Fairness and the Future: evaluating extreme technological risks

Simon Beard – Centre for the Study of Existential Risk

This paper will consider the ethics deploying dual use technologies that pose potentially existential threats and other catastrophic effects at the global scale.

In particular, it will focus on technologies where there is a reasonable degree of certainty about the size of its potential costs and benefits (for example human extinction) but a high degree of uncertainty about the likelihood of these occurring. In such cases it is often desirable to seek to reduce the degree of uncertainty about this risk profile and so allow us to evaluate these risks against the expected benefits from the activity. However, in this article I will argue that there are important distributional concerns about global catastrophes that may be easier to establish when evaluation of such technologies, and that may be more significant in determining whether they should be deployed than the size of the risks involved. I will characterise these concerns in terms of 'fairness', an overarching set of distributional concerns that are grounded in the claims of individuals not to be worse off than they might have been.

In the first section of this paper I will briefly introduce three examples of dual use technologies that pose significant, yet highly uncertain, existential threats: gain of function' research on potentially pandemic pathogens, geoengineering via solar radiation management and the pollution of carbon dioxide and other greenhouse gasses.

In the second section of this paper I will introduce an overarching conception of fairness that I will use and argue that our intuitions about fairness can be captured by a simple set of principles whose moral importance is marginal on their own, but becomes very significant when they are violated together, compounding the badness of individual instances of unfairness.

In the third section of this paper I will apply my model of fairness to the dual use technologies I described in section 1 and show how concerns about fairness may give us reason to rule some of these technologies in and others out, even given the high degree of uncertainty surrounding their risk and benefit profiles.

Finally, in section 4 of this paper I consider a challenge to such fairness based evaluations of globally catastrophic risks based on the so called 'non-identity problem'.

1 – Dual use technologies presenting potentially extreme risks

The following three technologies are all carry potentially global catastrophic and / or existential risks. However, in each example the probability of these consequences is not known. In such cases it is common to see arguments that we should adopt a precautionary approach, either avoiding such technology entirely or at least waiting until the size and likelihood of its risks are better understood. However, as I will argue below, such an attitude misses at least one important aspect of these behaviours, the size and distribution of their expected benefits. As I will argue in the following section, the moral importance of who is likely to benefit from these technologies, and by how much, is such that it can swamp the uncertainty that surrounds the costs associated with these technologies, and this may allow us to make morally justifiable evaluations concerning them even under the current conditions of uncertainty.

<u>Gain of function research on potentially pandemic pathogens</u>

In order to understand the spread of pandemic pathogens it is possible for scientists to genetically modify pathogenic bacteria and viruses under laboratory conditions to enhance their pathogenicity and transmissibility. Such 'gain of function' experiments allow scientists to better understand how pathogens evolve and how they can be more effectively tackled. In 2013 scientists modified a strain of H5N1 influenza (avian influenza) making it transmissible between ferrets - and therefore theoretically transmissible between humans as well[i].

Such research poses a number of threats. One key danger lies in the possibility of laboratory escape of pathogens leading to a global pandemic. Depending upon the functionality of the pathogens that are released such a pandemic could produce a global catastrophe leading to the deaths of billions of

people and have knock on effects that pose potentially existential threats to humanity. The possibility of such a catastrophe is not seriously in doubt, but its likelihood is hard to establish. One prominent risk / benefit analysis of this research argued that the historical incidence of laboratory escapes and man-made pandemics should be used to estimate the likelihood that gain of function research could have catastrophic outcomes, producing an expected cost of such research between 2,000 and 1.4 million human fatalities per laboratory undertaking this kind of research per year[ii]. Replying to this analysis another researcher argued that in-fact every historical instance of a laboratory escape or man-made pandemic occurred under very different conditions to those of the proposed Gain of Function research, and that the likelihood of catastrophic risks was therefore much lower than this historical data would suggest. This study argued that the expected costs cost of such research would be more than a million times lower, potentially as low as 0.00002 fatalities per laboratory per year of research[iii]. Both of these estimates are extreme, but it is hard to establish accurate probabilities for such rare and unpredictable events as laboratory escapes. In the absence of more robust risk / benefit analyses Gain of Function Research remains highly controversial and is currently the subject of a moratorium in the USA and elsewhere.

