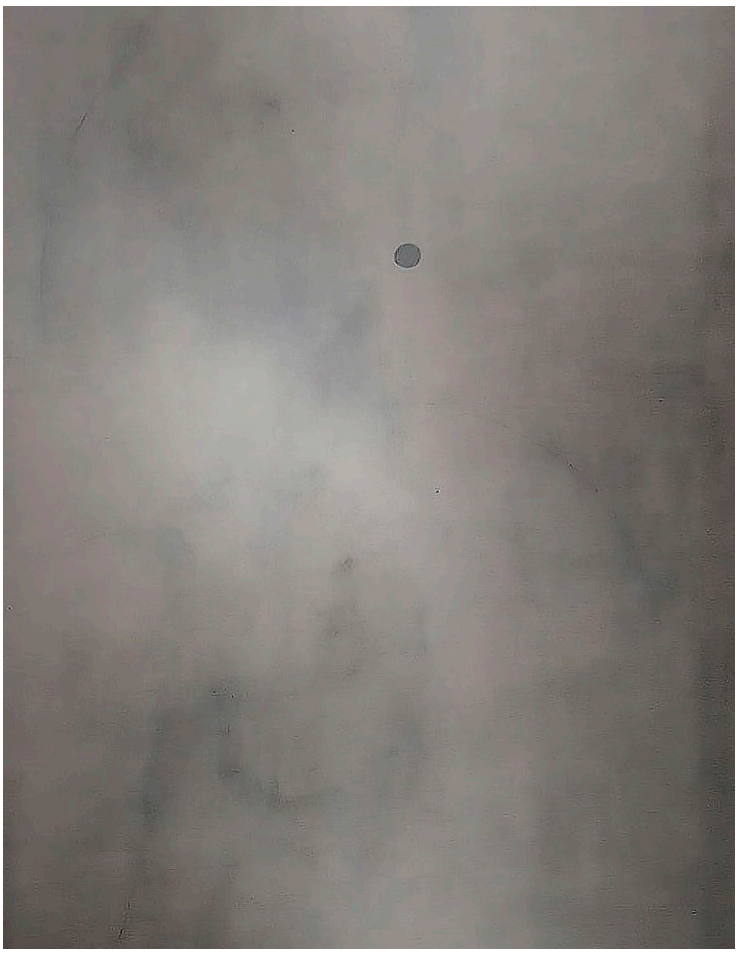
LUCA LO SAPIO

BIOETHICS AND THE ETHICS OF EXTINCTION

ABSTRACT: *BIOETHICS AND THE ETHICS OF EXTINCTION*
This article explores the idea of rethinking bioethics in terms of an Ethics of extinction. After showing the centrality to contemporary ethical debates of extinction, the article shows how bioethics was the first broad disciplinary reflection on the ethical consequences of human actions potentially bringing about extinction. In particular, it will focus on a widely debated issue in bioethics related to our inability as individuals and as a collectivity to deal with current existential risks.



1. *The Earth after* us

In *Dialogo di un folletto e di uno gnomo (Dialogue of an Elf and a Gnome)*[1], the poet and philosopher Giacomo Leopardi *proposes* a scenario that, in some ways, could be *typical* of a thought experiment designed to reason about the effects of the disappearance of sapiens[2].

[1] G. Leopardi, *Dialogo di un folletto e di uno gnomo*, in *Le operette morali*, La Feltrinelli, Milano 2014.
[2] See T. Pievani, *La terra dopo di noi*, Contrasto Editore, Roma 2019.

In a topical passage, the Elf says «all men are dead and the race is lost» and to the Gnome's question «Now how are we to know the news of the world?», the Elf replies «What news? That the sun has risen or gone down, that it is hot or cold, that here or there it has rained or snowed or blown the wind? For when men are gone, Fortune has taken off her blindfold, put on her spectacles, and set her wheel to a harpoon, and sits cross-armed, looking at the things of the world, without putting her hands on them anymore; there are no longer any kingdoms or empires that swell and burst like bubbles, because they are all blurred; there are no wars, and all the years resemble one another like an egg to an egg». To the Gnome's worried remark «nor will it be possible to know how many we are of the month, because no more lunars will be printed», the Elf replies «It will not be so bad, that the moon will not fail the way»[3].

