor – and the greatest fundamental challenge for thinking about and addressing the risk of existential catastrophe – lies in the fact that, by default, we have never experienced or witnessed one. Thus, Bostrom emphasizes, given that "there is no opportunity to learn from failure," the "reactive approach" must be abandoned in favour of a "proactive approach" when dealing with the threat of existential catastrophe. (2013: 27)

In this paper, I take as read that existential risk, as well global catastrophic risks generally, demand our serious attention, and that they do indeed pose special or unique difficulties to attempts to think and address them. However, I also propose that the question of *how* we think about existential risks, in light of these special difficulties, is of fundamental importance, and that

[1] References to this paper, originally published in the *Journal of Evolution and Technology*, use its own internal numeration of sections in reference to the pdf and html formats in which it circulates online. The scheme was subsequently refined to indicate that an existential risk would have to be not only global, but trans- and pan-generational, i.e. not only decimating humanity at the time of its occurrence, but destroying or terminally impairing all future human generations (see Bostrom and Ćircović 2008: 3 and Bostrom 2013: 17).

this question is not fully answered by elaborations of the various cognitive biases that have had and could always have an effect on such thinking, despite the importance of reminders such as those provided by Yudkowsky (2008).

Another way of putting Bostrom's statement that there is no scope for a reactive approach in relation to existential risks is to say that there is no opportunity here for adaptive change (which in some contexts would be termed "evolution") to take place. The difficulty we face is, in the conceptual vocabulary of Gregory Bateson, the impossibility not only of first-order learning by trial and error, but also of second-order learning, or "deutero-learning" (1972: 166-169). Because "error is always biologically and/or psychically expensive," organisms reduce the amount of trial-and-error learning necessary by "learning to learn" more efficiently: "we (and all other biological systems) not only solve particular problems but also form *habits* which we apply to the solution of *classes* of problems," allowing them to be "solved in terms of assumptions or premises, fewer in number than the members of the class of problems" (Bateson 1972: 274). The notion that existential risk demands a "proactive approach" could be re-stated in these terms: existential risk presents us with the challenge of learning how to acquire the adaptive effects of deutero-learning without the benefit of the first-order learning through which we have become accustomed to developing them – effectively, a problem of third-order learning.

Meeting such a challenge can be expected to entail the kinds of capacities, acquired through millennia of deutero-learning, that are often discussed as autonomous properties of human thought or mind (abstract reasoning, reflective intelligence, deductive and inductive logic, and so on). Yet it should also entail a certain wariness of the habits such properties entail, as well as the dangers of leaving out other facets of mind acquired by feedback loops such as intuition, feeling, and unconscious modes of engagement with the world. As Bateson puts it, part of the efficacy of those hard-programmed analytic and cognitive habits often taken to be essential to learning depends upon a kind of meta-habit of not examining them (1972: 274). The FHI approach to existential risk seeks to

jolt us out of one meta-habit of not considering the (human-caused) end of humanity as a problem that needs addressing – in the process asking us to re-examine other psychological biases affecting the way we think (or do not think) about this issue. Here I want to ask whether this approach, heuristically valuable though it may be, introduces its own set of potentially restrictive biases, to the neglect of other modes of thinking that might be valuable to the challenge of addressing existential risk. Primarily, I want to explore what gains, if any, there may be from considering existential risk through the lens of a process-based metaphysics such as Alfred North Whitehead's.

This is an experiment, and as such is not undertaken with any sure expectation of success. At the same time, it is, of course, not an exercise undertaken arbitrarily. There are a number of reasons, intuitively at least, for thinking that Whitehead – who describes the lectures which compose *Process and Reality* as an attempt not only to provide a coherent metaphysical system, but also to repudiate certain prevalent philosophical habits of thought – might be helpful in this context.[2]

To begin with, at the broadest level, we might anticipate that a process-based framework would be particularly well-suited to any attempt to establish the criteria for the occurrence of events of a certain class and their complex relationships to prior events. My hope is that Whitehead's conceptual scheme and vocabulary may help address some concerns I have about the downplaying of the processual dimension of risk and catastrophe effected by the categorial and probability-based schema of Bostrom. At the same time, Whitehead's enterprise is ultimately underpinned by what he terms a rationalist "adventure of hope" – the faith that there are no elements in experience that are not "intrinsically capable of exhibition as examples of general theory" (PR: 42) – placing it, at least broadly, within the same rational spirit that informs current and recent thinking on existential risk.[3] On this basis, we may have at least initial cause to hope that such a speculative philosophy would be

[2] Among the "nine myths and fallacious procedures" in Whitehead's list, probably most relevant here are the habitual "distrust of speculative philosophy," the "trust in language as an adequate expression of propositions," and the "belief that logical inconsistencies can indicate anything else than some antecedent errors" (PR: xiii). References to *Process and Reality* are indicated throughout by the initials PR followed by a page number.

[3] Cf. Whitehead: "Rationalism is an adventure in the clarification of thought, progressive and never final. But it is an adventure in which even partial success has importance" (PR: 9).

of value in attempting to explore an area in which uncertainty must be accepted and respected as fundamental – the consideration of necessarily future and unprecedented events – yet in which greater accuracy of understanding is nevertheless urgently desired and sought.

2. The Becoming of Existential Risks

Thinking about existential risks is necessarily a speculative undertaking. But there are many different ways of being speculative, and the very conditions of a problem that requires speculation seem to point to the value of trying different modes, while exercising our best intuitive and intellectual estimations as to which might bear fruit.