Another morally important feature of gain of function research are the expected benefits. These too are hard to assess. However, there are two likely scenarios for how this research could be used. In the first of these the research will be used to understand what makes certain kinds of influenza virus cause global pandemics and therefore to prevent such pandemics from occurring in the future. Since the 18th century there have been 6 such pandemics, each of which would, if it had occurred today, produce in excess of 2 million deaths. Indeed, just one of these, the influenza outbreak of 1918, was responsible for over 50 million deaths, though the global population in 1918 was less than a quarter of what it is today. Due to their severity it is estimated that the long run expected deaths from such pandemics average out to around 700,000 per year[iv], with most of this these expected to be amongst the worst off people in impoverished countries.

On the other hand, in the alternative scenario the research is used to provide quicker and more effective treatments for seasonal influenza or is used to create treatments that are only aimed at treating individual cases, rather than preventing an entire pandemic. Whilst seasonal influenza affects many people each year, especially in impoverished countries, its expected costs are significantly lower than pandemic influenza. Nevertheless, significant amounts of money are currently spent on the prevention and control of seasonal influenza, especially in wealthy countries, so that this kind of research may be more economically viable than a wide scale global effort to eliminate pandemic influenza strains for good. Furthermore, since such research has more modest aims it is more likely to succeed without producing any incidence of pathogen escape. Together these considerations suggest that this scenario is more likely to reflect the way that such research will be conducted in the near future, implying that the benefits of such research may be distributed heavily towards those most able to afford to pay for them[v].

Geoengineering via Solar Radiative Management

In order to reduce the amount of heat energy in the earth's atmosphere, which causes climate change, it is possible to inject aerosol particles into the stratosphere that reflect a portion of the sun's energy back out into space. Such technology, which mimics the effects of natural volcanic eruptions, has been proposed as a means of preventing, and partially reversing, anthropogenic climate change.

Accepting for the moment that such geoengineering would be successful, the precise risk I will consider in this paper relates to the interruption of the aerosol injection. Whilst climate change from greenhouse gas emissions can be expected to last for hundreds or thousands of years, aerosols injected into the upper atmosphere are soon rained out and thus cease to produce any cooling effect after a year to 18 months. If aerosols were injected into the atmosphere merely as a mask for the global warming produced by greenhouse gas emissions and the flow of aerosols was subsequently interrupted, this could lead to very rapid warming of the earth's atmosphere. Rather than 2 to 3 degrees of global warming occurring over a period of 50 to 100 years, which will be exceedingly dangerous and hard to adapt to, this level of warming could then occur over a period far too short for any hope of successful adaptation, leading to a global catastrophe and potentially an existential threat[vi]. Such interruption might be triggered by a social catastrophe that removes our ability to continue injecting aerosols into the stratosphere in sufficient quantities, such as a global pandemic or nuclear war, or from some unanticipated side effect of the process that renders it too costly to continue.

Geoengineering of this kind can therefore only be considered 'safe' if it is used merely as a means of adapting to changing levels of greenhouse gasses by masking some of their worst effects. It should not be seen as a safe and reliable alternative to climate change mitigation that would allow high levels of emissions to continue. Policy makers therefore face a choice between allowing geoengineering but continuing to cut greenhouse gas emissions (safe, but highly costly), allowing greenhouse gas emissions to continue to grow (low cost, but with potentially catastrophic consequences if the flow of aerosols is interrupted) or implementing geoengineering alongside reductions in carbon dioxide emissions, masking some of the worst effects from present and future climate change.

Greenhouse gas emissions

A final technology that I will consider is the emission of climate changing greenhouse gasses itself. The implications of this are harder to model succinctly[vii]. However, in essence the existential risks posed by climate change are significantly increased by 1) the amount of greenhouse gasses that will be emitted and 2) the amount that is spent on adapting to global climate change. The key question that I will consider in this paper however is not how much money is spent on either of these things, but whether their cost is born disproportionately by wealthy countries, by poor countries or by all countries in proportion to their ability to pay. Most accounts of climate justice imply principles such as the 'polluter pays', that in practice, though often not in principle, imply that costs are to be born predominantly by richer countries. Some accounts however argue that since poorer countries have the most to lose and are currently the lest efficient polluters that they should bare the most costs[viii].