Nature is indifferent to human affairs: this is the apologue of Leopard's *Diaologue*. Nature, as such, does not have a point of view. Therefore, it would "react" to our *cosmic disappearance* with *sovereign indifference*. However, that *Nature does not have a point of view*, and that our disappearance would be *indifferent* for It[4] does not *ipso facto* tell us that our cosmic demise wouldn't be bad, *not only from the point-of-view of* those who disappear, but also *sub specie aeternitatis*[5].

Indeed, as will become clearer in the next paragraphs, in recent times a more structured reflection on sapiens extinction has developed and a range of ethical questions about whether extinction is bad and wrong have begun to be examined.

In particular, three major positions have emerged within the debate:

[3] *Ibid.* (The translation from italian is mine).

[4] It refers to Nature.

[5] On the problem of extinction and the *appropriateness*, in moral terms, of delaying it, there is a vast literature that, among other things, emphasises the need to distinguish between *being extinct* (understood as a fact) and *going extinct* (understood as a process). Among others see E.P. Torres, *Human extinction. A history of the science and ethics of annihilation*, Routledge 2023; W. MacAskill, *What we owe the future*, Oneworld Publication, London 2022; E.P. Torres, *Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks*, Pitchstone Publishing 2017; D. Parfit, *Reasons and persons*, OUP, Oxford 1984.

1) extinction is bad since it deprives us of the possibility of developing future plans and/or deprives Nature of its self-conscious realisation capable of shedding light even on Nature itself;
2) Extinction is not bad in itself but depends on how it occurs and under what conditions;
3) Extinction is a good because it eliminates suffering from the world or because it frees the world from a species that harms it beyond measure.
In the next paragraph, following Torres' analysis, I will show how reflection on extinction has developed recently. However, I already point out here that, although not explicitly recognised in these terms, bioethics, in some of its forms, was the first experiment in systematic reflection on the subject of extinction. Indeed, the *possibility* that our disappearance is an evil, *sub specie humanitatis and sub specie aeternitatis*[6], associated with the growing number of global catastrophic risks[7], has contributed, since the 1960s, to the birth of a multi-voiced reflection, of which bioethics has represented one of the major traces[8].

2. *The Extinction problem*

The focus on extinction has not always been at the centre of the scene. Despite a few exceptions, including Giacomo Leopardi's own reflections, we have to wait until the onset of so-called secularisation to witness the flowering of reflection on the extinction of our species in the Western culture.
Although there is not necessarily agreement on this point, following Torres' reconstruction, five phases or, as Torres defines them,

[6] Here I employ the distinction *sub specie humanitatis* and *sub specie aeternitatis* formulated by David Benatar in *The Human predicament. A candid guide to life's biggest questions*, OUP, New York 2017.
[7] By global catastrophic risk I mean an event whose occurrence could annihilate human life or drastically reduce our ability to maintain current standards of civilisation (cf. N. Bostrom, *The vulnerable world hypothesis*, in «Global Policy», 10, 4, 2019, pp. 455-476).
[8] Other important authors are Günther Anders and Karl Jaspers.

five moods can be identified with respect to the theme of extinction.

1) Indestructibility (ancient times to the 1850s): The concept that human beings are inherently everlasting, in some sense, and cannot be destroyed, can be found in various cosmological theories and mythological systems that have existed since the time of the Presocratics in Ancient Greece. However, this doesn't imply that ancient civilizations never considered the possibility of a universe without us. In cases where they did, it was typically believed to be a temporary situation. In other words, they acknowledged the potential for our extinction in a limited sense, but rejected the notion that we could vanish permanently.

Nonetheless, the belief in our indestructibility took on a more radical form with the rise of Christianity and its dominance over the Western worldview. Over a span of approximately 1,500 years, during this period, it was widely regarded as impossible for humans to be extinguished in any naturalistic way. This belief provided a sense of "Comfort" and "perfect security," as described by notable individuals writing at the end of this era. It reassured people that regardless of what the future held or the catastrophes that might befall humanity, we would ultimately endure forever.