The particular speculative mode developed by Bostrom and taken up by a number of other thinkers of existential risk begins with the attempt to categorize types of existential risk (within a more general categorial scheme of types of risk). This then forms the basis for making judgments about the probability of the (primarily near-future) occurrence of these types of risk, and for thinking about ways of lowering these probabilities. This approach has clear heuristic value in establishing existential risks as not only demanding, but also amenable to calculation, analysis, and planning in ways that may conceivably translate into pragmatic policy-making and other forms of collective preventive action.

This may well be a viable direct route to locating existential risks within the spheres of research and policy we can reasonably deem most likely to have a chance of mitigating them. Yet it nevertheless encourages us to neglect certain aspects of possible existential catastrophes – primarily their processual character – in ways that I would suggest could be to the detriment of attempts to address them in the long run. Experimenting with restoring this processual dimension by describing existential catastrophe and existential risk in Whiteheadian terms is not a matter of "correcting" an oversight: there is nothing in Bostrom's schema that takes any aspect of reality to have a

fundamentally non-processual mode of existence. Rather, the worry is that this schema lends itself to ways of thinking that tend to bracket or neglect the processual, a reduction that is commonplace and efficacious in many contexts of everyday and scientific thinking, but which can on certain points have a nontrivial effect. Identifying and seeking to recover the nontrivial losses in this reduction is the first aim of the following redescription of existential catastrophe and risk in Whiteheadian terms.

Henri Bergson, another process metaphysician, often referred to this kind of reduction as a "spatialization" of what are more aptly considered temporal aspects of reality. Whenever we treat something of a temporal or processual character as though it were divisible into homogeneous units – whether implicitly in thought and language, or with the aid of diagrammatic forms that represent time as a spatial dimension – we spatialize it. Whitehead agrees with Bergson that spatialization is reductive, in that it promotes an analysis of the world "in terms of static categories," even as it is simultaneously "the shortest route to a clear-cut philosophy expressed in reasonably familiar language" (PR: 209). In many cases and contexts, this reduction can be considered irrelevant or trivial. But whether as a direct impetus or indirect influence, the spatializing habit can have a nontrivial effect on attempts to understand the fundamental nature of reality, or any aspect of reality in which this fundamental nature is at stake: "The simple notion of an enduring substance sustaining persistent qualities, either essentially or accidentally, expresses a useful abstract for many purposes of life. But whenever we try to use it as a fundamental statement of the nature of things, it proves itself mistaken" (PR: 79). Bostrom's categorial and typological framework provides a "useful abstract" in this sense, but should not be taken as a "fundamental statement of the nature" of the things in question.

Let us try the experiment of putting existential risk and catastrophe in Whiteheadian terms. At the heart of Whitehead's metaphysics is what he refers to as the "concrescence" of "actual entities". Anything conceivable as a singular unit or object, any Cartesian *res vera* (PR: xiii) can, in

Whiteheadian terms, be considered an actual entity or an actual occasion.[4] The reason Whitehead uses the terms "entity" and "occasion" interchangeably is that an actual entity "exists" only at the moment its concrescence is "satisfied" – that is, the moment it is realised by or in relation to some other entity as an entity or thing: it is already passing out of being at the moment of this satisfaction, which is only identifiable in relation to it. Given that actual entities are the basic units or things of Whitehead's universe, anything conceived as having some reality from a human perspective – anything we are likely to treat as a thing, entity or occasion in the more everyday senses of these terms – is likely to consist in Whiteheadian terms, as a collection or convergence of many different actual entities. Perhaps in implicit recognition of this fact, Whitehead reserves a separate terminology for such a collection, referring to it as a "nexus" or "society": "In our reference to the actual world, we rarely consider an individual actual entity. The objects of our thoughts are almost always societies, or looser groups of actual entities" (PR: 198). A molecule, for example, must be "some kind of nexus of actual occasions" (PR: 73). The same applies to a person. To the extent that we conceive of a person as having consistence over time (and, indeed, to the extent that they *do* have some such consistence or persistence, regardless of our perception or conception of them) as somehow being the same "person" at one moment after another, and having done so throughout their lifetime, we are considering a nexus. This kind of nexus – as e.g. a persistent object or person – is for Whitehead a "society," in that the actual entities – or subordinate nexuses – which constitute it are related by a "social order," e.g. by which one cell in a body is replaced by a cell of a corresponding type, or by which organs and mental activities continue to interact in such a way as to refer to or prehend a particular body or person that is, through this prehension, grasped as the same from one moment to the next, day by day and year by year.[5]

[4] Cf. Whitehead: "The actual entity never moves: it is where it is and what it is. In order to emphasize this characteristic by a phrase connecting the notion of 'actual entity' more closely with our ordinary habits of thought, I will also use the term 'actual occasion' in the place of the term 'actual entity'" (PR: 73). On the one exception to this equivalence, see note 10 below.

[5] Social order corresponds to "that complex character in virtue of which a man is considered to be the same enduring person from birth to death" (PR: 90).

Since both existential risk and existential catastrophe are complex phenomena (whether considered from a metaphysical or anthropic perspective), I will use these terms (nexus, society) to describe them here. However, the fundamental metaphysical relationship between concrescence and actual entity/occasion gives us the fundamental structure involved in a process-oriented approach: anything constituted, recognized or perceived as a particular "thing" is what it is only as the realization of a process, a becoming, and not as having some static or eternal existential status outside of this process.