Evaluating these technologies

In what follows I will attempt to justify the following conclusions about these three technologies, despite the significant degree of uncertain surrounding each of them: 1) that gain of function research is probably justified so long as it is conducted to find ways to prevent influenza pandemics but almost certainly not if its main goal is merely to treat influenza or to prevent seasonal influenza, even if the cost benefit ratios of the two kinds of research are broadly the same, 2) geoengineering via solar radiation management is justifiable so long as greenhouse gas emissions are reduced quickly and if any additional emissions are focused towards helping the poorest countries to develop and 3) that the costs of climate change should be bourn disproportionately by the wealthiest countries. My argument is that to do otherwise would be unfair. By 'unfair', I mean that they would contradict a specific set of principles that I take to express something close to our innate feeling about what is fair and unfair, and that I set out below.

2 – Principles of fairness

By fairness I mean, roughly, moral reasons for evaluating an outcome based not upon the good of people in it, but on their relationships. Such relationships can be real, social, relationships (social and procedural justice) or implied, proportional, relationships such as 'worse off than' or 'equally well of as' (distributional justice). In this paper, I shall primarily consider fairness in terms of proportional justice. Note that whilst social relationships can only exist between people or things contained within a single outcome, implied relationships can exist between people in the same outcome (being better or worse off than others) different possible outcomes (being better or worse off than one 'might have been') or purely hypothetical outcomes (being better or worse off than one 'deserves to be').

Fairness in this sense is important for a number of reasons. Firstly, it has been shown to be a significant motivational force on human behaviour. Individuals often react more strongly to how they fare in relation to others, in relation to how they might have fared or in relation to how they feel they deserve to fair than to how well off they actually are. Whether or not we believe we can produce the 'correct' evaluation of an outcome without such considerations of fairness therefore, if we wish to actually implement our evaluations within a social or political context such considerations should be included and made explicit[ix]. Secondly, without considerations of fairness, moral evaluations struggle to take account of the moral standing of individuals within the aggregate value of an outcome and face the 'separateness of persons' objection[x]. Within the context of evaluating future technologies this manifests itself as a tendency to focus on whether total costs are greater than or less than total benefits without considering the distribution of costs, and especially, the distribution of benefits. As I will argue, these issues are of great significance in our evaluation of these technologies.

As previously mentioned, there are two aspects to fairness: Social justice, which deals with the value of real social / procedural relationships, and proportional / distributive justice, which deals with the hypothetical relationships 'better than' and 'worse than', which may exist between an individual and some other real, possible or hypothetical person. In this paper, I will only be considering proportional / distributional justice[xi].

The conception of fairness that I present draws on two important theories in proportional justice. The first of these, Luck Egalitarianism, has proven to be highly compelling and influential, although it has faced a great many criticisms. According to Luck Egalitarianism, it is unjust for one to be worse off than others through no fault or choice of one's own[xii]. One key insight that I take from this is that proportional justice is not simply about any single kind of implied relationship, but rather represents a unifying conception of fairness that brings together ideals that depend upon quite different kinds of relationships, such as actual relationships of equality and inequality and hypothetical relationships of proportional desert.

The second theory that I draw on is a recent, and highly successful, conception of fairness that has been developed to deal with many of the critiques levelled at Luck Egalitarianism, known as the Competing Claims view, which has been developed by Mike Otsuka, Alex Voorhoeve, Mark Fleurbaey and others. According to this view:

"we decide between alternatives by considering the comparative strength of the claims of different individuals, where (i) a claim can be made on an individual's behalf if and only if his interests are at stake; and (ii) his claim to have a given alternative chosen is stronger: (iia) the more his interests are promoted by that alternative; and (iib) the worse off he is relative to others with whom his interests conflict."[xiii]

There are two key insights that I draw from this theory. Firstly, that unfairness only exists when somebody's interests are at stake, i.e. they are worse off than they might have been (and not simply when they are worse off than others or worse off than they deserve to be). Secondly, that the badness of different aspects of unfairness combine together productively rather than representing distinct elements of an outcome whose values one merely adds together.

Combining these key insights, I therefore proposed a combined conception of unfairness characterised by the following three conditions:

1) It is unfair to be worse off than one might have been[xiv]
2) It is more unfair to be worse off than one deserves to be, except where one could not have been better off
3) It is even more unfair to be worse off than others, except where one either could not have been better off than one is or does not deserve to be so

Fairness in this sense is not strictly egalitarian, although it does show a preference for equality as the worst forms of unfairness are predicated on its existence. This is because, when the worst off cannot be made any better off than they are then there is nothing bad about the inequality that exists between them and others. For instance, it is not unfair that many ancient Egyptians were worse off than ourselves, because they cannot be made any better off. More importantly fairness of this kind involves multiple characteristics that interact in a highly productive manner. The unfairness of simply being worse off than one might have been makes very little difference to the value of an outcome. Only when it is combined with the unfairness of being worse off than one deserves, especially when one is also worse off than others, does the effect of unfairness on the value of an outcome become significant, and even begin to act like a 'trump' to other kinds of value, such as the maximisation of total welfare.