2) Existential Vulnerability and Cosmic Doom (1850s to the mid-twentieth century): The journey began with the revelation of a scientifically validated mechanism of demise, known as the Second Law of thermodynamics. This principle dictates that our earthly or celestial dwelling will progressively become less hospitable to life until it becomes completely uninhabitable. Although physicists initially projected this outcome to be millions of years away, the Second Law unexpectedly revealed that our extinction is not only possible but also ultimately unavoidable. This dual realization inflicted a profound sense of hopelessness upon many individuals, prompting them to question the purpose and significance of life. The backdrop for this existential shift was the waning influence of

religion on notions of human nature and the future of humanity. While the foundations of religious beliefs had already been significantly weakened, the fundamental ideas concerning existence and the end times remained largely intact. Consequently, this paradigm shift, coupled with the decline of Christianity, unleashed a torrent of captivating and imaginative speculations regarding the potential ways in which humanity could meet its end. However, among the scientific community (or natural philosophers), only the Second Law was widely acknowledged as a legitimate threat to our survival.
3) Impending Self-Annihilation (1945/mid-1950s to the 1980s/early 1990s). The shift described here began to emerge around the same time as the onset of the Atomic Age in 1945. However, it wasn't until the latter half of the 1950s that it truly took shape, as leading scientists recognized the potential for even a small-scale thermonuclear conflict to spread deadly amounts of radioactive particles across the entire planet. Over the following decades, a multitude of credible anthropogenic catastrophe scenarios emerged, encompassing various aspects such as nuclear weapons (e.g., the nuclear winter hypothesis), environmental contamination and degradation due to pollution and overpopulation, the threat of runaway climate change, and speculative dangers associated with biological weapons, self-improving artificial intelligence, and atomically precise technologies.
4) Nature Could Kill Us (1980/early 1990s to the late 1990s/early 2000s). The realization emerged from the understanding that natural events such as asteroids, comets, and volcanic supereruptions possess the capacity to impact the entire planet and trigger widespread extinctions, where numerous species vanish within relatively short geological timeframes. Before this realization, which began to take shape in the 1850s and persisted throughout the Cold War era, it was widely accepted that natural catastrophes were confined to specific regions of our planet. However, a significant transformation occurred during the 1980s, coinciding with the

decline of a prevailing scientific framework known as uniformitarianism. This transformation was largely influenced by groundbreaking research demonstrating that non-avian dinosaurs became extinct approximately 66 million years ago due to a colossal asteroid colliding with Earth. The demise of uniformitarianism and the emergence of a disconcerting new framework called neo-catastrophism revealed that our existence is not secure in a benign universe; rather, we are just as susceptible to abrupt annihilation from natural perils as the dinosaurs were. Sooner or later, Nature will endeavor to eradicate its own creation.

5) The Worst Is Yet to Come (late 1990s/early 2000s to the present). Unlike the preceding three shifts in mood, this particular shift was not motivated by the discovery of new methods of destruction. Instead, it was instigated by two significant developments. Firstly, a groundbreaking philosophical perspective emerged regarding the moral significance of averting our own extinction. This perspective directly inspired endeavors to outline a comprehensive understanding of our existential predicament, which can be referred to as the "threat environment." This endeavor involved, to some extent, a shift towards futurism, focusing on emerging and anticipated risks stemming from advancements in biotechnology, synthetic biology, molecular nanotechnology, and artificial intelligence, including the concept of “artificial superintelligence”.

The second factor that triggered this shift was recent research in the field of environmental sciences, revealing the imminent and catastrophic dangers posed by human-induced climate change, global biodiversity loss, and the ongoing sixth mass extinction event. Simultaneously, it became evident that humanity, particularly the Global North, has set in motion a new geological epoch referred to as the “Anthropocene”. This epoch signifies our irreversible impact on the geological record. At the core of this prevailing sentiment was a terrifying suspicion that, despite the dangers experienced in

the twentieth century, the twenty-first century holds even greater perils. In simpler terms, the worst is yet to come[9].

3. *Existential risks and the Ethics of Extinction*

In this context, the concept of existential risk became increasingly important. «An existential risk is one that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development»[10].

Nick Bostrom, in 2013, proposed this definition of existential risk, pointing out, in the same article, that today it is not so much the risk of a natural disaster that we should be concerned about as the risk related to human activities.

Indeed, the unique nature of human action casts us to worry not only and not so much about global catastrophic risks of natural disasters as about the risks associated with human actions intentionally or unintentionally aimed at destroying the planet. An articulated taxonomy of the existential risks our species is currently facing can be found in J. Leslie's *The end of the world*[11]. In particular, Leslie underlines that at least three categories of risks are at stake: 1) Risks that are already well known; 2) Risks that are often unrecognised (among these a first sub-typology is that of natural disasters, a second sub-typology is that of man-made disasters); 3) Risks arising from the acceptance and spread of certain philosophical ideas. Leslie outlines a range of well-known risks encompassing various domains. In terms of potential catastrophes, these include:

[9] E.P. Torres, *Human extinction. A history of the science and ethics of annihilation*, cit., pp. 8-10.