For any putative existential catastrophe – say, the result of an impact event between our planet and another astronomical object – even if we limit our attention to those aspects of the impact with a direct destructive effect on human life, we treat it as a nexus of occurrences ordered by their common relation to the occasion of the impact. This type of nexus is what Whitehead terms an "event."[6] The continued existence of the impacting object itself, over time (as of the Earth, or any organism or object whatsoever) would be a nexus of the "social" type; and the nexus which included its position and trajectory relative to Earth prior to the collision would of course be one societal nexus particularly pertinent to any attempt to prevent the impact (though, likewise, one could consider the nexus of its salient effects, including powerful winds, shock waves, thermal intensification, tsunamis, earthquakes, all of which could be analysed further as societies with subordinate nexuses and societies, down to the molecular level, or equally to the biographical level of their effects on groups of human and non-human societal nexuses).

We could in theory extend our description to infinitesimal degrees in these terms. What I want to point to is that, for any putative existential catastrophe – for an astronomical impact event

[6] "I shall use the term 'event' in the more general sense of a nexus of actual occasions, inter-related in some determinate fashion in one extensive continuum. An actual occasion is the limiting type of an event with only one member" (PR: 73).

as much as for a terrorist attack using nanotechnology, or a nuclear holocaust – there is a societal nexus corresponding to the process of which it is the posited outcome. There are some species of existential catastrophe for which this would be largely irrelevant to us – e.g. the scenario described by Bostrom in which it turns out we are living in a simulation, and those running it decide to shut it down (Bostrom 2002: 4.3; 2003). But for most of the species of existential catastrophe in whose culminating process we might hope to intervene, it is the nexus or society of its coming-to-be that is likely to be most worth our attention.

This nexus or society can be understood as the material or actual set of processes – consisting in subordinate nexuses and others within those, to whatever degree of detail we find it helpful to try to identify them – corresponding to any particular estimation of an existential risk. That is, estimating the probability of an existential risk can be considered an estimation of the likelihood that one such nexus exists, is in process, towards its eventual satisfaction. Attending to this nexus, seeking to locate it, speculating as to and seeking out its components and their advance, would be the operation of an attempt in these terms to mitigate a given species of existential catastrophe.

Bostrom and others' identification of existential catastrophe and risks as particular events, amenable to categorization and probabilistic calculation, has heuristic value in calling for attention to, and beginning to search for advantageous ways to develop mitigating strategies against them. However, such an approach simultaneously (if inadvertently) encourages us to think about given existential catastrophes in binaristic or atomistic terms, as possibilities which will either come to be the case or not, in a manner which risks diverting our attention away from the processes by which this might occur, and which make their probabilities dynamic over time.

One retort to this might be that it should be perfectly reasonable to expect us to be able to employ probabilistic thinking and more concrete, process-sensitive analysis respectively in their proper contexts. That is, we should be able to apply the former in contexts where we are concerned

with the concept and likelihood of existential risk(s), and the latter in contexts where we are more concerned with intervening in factors that seem to be converging to increase the probability of some specific existential catastrophe occurring. I would suggest, however, that this would be an extremely difficult distinction to maintain in any sustained way. Not all of the feedback loops that go into the development of our thinking and reasoning (indeed, not even the majority of them) are conscious – and neither, indeed, are the relevant factors affecting either our own or nonhuman decisions for the realization of some particular nexus such as a given existential catastrophe. It is neither mystical nor irrational to recognize that an invocation to base our evaluations on "reasons rather than untutored intuition" (Bostrom 2002: 4.3) *can* only ever be met in part, and *should* only be pursued as far as there is useful scope for acting on the basis of reason alone.

Let me offer an example of an area in which I think the probabilistic approach of the prevalent thinking on existential risk – that is, this tendency to encourage an approach which would lead us, consciously or unconsciously, into the habit of treating existential catastrophes as atomistic events isolated from the processes of their coming-to-be, rather than nexuses – can cause problems that might be addressed by the supplement of a more process-based approach.

A recurrent feature of discussions of existential risk to date is the drawing of relatively firm distinctions between terminal and non-terminal global catastrophic risks. This is a virtually inevitable result of Bostrom's proposal to categorize risks by type, according to discrete levels (rather than gradations) of intensity or scope. This might seem to be assuaged by the way existential risk is included within the larger category of global catastrophic risk, rather than set apart as a wholly independent category. In Bostrom's original schema (2002: 1.1), there are risks that are considered global but non-terminal (endurable), such as the thinning of the ozone layer, and terminal risks that are not considered global, such as genocide: a putative catastrophe would have to be both global and terminal to be placed in the existential risk category. In a later, revised scheme (Bostrom and Ćirković 2008: 3; 2013: 17), the global category has been further subdivided to include categories of

risk that are trans- and pan-generational, and the "terminal" category has been replaced with "crushing." On the one hand, this revision indicates a sensitivity to the range of possible catastrophic events that have a non-negligible chance of occurring, and the fact that a catastrophe can be devastating for large portions of humanity without qualifying as existential. On the other hand, however, it functions to reinforce the "special" status of existential risks and the sense that, however great another catastrophe might be, an existential catastrophe should always be of massively greater concern. This is also reflected in the edited collection of essays *Global Catastrophic Risks* (Bostrom and Ćirković 2008), where the editors recognize the sensitivity (and controversy) surrounding the question of how much worse an existential risk should be considered compared to a non-existential global catastrophic risk, broadening their scope in order to "lay a broader foundation of systematic thinking about big risks in general" (2008: 4). Nevertheless, the core of the approach remains that "existential risks share a number of features that mark them out as deserving of special consideration" (2008: 4).