These features of my account seem to reflect many people's intuitions about fairness. For instance, there are many examples of inequality in the literature that are not taken to be unfair, two classic examples are 'levelling down' and 'mere addition', in both of these some people are worse off than other but could be made no better off. If one accepts a view such as that which I suggest here then there is therefore no unfairness in such cases, despite their significant inequality.

Alex Voorhoeve and Mark Fluerbay give another related case, which I summarise below, where the numbers represent the distribution of individual welfare:

Alternative	Person	State of the world (equiprobable)	
		S1	S2
Non-Risky treatment	Albert	1	1
	Bob	0.65	0.65
Risky treatment	Albert	0.95	0.6
	Bob	0.6	0.95

In this case the Non-Risky treatment produces more welfare overall and a no less equal distribution of this welfare than the Risky treatment, yet there seems to be something unfair about the Non-Risky treatment when compared with the Risky treatment. This shows, once more, that there is more to fairness than a concern for equality.

On my theory however, whether the inequality in each case is unfair depends on whether those who are worst off could be better off. In one state of affairs, S1, the worst off person would be slightly better off if we chose the non-risky treatment than the risky treatment, however in the other they would be much better off. Whilst, overall, people are better off if we select the non-risky treatment than the risky treatment, and in no case is their actually less inequality from choosing the non-risky treatment, it is in the interest of the worst off person, Bob, that we select the risky treatment, and this makes selecting the non-risky treatment less fair as a result.

Note however that in cases like this, such unfairness is not taken to be a significantly bad feature of an outcome. Voorhoeve and Fleurbaey speculate that many who hold the competing claims view may still prefer the less fair 'non-risky' treatment to the fairer 'risky' treatment, due to the small increase in overall welfare, and this is a belief that I share.

Another case in which there is acknowledged unfairness, but this makes little difference to our evaluation of outcomes is presented by Larry Temkin in his 'progressive disease 3'[xv]. Let me present a functionally identical but simplified version of this case here.

Wheel of fortune: imagine that the allocation of resources amongst a group of people is to be established as follows. Everyone has a marker placed upon a wheel of fortune. The wheel is spun and whoever's marker ends up in the first position gets the most resources, whoever's marker ends up in the second position gets second and so on until whoever's marker ends up in the last position gets least. We cannot alter this method of allocating resources and we know that the people themselves will not reallocate resources once they have been allocated between them in this way.

If I am offered the chance to spin the wheel, but I know that it has already been spun by somebody else, fairness does not seem to count in favour of me spinning the wheel or against it. However, if instead I am offered the opportunity not to spin the wheel, but rather to move it on one place then it seems unfair for me to do so. This is because the relative gains of 99 people moving to a position 1 lower will be relatively small, but the losses of the person who had previously been in the first position and who will now move to the last position are very much larger. The strength of this persons claim not to be moved therefore outweighs the much weaker claims of everybody else to be moved. However, if I were told that by moving the wheel on 1 place I would also add some small amount to the quantity of resources to be allocated then this would easily outweigh my reasons of unfairness of not moving the wheel on 1 place.

These cases all seem to concur with the notion that there is no unfairness where people's interests are not at stake, i.e. they cannot be made better off, but that the unfairness of merely being worse off than one might have been is trivially small unless combined with other aspects of unfairness as well. On the other hand, there are certain cases in which unfairness is taken to be so bad as to effectively trump most other moral values. One classic example of this is Thomas Scanlon's case of Jones' hand

"Suppose that Jones Has suffered an accident in the transmitter room of a television station. Electrical equipment has fallen on his arm, and we cannot rescue him without turning off the transmitter for fifteen minutes. A World Cup match is in progress, watched by many people, and it will not be over for an hour. Jones's injury will not get any worse if we wait, but his hand has been mashed and he is receiving extremely painful electrical shocks. Should we rescue him now or wait until the match is over?"