[10] N. Bostrom, *Existential risk prevention as Global priority*, in «Global Policy», 4, 1, 2013, pp. 15-31.

[11] J. Leslie, *The end of the world*, Routledge, London 1996, pp. 3-13. For further proposals for classification of existential risks, largely overlapping with the above, see M. Rees, *On the Future. Prospects for humanity*, Princeton University Press 2018; T. Ord, *The Precipice. Existential risk and the future of humanity*, Hachette Books, New York 2020.

- Nuclear war, posing a significant threat to global stability and survival.
- Bacteriological warfare or acts of terrorism and crime, which could unleash devastating epidemics or intentional harm.
- Chemical warfare or acts of terrorism and crime, with the potential to cause widespread harm and destruction.
- Destruction of the ozone layer, leading to harmful levels of ultraviolet radiation reaching the Earth's surface.
- Emission of greenhouse gases and subsequent global warming, contributing to climate change and its adverse effects.
- Pollution-induced poisoning, resulting from the release of toxic substances into the environment.
- Pandemics, such as the outbreak of highly contagious diseases with severe consequences.

Furthermore, there are often overlooked hazards originating from natural sources, including:

- Volcanic eruptions, capable of causing significant disruptions and environmental impacts.
- Asteroid or comet impacts, posing a threat to life and infrastructure on a large scale.
- Extreme ice ages triggered by passing through an interstellar cloud, potentially leading to widespread ecological disruptions.
- Close Supernova explosions, which could have detrimental effects on the Earth's biosphere.
- Other massive astronomical explosions, causing potential hazards to our planet.
- Unpredictable collapses of complex systems, with potential cascading effects.
- Unforeseen and unimaginable events that fall beyond our current understanding.

In addition, there are often unacknowledged hazards associated with human-made disasters, including:

- Insufficient preparedness for raising future generations, leading to potential social and economic consequences.
- Disasters resulting from the engineering and subsequent release of genetically modified organisms into the environment.
- Disasters caused by the replication and release of nanomachines, with possible unintended consequences.
- Potential risks stemming from the development and deployment of Artificial Intelligence technologies.
- Disasters arising from the alteration of delicate ecological balances due to specific agricultural practices.
- The hypothetical creation of a new Big Bang in a laboratory setting.
- The theoretical possibility of a transitional phase that could lead to the annihilation of everything.
- The potential destruction of our civilization by extraterrestrial intelligences.
- Unknown and unpredictable risks that defy current comprehension.

Lastly, there are risks associated with the acceptance and propagation of philosophical and religious ideas, including:

- The placement of individuals with dangerous philosophical or religious ideas in positions of strategic influence for global governance.
- The influence of Schopenhauerian pessimism, which can shape negative worldviews and impact decision-making.
- Ethical relativism, emotivism, prescriptivism, and other doctrines that challenge the existence of objective values.
- Negative utilitarianism, a moral perspective that emphasizes minimizing suffering over promoting well-being.
- Granting moral weight solely to existing individuals, potentially overlooking future generations' interests.
- The problematic application of human rights theory, which may neglect risks associated with overpopulation in the pursuit of an unconditional right to procreation.

- Overconfidence in the rationality of others, as exemplified by the prisoner's dilemma.
- The notion that a state has the right to retaliate with nuclear attacks in response to a nuclear assault, driven by a desire for revenge[12].

These extensive lists emphasize the breadth and complexity of the risks we face across various realms, urging us to take proactive measures to mitigate and address them effectively. From this taxonomy it is possible to understand that, although the probability of occurrence of a global catastrophic risk could be not so high, it is at least not negligible and the values at stake are high.

4. *Bioethics as an Ethics of extinction*

The History of bioethics, as some reconstructions point out, is to be explored *from* (at least) two trajectories: on the one hand, a *narrow bioethics*[13] or biomedical ethics[14], whose main goal is to investigate the human conduct in the field of life sciences and health care in the light of moral principles and values[15]; on the other hand, a *Global Bioethics*[16] or Ecological bioethics, whose main goal is to investigate the human conduct *in relation to the* environment and non-human living beings in the light of moral principles and values.