An argument that has been used repeatedly to emphasize the greater importance of existential risks over others is Derek Parfit's reasoning that the difference between a nuclear war that destroyed 100% of humans, and one that destroyed 99%, would be far greater than the difference between the nuclear war that destroyed 99% and the avoidance of such a war altogether. (That is, the survival of a tiny number of humans with the potential to propagate the species into the future is infinitely preferable to the survival of none.) This is based on the reasoning that the eradication of 100% of humanity should be taken to include all possible future generations, whereas the eradication of 99% would not. This sets up an oppositional relationship between existential and non-existential risks: "One might consequently argue that even the tiniest reduction of existential risk has an expected value greater than that of the definite provision of any 'ordinary' good, such as the direct benefit of saving 1 billion lives" (Bostrom 2013: 19).

This kind of reasoning, often couched in terms of utilitarianism, also involves a degree of game theory, whereby the value assigned to a smaller quantity of lives, e.g. 1 billion, is much smaller than that assigned to a much larger quantity, e.g. 10^{16} (Bostrom 2013: 18), thus making sacrificing the former to preserve the latter ethically preferable to preserving the former at the (possible eventual) expense of the latter. However, while there may be no urgent need to refute this in theory,[7] its application in practice would depend on the emergence of conditions in which this binaristic choice could be actualized. Whether and how likely such conditions are to emerge seems to me much more open to question. Furthermore, the possible argument that even situations in which this *seemed* to be a binary choice would require us to act for the greater good, given the stakes, are suspect, given the extent to which actions taken in the name of the greater good have throughout human history led to the destruction of life on massive scales, and could thus at least as easily be expected to contribute to an existential catastrophe as mitigate one.

While situations can be conceived in which these conditions are met – such as Bostrom's imagining of a "rogue state" scenario in which a preemptive strike against a sovereign nation is necessary to prevent it causing an existential catastrophe (2002: 9.3) – it seems likely that for most varieties of existential risk, there is a good chance that many non-crushing global catastrophes could form part of the nexus constituting the coming-to-be of a particular existential catastrophe. A straightforward example would be global warming. Bostrom gives the thinning of the ozone layer as an instance of an endurable (i.e. non-existential) global catastrophic risk. But this delineation can only be made after the fact: had the "ozone hole" not been recognized and made the target of direct

[7] There are many grounds for at least questioning this reasoning, or at least the degree of difference derived from it between the destruction of human lives on massive contemporary versus pan-generational scales. For example, the inclusion of all potential future generations within the calculation of the number of lives destroyed seems to presume that no other factor intervenes in the near future to prevent the appearance of those 10^{16} lives. In another vein, to suggest that the destruction of a potential future life is equivalent in value to the destruction of an extant life is dangerously close to arguments that equate abortion with killing – and could even be taken to imply that all of us have an ethical duty to procreate as much as possible. However, I see no reason to disagree in absolute terms with the idea that a pan-generational catastrophe would be considerably worse than one affecting a more limited (though great) number of lives: the concern here is how one determines one or the other *in advance* to be an outcome of particular circumstances or actions, and what actions this is then used to support or justify.

global action in the 1980s, or had the Montreal Protocol not succeeded, the production of ozone-depleting chemicals might have been the primary factor in the occurrence of an existential catastrophe. While it is correct that a short-term thinning of the ozone layer can be considered endurable, while a longer-term erosion, within certain quantifiable parameters, would be crushing or terminal for humanity, placing them on opposing sides of a categorial line encourages us to think them in opposition, neglecting the way in which each is implicated within the other: an existential ozone-depletion catastrophe arises out of a non-existential one, and addressing the latter reduces the risk of the occurrence of the former.

For other species of existential risk, the possible relations are likely to be much more complicated, but still in ways that I think the categorial approach may encourage us to neglect. For example, the "rogue state" scenario, or one arising from the "deliberate misuse of nanotechnology" to "eat up the biosphere or destroy it by other means such as by poisoning it, burning it, or blocking out the sunlight" (Bostrom 2002: 4.1) could be considered a societal nexus partially constituted by numerous prior global catastrophes whose mitigation might have led to its avoidance. In this hypothetical scenario, it seems highly plausible that, as is the case with many historical acts of terrorism, those behind the deliberate misuse of nanotechnology might be responding to certain geopolitical conditions of injustice and inequality: these might include military invasions and conflicts, widespread poverty, the uneven global distribution of the effects of climate change – all enabled or permitted by other human institutions and powers with the capacity to intervene. Furthermore, it should be possible to identify a number of putative existential catastrophes which, though differing in kind, might include these diverse global catastrophes as significant elements within their concrescence. There are likely to be many non-terminal global catastrophic risks that would constitute elements in the nexus of a number of putative existential catastrophes in formation. It would, therefore, make sense to direct attention towards identifying and addressing these, and viewing this not as a subtraction of resources and attention from the putative actual

entities that constitute given putative existential catastrophes, but as part of the wider challenge of addressing existential risks.

This may seem, from the prevalent existential risk perspective, to advocate what Bostrom denigrates as frittering away altruism on "a plethora of feel-good projects of sub-optimal efficacy" (2013: 19). There are several reasons in addition to the above for suggesting this is not (or at least not necessarily) the case, and that there are additional benefits to the task of reducing at least some species of existential risk, direct and indirect, in tackling such global ethical challenges as world poverty, health, poor living conditions, environmental damage, social inequality, military conflict, and other non- or not-yet-existential threats.