In this case the fact that Jones is being made much worse off gives him a claim of unfairness that is supposed to trump our concern for the benefits of prolonging the football match. However, I find it dubious that Jones' claim is supposed to trump any kind of benefit. For instance, consider that instead of watching a football match Jones' had was being mashed by apparatus that would deliver high quality educational content to many of the world's worst off children or that would preserve the entire recorded history of world football from being erased. What make's Jones' claim act like a trump in this case is that his suffering is so much worse than the other people's benefits. Note also that in this case it is not simply the suffering, but the fact that it makes Jones' worse off that is important. If Jones' was himself a football fan and if saving his hand would mean that he also lost out on his ability to watch the game we might believe that we no longer had a reason to save him, even though the additional benefit that Jones' would receive is very slight, compared to the enjoyment of everyone else.

What I take from this case, and others similar to it, is that claims of unfairness can become significantly more or less important depending upon factors other than the extent to which an individual is being made worse off, and that whilst inequality is not valuable in itself, even relatively small differences in the distribution of goods and benefits can turn a claim of unfairness from being virtually trivial to being a claim strong enough to trump very significant benefits to others.[xvi]

3 – How the principles of fairness evaluate the dual use technologies

Technologies that pose existential, or globally catastrophic, risks inherently impose risks upon everyone, and thus give every individual on earth a potential claim of unfairness against their deployment. Whether, in weighing these claims against the claims of those who also stand to benefit from the technology's deployment, we ultimately judge their deployment to be fair or unfair will depend crucially not only upon size and distribution of these risks, but also on the distribution of benefits across these people. Assuming that their benefits can be expected to outweigh these expected costs there are three possibilities

1) The benefits will be distributed so that every individual is expectedly better off than they would otherwise have been. In this case the complaints of unfairness created by the distribution of risks and benefits are effectively neutralized
2) The benefits will be distributed in a way that does not mean that everybody can expect to benefit, but they will tend to benefit the worst off (or more deserving) more than the better off (or less deserving) - in this case there will still be potential claims of unfairness from those who cannot expect to be made better off, but these are likely to be trivial
3) The benefits will be distributed in a way that does not mean that everybody can expect to benefit, but they will tend to benefit the better off (or less deserving) more than the worse off (or less deserving) – in this case there will be complaints of unfairness but these are likely to be trumping, or close to trumping.

This issue can be most readily seen in the case of gain of function research. Whilst the expected costs of such research will affect everybody, the expected benefits will be highly dependent upon what kind of research is undertaken. If research is used to find effective treatments for pandemic influenza and could prevent even a single pandemic on the scale of the 1918 influenza outbreak, then it is likely that the research would be of considerable benefit to all of humanity and would not present any problems of unfairness. On all but the most pessimistic assessments of the risks associated with this kind of research therefore it can be justified based on its substantial net benefits, and should be allowed to take place[xvii].

On the other hand, if the research were used to find more effective treatments for seasonal influenza, then it could still impose expected costs upon everyone, but would only now be of benefit to a minority, those who are wealthy enough (and concerned enough about the state of their health) to seek treatment for seasonal influenza. Whilst potentially economically lucrative such treatment would not constitute a benefit to all of humanity and would disproportionately benefit the best off. Even if the research turned out to be relatively safe therefore, so that the expected benefits outweighed the expected costs, this particular distribution of these costs and benefits could create claims of unfairness that were trumping, or near trumping, thus effectively ruling out such research as unjustifiable.

The case of geoengineering via solar radiation management presents more challenges for an analysis in terms of fairness. This is because whilst the geoengineering itself can be expected to benefit

everybody by reducing the need to adapt to climate change, it will primarily benefit those areas that would expect to find it hardest to adapt to climate change, which are predominantly poorer less well-off countries and poorer more marginal groups within richer countries[xviii]. Given that the existential threats posed by such geoengineering would affect everybody, there is therefore a prima facie case for arguing that this might be unfair on wealthier countries. However, as I have argued, such complaints would tend to be more trivial and less trumping because they would affect those who are already best off (and arguably least deserving), making this kind of geoengineering justifiable so long as its expected benefits outweigh its expected costs to a reasonable degree.

However, if one conceptualizes the policy choice as not being merely between introducing geoengineering via solar radiation management and not doing so, but between the three options of introducing this kind of geoengineering with or without reductions in carbon emissions and not introducing it, then the nature of the claims of fairness changes. Whilst poor countries stand to benefit more than rich countries as a result of the geoengineering itself, it is likely that rich countries will benefit somewhat more than poor countries from continuing to pollute at levels that make this kind of geoengineering dangerous. This would render this course of action predominantly beneficial to these countries and thus make the risks it imposes significantly unfair.