«Potter's bioethics does not coincide with the bioethics we all knew from the 1970s onwards, i.e. focused on the new frontiers of biomedicine, and carried out with excellent means by Hellegers and the Kennedy Institute at Georgetown University in Washington DC»[17],

[12] J. Leslie, *The end of the world*, cit.
[13] K. Ferguson, *The health reframing of climate change and the poverty of narrow bioethics*, in «Journal of Law, Medicine and Ethics», 48, 4, 2020, pp. 705-717.
[14] See T. Beauchamp, J. Childress, *Principles of Biomedical Ethics*, 7ª ed., OUP, New York 2013.
[15] This is Warren Reich's definition of bioethics introduced in *Encyclopedia of Bioethics'*, first edition.
[16] See V.R. Potter, *Global Bioethics*, Michigan State University Press, Michigan 1988.
[17] G. Russo, *La Bioetica di Van Rensselaer Potter*, in V.R. Potter, *Bioetica. Ponte verso il futuro* (1971), Sicania, Messina 2000, pp. 24-25.

rather it identifies with «a bioethics where the quality of man's physical life (*Medical Bioethics*) was cybernetically coordinated with the quality of environmental and ecological life (*Ecological Bioethics*)»[18].

In the article *Bioethics. The science of* survival[19], considered by many to be the birthplace of bioethics, Potter states that «humanity is urged by the urgency to develop a new form of wisdom that provides "the knowledge of how to use knowledge" for the survival of humanity and the improvement of the quality of our lives. This concept of wisdom as a guide for action - the knowledge of how to use knowledge for the collective good - could be called the "science of survival", no doubt being the pre-requisite for improving the quality of our lives»[20] .

A few years after the publication of *Bioethics. The science of survival* and *Bioethics. A bridge to the future*, it was the turn of Hans Jonas with his *Principle of Responsibility. An Ethics for Technological Civilisation*[21]. Here, the German philosopher writes that there is a need for «an imperative appropriate to the new type of human action and oriented to the new type of acting subject, it would sound something like this: «Act in such a way that the consequences of your action are compatible with the permanence of genuine human life on earth», or, translated into negative «Act in such a way that the consequences of your action do not destroy the future possibility of such life», or, translated simply: "Do not endanger" "the conditions of humanity's indefinite survival on earth", or again, translated positively: "Include in your present choice the future integrity of mankind as the object of your will"»[22].

[18] *Ibid.*, p. 25.
[19] V.R. Potter, *Bioethics. The science of survival*, in «Perspectives in Biology and Medicine», 14, 1, 1970, pp. 127-153.
[20] *Ibid.*
[21] H. Jonas, *The imperative of responsibility. In Search of an Ethic for the Technological Age* (1979), University of Chicago Press 1984.
[22] *Ibid.*

The common thread running through bioethics is the connection between The Science of Survival and The Responsibility Principle, which emphasizes the importance of delving into a thought process I call "the Ethics of extinction". Recognizing the profound existential threat posed by human actions compels us to deeply contemplate the notion of extinction, to the extent that it validates the claim that bioethics, particularly in its Potterian and Jonasian forms, encompasses an Ethics of extinction. The term "Ethics of extinction" serves to highlight two essential points. Firstly, it underscores that even when extinction is not explicitly mentioned, it remains the primary focus of investigation. Secondly, it emphasizes that ethical reflection is structured through the interpretative lens provided by the imminent existential threat we face. By employing the expression "Ethics of extinction", we draw attention to the overarching theme of extinction and acknowledge that ethical considerations are shaped by the perspective of being under such a profound threat. In fact, after a period of partial obscurity, in which the term bioethics was not even associated with the name of Potter[23], we are now faced with a *rainassance of* Potterian bioethics. In simpler terms, the field of global bioethics has experienced a resurgence in importance compared to biomedical ethics. As a result, many analytical authors who were previously focused on biomedical ethics and clinical bioethics have shifted their interests towards global bioethics issues. Particularly, they have become increasingly concerned with the subject of human extinction and exploring potential strategies to prevent it. This has led to an intriguing convergence between the thoughts of renowned thinkers such as Potter and Jonas, and the contemporary analysis conducted by authors like Nick Bostrom and Julian Savulescu, creating an unforeseen overlap between their reflections.

[23] H.T. Have, *Encyclopedia of Global Bioethics,* Springer, 2016, pp. V-VI.