For one thing, as Bostrom notes, addressing many species of existential risk is likely to require a lot of advocates and resources. He thus expresses the hope that "some of the global movements that emerged over the last half century – in particular the peace movement, the environmentalist movement, and various global justice and human-rights movements – will increasingly take on board more generalised concerns about existential risk" (2013: 27). Surely recruiting the voices, efforts, and resources of those committed to such movements is likely to be facilitated by including the challenges of those various forms of global catastrophe in which their primary interests reside, as part of, rather than in competition with, the challenge of addressing existential risk. The notion of asking activists and oppressed groups to set aside their primary concerns in favour of the "greater good" that is the survival of humanity as a whole seems to me not only politically or ethically unjustifiable, but also hopelessly impractical – especially given that, for many such groups it is precisely the struggle over who or what counts as "human", and who is viewed as representative of humanity that is at stake.[8]

Secondly, while some form of game theory might eventually come into play, depending on the probability and type of existential risk in question, there is and will continue to be a great degree

[8] See Sylvia Wynter (2003) for an extended account of the epistemology and politics of what she sees as the ongoing struggle in modernity between the ethnoclass "Man" and humanity as a species.

of ignorance shrouding all thinking about future risk, regardless of the partial achievements of efforts to pierce it. For this reason there must be some ethical value in the intuitive judgment that any kind of suffering on a mass scale is not only worth addressing in itself, but also as a possible precursor of a more totalizing existential catastrophe. Bateson developed his thinking on deutero-learning in complementary response to a paper by Margaret Mead, in which she advocated that researchers work "in terms of values which are limited to defining a *direction*" rather than towards "defined *ends*" (quoted in Bateson: 159). It seems to me that such a "directional" approach must at least be part of our thinking of and approach to existential risk, such that any kind of global catastrophe should at least be evaluated as a potential element in a putative future existential catastrophe. There would, of course, equally be much danger in making this a fixed law or presumption, and it could promote the very kinds of biased judgment that Bostrom and Yudkowsky (2008), for example, want us to avoid. But we may bear various forms of cognitive bias in mind and look for them in our thinking without abandoning intuition altogether.

Finally, and perhaps most simply, even if we identify it as a reasonable probability that a number of global catastrophic risks are *not* going to play a part in the later occurrence of an existential catastrophe, there is arguably even greater reason to see some of these as potential elements in the development of situations that would be *worse* than the actualization of an existential risk. Bostrom includes such scenarios in his categorial scheme under the designation "hellish." Examples of scenarios that could be considered worse than the eradication of humanity include "permanent and extreme forms of slavery or mind control" (Bostrom and Ćirković 2008: 4), "horrible incurable diseases" (Bostrom 2013: 28), and extreme, permanent totalitarian regimes. As Bryan Caplan notes, "it is tempting to minimize the harm of a social disaster like totalitarianism, because it would probably not lead to human extinction… But perhaps an eternity of totalitarianism would be worse than extinction" (2008: 517).

At the very least, all of this would seem to point to the value and efficacy of addressing global catastrophic risk holistically in such a way as to include its existential and non-existential varieties. Nothing in the categorial and probabilistic scheme of Bostrom and others directly opposes this, and indeed, there are moments at which it is advocated. This may be taken as manifest in the decision to publish a book on global catastrophic risk rather than existential risk, for example. However, to the extent that the latter ultimately comes across as a book on existential risk with some attention paid to the other sub-categories of global catastrophic risk, and to the extent that notions which seem to oppose this special category to others crop up repeatedly in discussions of the topic, the mode of thinking that underpins them seems to bring with it the kinds of risks I have pointed to above.

Whitehead's system and vocabulary are of course not the only way to get to this position. However, I would suggest that by encouraging us to think carefully about the relation between being and becoming – in putting us within a certain process-sensitive mindset – they begins to affect the way we think about a given phenomenon or subject, such as existential risk, in subtle but potentially important ways. Beyond this, the detail of the conceptual vocabulary with which Whitehead develops his metaphysics, and which I have barely touched on thus far, offers great scope for more careful description of particular putative or potential existential threats in processual terms, should the attempt be deemed worthwhile.

3. Technoscientific Bias and Non-scientific Resources: Propositions

Before concluding, I would like to consider from another angle, but one also partially informed by Whitehead, the special problems existential risk poses to being thought.

I noted in the introduction to this paper that Bostrom posits an "aversion" to thinking about such a depressing topic as one of the reasons so little attention has been directed to the possible

occurrence of existential catastrophes. A possibly related factor, highlighted by Yudkowsky (2008: 105-7), is "scope neglect." This is the phenomenon whereby people treat a small number of negative occurrences (e.g. deaths) as though they were worse than a much larger number. Such thinking (or feeling) was expressed by Kurt Tucholsky, among others, in attributing to an imagined French diplomat the following statement: "The death of one person: that's a catastrophe. A hundred thousand deaths: that's a statistic" [9] (1932: 148).