In effect the fact that this kind of geoengineering imposes risks upon all people equally means that it would be unfair for wealthy countries to benefit from implementing it, unless this can be shown to be in the interest of poorer countries as well. This would seem to imply either that action must be taken to remove the risk of potential global catastrophe associated with this (and other) kinds of geoengineering or the benefits from them should be allocated predominantly to smaller countries, for instance by allowing them to make limited additional emissions of greenhouse gasses in order to encourage development or by compensating them for not doing so[xix].

This brings us to the final kind of technology I am considering in this paper, greenhouse gas emissions themselves. In this case the situation is made even more complicated by the fact that such emissions are already taking place and, in conjunction with other kinds of environmental change, causing significant amounts of harm, in particular to many of the worst off people living in marginal environments.

Apart from the desperate need to reduce such greenhouse gas emissions and curb climate change two key issues are whether climate justice should be forward looking, or whether it should take account of past emissions, and how the costs and benefits of emission reductions should be distributed. In discussing these two issues I will not even attempt to do service to the rich and considered literature that exists on this subject, but merely to consider what concerns about fairness and proportional justice of the sort that I outlined above might have to say on the issue.

Firstly, note that on the account I have offered here fairness is not grounded so much in a static account of how well off people are, but in a dynamic and relational account. What matters are whether people are being made better or worse off and whether they stand in relations of fairness to others. One thing that is immediately apparent when we consider fairness in this way is that the current situation is not fair. People are being made worse off by climate change who have not benefited significantly from industrialization, and there exists an established global energy market that works to the advantage of many of the better off whilst disenfranchising, and in many cases actively oppressing, the worse off. Such features of the current global economy are widely acknowledged but often removed from debates about climate change on the grounds that they are not directly causally connected with changes in the climate. For instance, it can be argued that individual polluters did not intend to produce such an unfair starting position for climate change negotiations and therefore cannot be held responsible for it. However, it is not necessary for a situation to be proportionately unfair for its participants to have intended it to be so (rather this would be a matter for social or procedural justice). The mere fact that one has ended up better off than another 'through no fault or choice of their own' is often taken as an instance of proportional unfairness in need of some form of restitution. In this case the fact that some have done well via the pollution of greenhouse gasses and ended up much better off than they would otherwise have been, whilst others have done badly as a result of an unfair global economic and industrial systems, the direct oppression of many poor people by resource extractors and the initial effects of climate change. Where this has supported a highly unequal distribution of global resources, makes the continuation of the current economic system robustly unfair.

It also seems clear that a concern about fairness does not necessarily lead us to the conclusion that all should benefit from efforts to tackle climate change. Since inactivity will tend to harm those who are already worst off it is clear that failing to tackle climate change by reducing emissions would be unfair, and that such claims of unfairness would be very strong, able to trump other claims when weighed against them. At most the burden of responding to these claims of unfairness would fall equally on everyone, giving us all equal reason to act to bring about reduced greenhouse gas emissions, despite the fact that this would benefit some more than others. However, given the background conditions of unfairness it seems hard to justify the view that those who have benefited from climate change, and who benefit most from current levels of pollution, should do most to reduce the harm from these emissions. Even if this were to make the world's wealthiest worse off than they would otherwise be, the fact that these people would still be better off than the rest of the world's population, and that they are unlikely to deserve to be as well off as they are means that any claims of unfairness resulting from this sort of harm would be morally less significant than those arising from the existing claims of unfairness or those that would be produced by allowing climate change to continue.

4 – Objections from non-identity

The account of fairness that I have developed in this paper is grounded on the idea that people are made worse off than they would otherwise be. Such accounts often face objections on the grounds that future people cannot be harmed in this way – that anyone who does not yet exist cannot be made worse off than they would otherwise have been by our current actions, because these actions will in part determine the environment in which they come into existence and hence their identity. This is known as the non-identity problem[xx].

In order to respond to this objection, we should distinguish two sense in which one can be worse off than one might have been

1 – one can be worse off specifically because of the action under consideration
2 – one can be worse off than one would be in some alternative possible outcome

The essence of the non-identity problem is to claim that as a result of our present actions one cannot be worse off than one might have been under the second sense, because all the possible outcomes one might face are ones in which we would already have acted the way that we have. Our current actions therefore cannot be said to harm a future person, and thus to treat them unfairly.