Indeed, Bostrom's Maxipok principle echoes the imperative of responsibility of Jonas. For instance, Bostrom argues that the potential loss of expected value caused by a catastrophic event with existential consequences is so immense that the primary consideration when acting out of a collective concern for humanity should be to reduce existential risks. It can be helpful to adopt the following guideline for such morally impersonal actions: maximize the probability of achieving an "OK outcome", defined as any outcome that avoids existential catastrophe (Maxipok).

Maxipok is best understood as a practical rule or initial suggestion, rather than a universally valid principle, as there are certainly moral objectives other than preventing existential catastrophes. The principle's value lies in aiding prioritization. Altruistic efforts are not so abundant that we can afford to squander them on numerous projects that have suboptimal effectiveness. If promoting existential safety for humanity yields expected benefits on a significantly larger scale compared to other contributions, it would be wise to concentrate on this highly efficient form of philanthropy.

It's important to note that maxipok differs from the popular maximin principle («Choose the action that has the best worst-case outcome»). Since we cannot completely eliminate existential risks, as we could be wiped out at any moment due to a vacuum phase transition triggered in a distant galaxy billions of years ago, applying maximin in this context would mean selecting actions based on the assumption of impending extinction. Maximin would thus suggest that we should all indulge and enjoy ourselves as if there were no tomorrow. While this implication might be tempting, it is implausible[24].

[24] N. Bostrom, *Existential risk prevention as Global priority*, in «Global Policy», cit.

5. *Longterminism and the future generations*

In his text *What we owe the future*[25], the Scottish philosopher MacAskill points out that our era is, compared to those that preceded it, unprecedented. Not only for what is new, in positive terms (a better quality of life, longer life expectancy, less poverty, greater spread of democracy, etc.) but also for our ability to impact on the balance of our planet, through behaviour whose aggregate effect will have an impact on present and future generations.

With this in mind, Greaves and MacAskill propose the concept of axiological longtermism, which suggests that, in many decision scenarios, the most favorable option beforehand corresponds to the best probability distribution over future events starting from a distant future date, denoted as "t". They argue that a stakes-sensitive argument can be used to derive deontic longtermism from axiological longtermism. Deontic longtermism asserts that, in a broad range of decision contexts, the morally right option to choose is the one that aligns with the optimal probability distribution over future events from the same "t"' date. The argument relies on the Stakes Principle, which states that when the importance of the values at stake is significant, non-consequentialist restrictions and prerogatives become relatively insignificant, implying that the best course of action is simply the one with the highest utility[26].

There are numerous versions of longterminism[27]: some radicals claim that the welfare of future generations is more relevant than the welfare of present generations.

---

[25] W. MacAskill, *op. cit.*

[26] A. Mogensen, *Staking Our future: deontic longterminism and the non-identity problem*, GPI Working Paper - No. 9-2019.

[27] There are different versions of longterminism and a bunch of authors arguing for its main conclusions. Apart from the aforementioned MacAskill see also N. Beckstead, *On the overwhelming importance of shaping the far future*, Ph. D. thesis, Rutgers University Graduate School, 2013; N. Beckstead, *A brief argument for the overwhelming importance of shaping the far future*, in H. Greaves & T. Pummer (Eds.), *Effective altruism: Philosophical issues*, Oxford University Press, 2019, pp. 80-98.

The argument in favor of longtermism is rooted in the recognition of the vast magnitude of the distant future. To be more precise, the potential for both immense value and immense suffering in the far future of civilization originating from humans surpasses that of the near future. This assertion is supported by two key factors. Firstly, there is the aspect of duration. By adopting any reasonable demarcation between the near and far future (e.g., a span of 1000 or 1 million years from the present), it is plausible that our civilization could persist for a significantly longer period than the near future by several orders of magnitude. Even if we conservatively assume that our civilization will cease to exist when the increasing solar energy output renders Earth inhospitable for complex life as we currently know it, we could potentially endure for approximately 500 million years. Secondly, the spatial scope and utilization of resources come into play. If our descendants embark on interstellar colonization endeavors, even at a fraction of the speed of light, they would eventually settle in a region of the universe and harness an abundance of resources far surpassing our present capabilities. These two factors combined indicate that the distant future holds tremendous potential for either exceptional value or significant suffering[28].

The problem with this perspective, even in its less radical versions, is that it clashes with our inability, at least at the individual level, to take action to improve the living conditions of merely potential entities, distant in time and incapable of reciprocation, the latter being, according to some authors, the condition for being able to claim rights[29].