Among speculations as to the reasons for this bias, which has been documented in a number of psychological experiments, Yudkowsky cites a common saying in this field of study, that "people do not evaluate events, but descriptions of events" (2008: 114). This should prompt us to consider the implications of the particular conditions underpinning the Bostrom/FHI mode of describing existential risk in comparison to other such modes. The hypothesis of this paper thus far has been that describing large catastrophes in terms of their probability within a categorial scheme of risk has nontrivially different effects on the way we evaluate it compared to a more process-sensitive mode of description (many of these effects being heuristically valuable, but a few giving possible cause for concern). Other modes of describing occurrences that threaten the extinction of humanity, as found in the many narratives constructed on the theme in mythology, religion, literature and other media, will have different conditions and implications for the ways we evaluate them. Yudkowsky, in fact, highlights this sphere, though primarily, it seems, in order to associated it with flawed thinking, when he suggests that "the cliché phrase *end of the world* invokes the magisterium of myth and dream, of prophecy and apocalypse, of novels and movies" (2008: 114). Faced with prospects of destruction on scales that literally transcend their capacity for understanding and rationality – "the brain cannot multiply by 6 billion" (Yudkowsky 2008: 115) – humans turn to the sphere of the transcendent to look for ways to deal with them. This turn to the transcendent is generally seen as a hindrance or distraction within research emerging from the FHI. The prevalent existential risk

[9] There are many versions of this expression: a very similar statement is commonly (possibly apocryphally) attributed to Joseph Stalin, while Yudkowsky (2008: 106) cites Hungarian physiologist Albert Szent-Györgi's statement that "I am unable to multiply one man's suffering by a 100 million."

approach represented by Bostrom et al seeks to bring existential catastrophe back within the immanent realm of thinkability by developing means of rendering it calculable and amenable to analysis, and exposing the cognitive biases that form obstacles to this enterprise. However, we might also wonder whether wholly avoiding this recourse to the transcendent is either possible or desirable as part of the challenge of mitigating existential risk.

There is a clear techno-scientific bias in existential risk thinking to date, manifest first of all in the general position that the kinds of existential risks with which it is principally concerned are new to human history, dating roughly from the appearance of the possibility of global nuclear war in the mid-twentieth century (Bostrom 2002: 2). This effectively brackets out from the category of significant existential risks any perceived threats to the survival of humanity that have arisen in past religious and mythological contexts, such as large-scale floods, earthquakes, volcanoes and the divine agencies often taken to be behind them. But such a dismissal would seem to run counter to Bostrom's own recognition, in justifying the speculative dimension of his own approach, that: "*If we don't know whether something is objectively risky or not, then it is risky in the subjective sense.* The subjective sense is of course what we must base our decisions on" (2002: 2, emphasis in original). Nevertheless, the result is that past cultural responses to such risks are effectively ruled out as potentially useful resources for facing current and future challenges of existential risk, even if Bostrom recognizes that religious responses may not have been "unreasonable" within their historical cultural contexts (2013: 29, note 35). This bias is underscored in the list of possible "general improvements" Bostrom hopes may increasingly help mitigate existential risk: "developments in educational techniques and online collaboration tools, institutional innovations such as prediction markets, advances in science and philosophy, spread of rationality culture, and biological cognitive enhancement" (2013: 28). While it is reasonable to assume that techniques such as "sacrificial offerings, persecution of witches or infidels, and so forth" (2013: 29, note 35) are unlikely to be of much use in facing existential risk in the future, to regard superstition-based ritual

as the only potential resource to be found in such contexts (or to dismiss such contexts on the basis that they *include* such elements) seems needlessly restrictive.

How might we envisage an approach that would be neither exclusively rationalist in this way, nor limited to impractically superstitious responses to the transcendent dimension of existential risk? That is, how might we go about rejecting or overcoming the apparent dichotomy of immanence and transcendence that this opposition implies, and which is found widely in modern thought (often manifest in excessive rationalism or scientism on the one hand, and dogmatic or obscurantist mysticism on the other)?

Whitehead's conception of God may initially seem to offer one prospect of addressing this problem. Despite being undertaken, as noted above, as a rationalist adventure, seeking to provide a descriptive system adequate to both objective and subjective experience as part of a single extensive continuum (like Bergson, Whitehead rejects a dominant opposition in nineteenth century philosophy between realism and idealism, anticipating later philosophers of immanence such as Gilles Deleuze and François Laruelle), Whitehead's metaphysics nevertheless has room for a God who/that is essentially not religious. However, the primary functions of Whitehead's God seem to be in enabling the creative advance of all existence, and in preserving or "saving" all that, by virtue of its processual coming-to-be, must also pass out of existence.[10] For this reason, it is arguably of little value in any endeavour concerned with affecting or (re)directing the direction or historical unfolding of specific situations as they impinge upon the human. Thus for help in addressing the question of existential risk, we must turn elsewhere.

[10] These are the respective functions, in summary form, of the two dimensions of Whitehead's God: a "primordial nature" constituting "the unlimited conceptual realization of the absolute wealth of potentiality" (PR: 343); and a "consequent nature" corresponding to the "realization of the actual world" (PR: 345), by which God constitutes the unity of all actual entities as they pass, "sav[ing] the world as it passes into the immediacy of his own life." (PR: 346)

Of greater value to the challenge of thinking about existential risk, I think, is Whitehead's discussion of propositions. For Whitehead, a proposition "is the unity of certain actual entities in their potentiality for forming a nexus" (PR: 24). The notion of attempting to address existential risks by postulating or trying to identify the nexus-in-formation that may be leading in the direction of a given existential catastrophe, contributing to or constituting its coming-to-be, as described in the previous section, could be described as a propositional mode. Whitehead terms the constitutive actual entities of a proposition its "logical subjects", while definite eternal objects (for example, the principle that humanity can come to an end, the fact of the passing or perishing of all actual entities, the potential for this or that mode of destruction) are its predicates. In this sense, a proposition has actuality in the actual entities it involves, as well as in the actual entity in which its thought or expression consists, and yet can still be the basis of logical or theoretical speculation in the senses in which philosophers, scientists, and other thinkers more commonly use the term. As Whitehead puts it, "a proposition is a new kind of entity. It is a hybrid between pure potentialities and actualities" (PR: 185-6).