It might be assumed that one cannot be worse off than one might have been under the second of these interpretations then one cannot be worse off than one might have been under the first interpretation either. On this reading the first interpretation merely selects from all the possible outcomes in which a person might have existed only those in which they are better or worse off as a direct result of our actions *and nothing else*. However, this is not the only way in which we might understand these claims. In order to appreciate this, it can be helpful to differentiate between identity contingent and non-identity contingent effects of our actions

An effect is identity contingent if individuals are affected as a necessary result of their identity. For instance, if our actions cause a person to be brought into existence who will necessarily be worse off, for instance because they suffer from a genetic disease.

An effect is non-identity contingent if individuals are affected in ways that are not related to their identity. For instance, a non-identity contingent effect would be more or less the same no matter who experiences it. An example of a non-identity contingent effect would be causing a bomb to explode in 100 years' time. This bomb would harm anybody who was near to it, irrespective of who they were or what they were like. The effects of this action are therefore not at all contingent on the identity of the person being affected by it[xxi].

If our actions have non-identity contingent effects and they cause a person to be worse off, then, even if they affect somebody who does not yet exist, there is still a sense in which they can be said to make that person worse off. This is because the conditional statement that that person would have been better off if they had existed in a world in which I did not act the way that I did could still be true. Note that this conditional statement will be true even if the antecedent claim, that the person could have existed in a world in which I had not acted in the way that I did, could not be true because my actions predated the coming into existence of this individual. If we accept that the truth of such a conditional

statement is sufficient to ground the claim that this person is worse off because of my action and that this further gives them a claim of unfairness on the grounds that they could have been better off, then we can defend the view that this person has been treated unfairly[xxii].

Of course we might deny that simply being made worse off by my actions is sufficient to show that one is worse off than one might have been, or that if there is no possible outcome in which one would actually be better off then there is no claim of unfairness from being made worse off by my actions. Yet it is not unheard of for conditional statements such as this to ground important moral facts. For instance, many people now share Harry Frankfurt's view that possessing free will, a very important moral consideration indeed, is not a matter of being in a position where one could actually have chosen to act differently than one has, but merely of being in a position such that it is conditionally true that if one's volition (ones second order desire about what one wants) had been different, then one would have acted differently[xxiii]. Whilst the cases of free will and fairness are not so similar to one another, I hope that the success of Frankfurt's theory of free will suggests that we should not rule out the moral significance of this kind of conditional truth out of hand.

Conclusion

In this paper I have argued that when considering dual use technologies that present existential or globally catastrophic risks it is not enough to focus on the potential costs associated with these outcomes, fairness demands that we consider the distribution of benefits as well. By exploring three different kinds of technology I have demonstrated how thinking about fairness might influence our decision making. In cases where a technology presents a threat of globally catastrophic or existential proportions I have argued that as such they can only be justified if their deployment is in the interests of everyone. Where there is already harm being done to the worst off and / or most deserving however, I have argued that fairness may not demand that everybody be benefitted in attempts to reduce or remove this harm. Rather I have argued that it is perfectly compatible with my account of fairness to take into account past and present injustices and to make the best off somewhat worse off, so long as this does more to help the situation of those who are worst off and who are currently being harmed the most. Finally, I have defended this view from the claim that it faces the non-identity objection on the grounds that even if it is sometimes true that there is no outcome in which an individual might be better off, they can still be made worse off because of our actions on the grounds that it can still be conditionally true that if they existed in a world in which we had acted differently they would have been better off.

---

[i] Imai, M., Watanabe, T., Hatta, M., Das, S.C., Ozawa, M., Shinya, K., Zhong, G., Hanson, A., Katsura, H., Watanabe, S. and Li, C., 2012. Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. *Nature*, *486*(7403), pp.420-428.

[ii] Lipsitch, M, Inglesby, T. 2014. Moratorium on research intended to create novel potential pandemic pathogens. mBio 5(6):e02366-14.

[iii] Fouchier, R. 2015. Studies on influenza virus transmission between ferrets: the public health risks revisited. mBio 6(1):e02560-14.

[iv] Fan, V, Jamison, D and Summers, L. 2016. *The inclusive cost of pandemic influenza risk* (No. w22137). National Bureau of Economic Research.