[28] C. Tarsney, *The epistemic challenge of longterminism*, in «Synthese», 201, 195, 2023.

[29] The other issue at stake is the unpredictability of the far future. This characteristic affects both the ability of individuals and institutions to make reliable decisions and the willingness of individuals and institutions to cope with scenarios different from the not-so-distant future ones. This paper won't be focused on this aspect.

Faced with this scenario, two issues need to be brought to the fore. The first deals with the fabric of our societies that do not seem equipped to deal with such situations, the second concerns the general inability of individuals to attend to and care about the long-term consequences of their very actions.

6. *The vulnerable world hypothesis*

The need to adopt different strategies to deal with the global challenges ahead is not only hampered by the problem of our inadequate moral psychology but also and first of all by a social structure that prevents the implementation of appropriate policies. In this paragraph I briefly propose the analysis of the Swedish philosopher Nick Bostrom, who connected his vulnerable world hypothesis with the idea that, from the perspective of global politics, our societies are characterised by a semi-anarchic base condition.

In his essay entitled *The Vulnerable World Hypothesis*, Nick Bostrom illustrates human creativity as a process of extracting balls from a giant urn. Throughout history, we have extracted a wide variety of balls, most of them white, some with shades of grey. These balls represent ideas, discoveries and technological inventions. To date, the cumulative effect of these innovations on the human condition has been mostly positive and may even improve in the future. However, what we have not yet extracted is a "black ball", i.e. a technology that, by its very nature, could destroy the civilisation that produced it. The failure to extract such a ball has so far been a matter of pure luck.

Should technological progress continue, it is practically inevitable that a "black ball" will be drawn. However, we usually adopt an attitude based on the hope that this will never happen.

To quote Bostrom's words: «The vulnerable world hypothesis holds that if technological development continues, there will come a point where a set of capabilities will make the destruction of the

civilisation that produced them extremely probable, unless that civilisation adequately distances itself from the semi-anarchic base condition»[30] .

The scenario presented by Bostrom with his Vulnerable world hypothesis is amplified by the presence of world order that makes the management of risks associated with technological development extremely problematic. For describing such order Bostrom introduces the term "semi-anarchic basic condition".

The term 'semi-anarchic basic condition' refers to a world order characterised by three main elements: 1) a limited capacity to develop preventive policies; 2) limited global governance; and 3) the existence of various motivations driving citizens and groups within a society, including selfishness, welfare, convenience, and even apocalyptic motivations of a residual part of the population, etc.

Firstly, there is a limited capacity to develop preventive policies, indicating a lack of comprehensive measures to address potential issues or challenges proactively. This suggests a reactive approach rather than a proactive one.

Secondly, there is limited global governance, implying a lack of centralized authority or regulatory structures at the global level. This results in a fragmented system where decision-making and governance responsibilities are dispersed among different actors and institutions.

Lastly, the existence of various motivations driving citizens and groups within a society is a defining characteristic of the semi-anarchic basic condition. These motivations can vary widely and encompass factors such as self-interest, the pursuit of individual or collective welfare, convenience, and even apocalyptic motivations held by a residual segment of the population. This highlights the diverse range of driving forces that influence human behavior within this world order.

[30] N. Bostrom, *The vulnerable world hypothesis,* cit.

In summary, the term 'semi-anarchic basic condition' refers to a global state where preventive policies are limited, global governance is lacking, and various motivations shape the actions of individuals and groups within society.

7. The limits of our common sense morality

The semi-anarchic base condition on the other hand is fostered, at least according to a certain line of interpretation, which I share in relevant part, by our inability to react effectively to the challenges around us. Within this framework it is interesting to analyze Persson and Savulescu's view. They firmly believe that while technology has advanced at an extraordinary pace, granting humans unprecedented power to benefit humankind but also to cause harm, the corresponding development of moral psychology has lagged behind. In other words, as technology has progressed, our moral understanding and ethical capabilities have not kept pace.

Persson and Savulescu argue that our moral psychology, which encompasses our innate moral instincts and cognitive processes, has largely evolved during the Pleistocene era. This period, spanning from about 2.6 million to 11,700 years ago, was characterized by a significantly different environment and set of challenges compared to the present day[31] .