Whitehead is critical, however, of the logic-centred philosophical approach to propositions that has treated them purely in terms of true/false expressions, geared towards the making of judgments. "The main function of propositions in the nature of things" he writes, is not to facilitate belief, "but for feeling at the physical level of unconsciousness. They constitute a source for the origination of feeling which is not tied down to mere datum. A proposition is 'realized' by a member of its locus, when it is admitted into feeling" (PR: 186). In terms of this distinction, the prevalent mode of discussing existential risk can hitherto be said to have been propositional in the narrower sense – seeking to establish the basis for making yes/no or true/false judgments and logical, calculable estimations of probability. If, however, we appreciate the value of propositions in Whitehead's sense, then we may look for propositions relevant to the challenge of thinking and addressing existential risk in the kinds of places that the probabilistic existential risk approach tends to exclude as irrelevant and/or irrational: in mythology, religion, intuition, literary, and other media

and narrative modes – Yudkowsky's separate "magisterium of myth and dream, prophecy and apocalypse, novels and movies." After all, there is no reason for seeing such "unscientific" realms as incapable of constituting useful resources (intellectual, affective, or otherwise) for thinking and addressing existential risk – any more than we would expect to find scientific or analytic thought and discourse free of either cognitive bias or speculation.

The propositions found in such cultural resources concerned with threats to humanity, from *Atrahasis* and other ancient flood myths to J. G. Ballard's *The Drowned World*, from Plato's *Timaeus* to *The Planet of the Apes*, may well all turn out to be "non-conformal" to the actual world of an entity concerned with it, rather than "conformal" ("non-conformal" and "conformal" being Whitehead's adaptations of "false" and "true"). It is quite likely that none will conform fully to the nexus of a given existential catastrophe (though by the time we knew this it would be too late, from a human-survival-oriented point of view, for it to matter); there is, however, plenty of scope for thinking that the propositions found in such loci might conform to some or other element in the actuality of existential *risk*, as the thinking, feeling, acting in relation to the possibility of such catastrophes – and that they may therefore be of value to attempts to shape or affect these responses and approaches. Even so, in contrast to the way propositions are deployed in a standard mathematical or logical treatise (such as Russell and Whitehead's *Principia Mathematica*), even failing to conform to actuality would not render a proposition without value within this broader perspective:

> The conception of propositions as merely material for judgments is fatal to any understanding of their role in the universe. In that purely logical aspect, non-conformal propositions are merely wrong, and therefore worse than useless. But in their primary role,

> they pave the way along which the world advances into novelty. Error is the price which we pay for progress. (PR: 187)[11]

Within this paradigm, error is of (great) potential value. In the context of an actualized existential catastrophe, error in the approach to mitigating it is terminal. But errors regarding, for example, the course of its development, the question of "which one will get us first," or the different factors in its concrescence, may all contribute to the population of a conceptual and affective picture of existential risk that may have diverse roles to play in our multi-modal attempts to mitigate it.[12] Every proposition brings something new within our scope:

> When a non-conformal proposition is admitted into feeling, the reaction to the datum has resulted in the synthesis of fact with the alternative potentiality of the complex predicate. A novelty has emerged into creation. The novelty may promote or destroy order; it may be good or bad. But it is new, a new type of individual, and not merely a new intensity of individual feeling. (PR: 187)

[11] On this basis, we might also consider as a probably nonconformal proposition with potential value as "lure for feeling," a version of Whitehead's God that would still have some capacity for the salvation of actual humans; this was a possibility raised in relation to environmental catastrophe in the presentation at the conference in Claremont that became the germ of this paper (*How Do You Make Yourself a Proposition? A Whitehead Laboratory,* Dec 1-3, 2016).

[12] Equally, while propositions drawn from an ancient mythological text may have little to contribute to the challenge, for example, of developing "ecophagic devices" to counter a nanotechnological catastrophe (though who knows?), they might easily offer something of value to the challenge of designing "new institutions that can maintain and administer centralized global power without becoming oppressive"(and the possible blindspots in the thinking in such design that could unwittingly give rise to further threats) – two possible approaches to nanotechnology as a global catastrophic risk suggested by Phoenix and Treder (2008: 497-8). In a similar vein, while scientific and analytic research (including, at least in passing, some of the work on existential risk) often recognizes that science fiction provides useful imaginary descriptions of possible emergent or future technologies that could give rise to global catastrophic threats, such as robotics and AI (Isaac Asimov), the technological singularity (Vernor Vinge), or nanotechnology (Neal Stephenson), we should not neglect what these and other less scientifically detailed works of science fiction might offer to pragmatic efforts towards risk mitigation in the propositions they offer touching on ethics, politics, culture, modes of thinking and feeling, myth, and so on.

This may be one general approach to tackling the problem described at the outset of this paper, of achieving the adaptive results of Bateson's deutero-learning without the benefit of first-order trial-and-error learning. This is a form of learning based on error without trial – on virtual error, or error as the general field of hypothetical possibilities from which actuality will continually emerge. It might be considered a cousin, as it were, of the species of thought experiment on which analytic thinking (including that of existential risk) often draws – but expanded to include feeling, error, uncertainty as valuable aspects of both the resource in question and the effects derived from it.