[v] Selgelid, M. 2016. *Gain of Function Research: ethical analysis.* Sci Eng Ethics (2016) 22: 957.

[vi] Baum, S. D., Maher Jr, T. M., & Haqq-Misra, J. 2013. Double catastrophe: intermittent stratospheric geoengineering induced by societal collapse. *Environment Systems & Decisions*, *33*(1), 168-180.

[vii] primarily because they are much better understood, and therefore harder to simplify

[viii] E.g. Posner, E and Weisbach, D. 2010. *Climate change justice*. Princeton University Press.

[ix] Dulebohn, J, Conlon, D, Sarinopoulos, I, Davison, R. and McNamara, G. 2009. The biological bases of unfairness: Neuroimaging evidence for the distinctiveness of procedural and distributive justice. *Organizational Behavior and Human Decision Processes*, *110*(2), pp.140-151. Narvaez, D. and Vaydich, J., 2008. Moral development and behaviour under the spotlight of the neurobiological sciences. *Journal of Moral Education*, *37*(3), 289-312.

[x] Otsuka, M. and Voorhoeve, A. 2009. Why it matters that some are worse off than others: an argument against the priority view. *Philosophy & Public Affairs*, *37*(2), 171-199.

[xi] Many accounts of dual use technology deal solely with procedural justice issues, hence I will not address these here. For an alternative account values that might fall into the category of social justice see Wolff, J. 2006. Risk, fear, blame, shame and the regulation of public safety. *Economics and Philosophy*, *22*(03), 409-427.

[xii] Temkin, L.S., 1993. *Inequality*. Oxford University Press.

[xiii] Fleurbaey, M., & Voorhoeve, A. 2012. Egalitarianism and the Separateness of Persons.

[xiv] Note that this claim does not simply reduce to the view that it is unfair that there is not more welfare in the world. It is not unfair that somebody is badly off because they are disabled when there is no possible 'cure' for their disability where the only other choice was that they never came into existence – although as I shall

discuss below there are cases in which it can be unfair for a person to be badly off even if the only alternative for them could have been non-existence.

[xv] Temkin, L. 2012. *Rethinking the Good*. Oxford. OUP p 441-442.

[xvi] One final point about fairness. It has been demonstrated by several authors, most notably Larry Temkin, that considering fairness as part of our evaluative mechanisms can prevent us from producing fully transitive orderings of possible outcomes, because whether one outcome is fair or not will depend very much on the other alternatives that are available. This point applies very much to the kind of fairness that I consider here and I am aware that this will make my views unattractive to some. Nevertheless, for the reasons I set out above I view fairness as a very important consideration from both normative and practical perspectives and I believe that its intransitivity is less problematic, and easier to justify, then that of other kinds of principles that are often invoked when assessing dual use technology, such as the principle of precaution and the Pareto principle.

[xvii] Note that in this case both the Gain of Function Research and the naturally occurring influenza may pose globally catastrophic or existential threats due the significant damage any kind of global pandemic could do to our interconnected global systems and the fragility of our current food / water / energy nexus. It is possible that once the global economy becomes more resilient an influenza on the scale of the 1918 influenza pandemic might cease to pose such risks, which would fundamentally alter the fairness of this research.

[xviii] Hortin, J. and Keith, D. 2016 Solar geoengineering and obligations to the global poor in Preston, C. ed *Climate Justice and Geoengineering*. Rowman and Littlefield 79-93

[xix] Preston, C. 2013. Ethics and geoengineering: reviewing the moral issues raised by solar radiation management and carbon dioxide removal. *WIREs Climate Change 4*. 30-31

[xx] Parfid, D. 1984. *Reasons and Persons*. OUP. 371-377

[xxi] I shall add that, in my opinion, if the only possible alternative to one person existing is nobody existing, i.e. if their existence is a matter of 'mere addition' then we have effected them in an identity contingent manner, thus satisfying the claim that such cases of mere-addition do not ground claims of unfairness. However, I recognise that such a claim is rather controversial.

[xxii] As I have already suggested, such claims based on merely hypothetical possible worlds already ground certain claims of unfairness, claims based on desert. To say that I am worse off than I deserve is, I take it, to say that I am worse off than I would have been if I got what I deserved, and this matters whether or not there is an actual possible world in which I get what I deserve.

[xxiii] Frankfurt, H. 1969. Alternate possibilities and moral responsibility. *The journal of philosophy*, *66*(23), 829-839.