Throughout the Pleistocene, humans lived in small hunter-gatherer communities with survival and reproduction being the primary concerns. Both our moral instincts and our ethical intuitions were shaped by the environmental conditions and social circumstances at that time. These instincts provided us with the ability to deal with issues such as cooperation, reciprocity and the sharing of resources within small groups.

Indeed, our common sense morality, shaped by our evolutionary history and personal experiences, may exhibit certain

[31] I. Persson, J. Savulescu, *Unfit for the future. The need for moral enhancement*, OUP, Oxford 2012, pp. 1-2.

characteristics that make it ill-suited for navigating the complex ethical challenges posed by technological development.

One such characteristic is the readiness bias, which refers to our tendency to be more afraid or cautious in situations where we have previously experienced negative outcomes. This bias is a product of our evolutionary adaptation to prioritize avoiding potential dangers and threats to our survival[32].

Furthermore, our common sense morality often operates within an action-omission model and a causality-based conception of responsibility. These schemas shape our moral judgements by emphasising the distinction between actively causing harm through performing an action and allowing harm to occur through omission[33]

Furthermore, our emotional responses and moral judgements are often influenced by a cognitive bias known as the 'near future bias'. This bias refers to our tendency to prioritise and be more emotionally engaged with outcomes or consequences occurring in the near future than those in the distant future.

When we are confronted with ethical dilemmas about technological developments and their potentials for benefit or harm, our emotional reactions tend to be more pronounced with immediate or near future consequences. We tend to be more likely to consider the immediate benefits of a given action or technology, while discounting or minimising the risks or long-term consequences associated with it[34].

Moreover, our moral feelings are spontaneously directed towards close relatives or friends, even if they are not spatially close to us. However, they are similarly not elicited for persons distant from us (affectively speaking), nor do they grow in relation to the number of persons involved in the scenario considered[35]

The parochial (myopic) character of common sense morality does not allow the activation of moral sentiments that go beyond the

[32] *Ibid.*, p. 19.
[33] *Ibid.*, p. 22.
[34] *Ibid.*, p. 27.
[35] *Ibid*, pp. 39-40.

prejudice of the near future, of availability, of causally based responsibility. This leads to what Persson and Savulescu, following Garrett Hardin, call the tragedy of the commons, i.e. the fact that each individual, by his actions, believing he is maximising individual interest, introduces, in the long run, behaviour that is detrimental to the group (of which he is a part), for example by depleting fundamental natural resources or polluting the environment[36].

8. *Concluding remarks*

Bioethics initially took the lead in addressing the existential risks confronting our species, as exemplified by the contributions of Potter and Jonas. These thinkers emphasized the urgency of formulating new principles and adopting novel perspectives to safeguard the survival of humanity. While it is valid to scrutinize certain aspects of Potterian and Jonasian bioethics, particularly regarding topics such as end-of-life choices, their works undeniably represent a groundbreaking milestone in contemporary philosophy. They introduced a well-structured examination of the grave threat of sapiens' extinction, marking a pivotal moment in intellectual discourse.

In this analysis, our primary objective was to highlight the significant impact of bioethical reflection in shedding light on the reasons behind humanity's inability to effectively address the challenges it faces. While delving deeply into this complex issue is beyond the scope of our current exploration, it is crucial to acknowledge the thought-provoking debates that have emerged from bioethical contemplation in recent years. Among these, two

[36] Buchanan and Powell present a different claim about our alleged moral inability to cope with certain scenarios. Indeed, they believe that exclusivism, that is, the moral attitude to care *exclusively* for the near and dear depends on the stimuli received from the environment and the specific way our *moral brain* has been trained and accustomed, not on a somehow inherent failure rooted in our evolutionary history (see A. Buchanan, R. Powell, *The evolution of moral progress. A biocultural theory*, OUP, New York 2018).

particularly intriguing discussions have revolved around the potential existential threat posed by AI[37] and the ethical considerations surrounding human enhancement[38].

However, our focus in this contribution has been to bring attention to an often overlooked aspect within the discourse on the ethics of extinction: the profound issue of the motivations that impede our ability to restructure our moral behavior. By exploring this aspect, we hoped to shed light on the underlying reasons that hinder the necessary transformation of our ethical conduct.

[37] See N. Bostrom, *Superintelligence. Paths, dangers, strategies*, OUP, Oxford 2014.

[38] See A. Clarke, J. Savulescu, C.A.J. Coady *et al.*, *The Ethics of human enhancement*, OUP, Oxford 2016.