On this basis, I would advocate mining the vast collection of cultural resources, both ancient and modern, relating to the theme of the end of the human, the form it takes and the ways humans and other beings respond, for propositions of potential value to the larger task of facing the challenge of existential risk, which is as much a psychological and cultural problem as it is a technoscientific one: in this sense, the conception of this project as multi- or transdisciplinary has not yet gone far enough. Even the most sceptic rationalist, one who deems such cultural resources as almost certainly irrelevant to this task, would accept, we might hope, that they have the potential capacity to help reduce existential risk by "*one billionth of one billionth of one percentage point*," which, according to Bostrom's calculation of the value of addressing existential risk at all, would be "worth a hundred billion times as much as a billion human lives" (2013: 19). Furthermore, though such an undertaking might require significant time and effort, nevertheless considered in relation to projects such as enhancing international counterterrorism initiatives, implementing comprehensive biosecurity strategies, developing global systems for overseeing nanotechnology research, or building "Noah's Ark" refuges[13] and seed banks, such research has the extra advantage of being, to put it simply, cheap.

We shouldn't expect to be able to anticipate exactly what benefits might be derived from such research (any more than one does with a given scientific experiment, so long as a working

[13] For a discussion, see Hanson (2008: 373-5).

paradigm and reasonable hypothesis that useful findings are possible has been established). But undertaking the endeavour would in itself imply a slight loosening of the techno-scientific/rationalist bias prevalent in existential risk thinking to date (this is not to say that this bias is not for the most part sensible and efficacious; it is in what it risks excluding, rather than what it includes and prioritizes, that I find some cause for concern). But we can speculate that the value of this loosening is one effect into which, through a series of feedback loops, we might expect to gain further insight and understanding as such an endeavour is pursued. In particular, we should at least entertain the possibility that the long-term survival of humans in some or other (likely posthuman) form will ultimately depend upon our ability to let go, at least to some extent, of our fixation on precisely this goal of human survival, or at least our treatment of it as an absolute imperative, and our seeming dependence on instrumental means of achieving it. Indeed, who is to say that this is not the "Great Filter" that has been proposed as bringing about the extinction of complex, intelligent lifeforms elsewhere in the universe, such that we have not yet encountered them?[14] Might it not be that the fixation of advanced technological societies or species on the scientific rationality and technological reasoning that they credit with having got them there, is precisely what repeatedly leads to their (self-)destruction through the (mis-)use of their technological accomplishments? This may very well *not* be the case; but it is a possibility that at least deserves to be included in our attempts to think about how to think about existential risk.

Works Cited

Gregory Bateson, *Steps to an Ecology of Mind* (Chicago: University of Chicago Press, 1972).

[14] See Bostrom (2002: 8.2); Ćirković (2008: 131-5); Webb (2002).

Nick Bostrom, “Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards,” *Journal of Evolution and Technology*, Vol. 9, No. 1 (2002). Reproduced and available online at: https://nickbostrom.com/existential/risks.pdf

Nick Bostrom, “Are You Living in a Computer Simulation?” *Philosophical Quarterly,* Vol. 53, No. 211 (2003): 243-255.

Nick Bostrom, “Existential Risk Prevention as Global Priority,” *Global Policy,* Vol. 4, Issue 1 (February 2013).

Nick Bostrom and Milan M. Ćirković, “Introduction,” in Bostrom and Ćirković (eds) *Global Catastrophic Risks* (Oxford: Oxford University Press, 2008), 1-29.

Bryan Caplan, “The totalitarian threat,” in Bostrom and Ćirković (eds) *Global Catastrophic Risks* (Oxford: Oxford University Press, 2008), 504-519.

Milan Ćirković, “Observation selection effects and global catastrophic risks,” in Bostrom and Ćirković (eds) *Global Catastrophic Risks* (Oxford: Oxford University Press, 2008), 120-145.

Robin Hanson, “Catastrophe, social collapse, and human extinction,” in Bostrom and Ćirković (eds) *Global Catastrophic Risks* (Oxford: Oxford University Press, 2008), 363-377.

Chris Phoenix and Mike Treder, “Nanotechnology as global catastrophic risk,” in Bostrom and Ćirković (eds) *Global Catastrophic Risks* (Oxford: Oxford University Press, 2008), 481-503

Kurt Tucholsky, *Lerne lachen ohne zu weinen* (Berlin: Ernst Rowohlt, 1932).

Stephen Webb, *If the Universe is Teeming with Aliens – Where is Everybody? Fifty Solutions to Fermi’s Paradox and the Problem of Extraterrestrial Life* (New York: Copernicus, 2002).

Alfred North Whitehead, *Process and Reality: An Essay in Cosmology*, ed. David Griffin and Donald Sherburne (New York: The Free Press, 1978).

Sylvia Wynter, “Unsettling the Coloniality of Being/Power/Truth/Freedom: Towards the Human, After Man, Its Overrepresentation – an Argument”, *CR: The New Centennial Review*, Vol.3, No.3 (Fall 2003): 257-337.

Eliezer Yudkowsky, “Cognitive biases potentially affecting judgement of global risks,” in Bostrom and Ćirković (eds) *Global Catastrophic Risks* (Oxford: Oxford University Press, 2008), 91-119.