Safeguarding the Far Future: A Broad Legal Intervention for Existential Risk Mitigation in Response to the LHC Controversy

Christopher Arvidsson

Doctor of Philosophy (PhD) Thesis

City, University of London
The City Law School

June 2023

DECLARATION

I confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

ABSTRACT

Legal interventions can assist in mitigating existential risks ('x-risks') traceable to anthropogenic processes, and examining events from the past can plausibly assist in the configuration of such interventions. To that effect, this thesis considers how a controversy from the past involving particle accelerators and the alleged risk of humanity's premature extinction (the 'LHC Controversy') can be used as a drafting guide for the law's response to certain risks within the x-risk landscape.

The research aim is to examine how the LHC Controversy can inspire the design of a broad legal intervention for x-risk mitigation. Exploring what can be learnt from the LHC Controversy in this way enables the thesis to contribute to the field of x-risk mitigation. Addressing the main aim involves an investigation of the x-risk landscape and different legal and non-legal aspects characterising the LHC Controversy.

It is argued that the LHC Controversy reveals the value of a broad legal intervention which targets the reliability of scientific work assessing a relevant x-risk. The legal mechanism advanced in this thesis ('the LHC-inspired intervention') is disconnected from probabilities of risk and the adjudication of competing scientific theories.

The LHC-inspired intervention entails evaluation of interconnected deficiency factors which can give credence to a concern that the underlying x-risk assessment is not, at present, as reliable as it should be. Broadly speaking, these deficiency factors require analysis of (i) the humans and organisations who consider the relevant x-risk and (ii) the possibility that they have based their conclusion(s) on flawed information or an incomplete state of knowledge or understanding.

TABLE OF CONTENTS

TABLE OF CASES

UNITED STATES

UNITED KINGDOM

NETHERLANDS

TABLE OF STATUTES AND BILLS

TABLE OF TREATIES AND CONVENTIONS

TABLE OF OFFICIAL PAPERS AND POLICY DOCUMENTS

-	United Nations, 'Our Common Agenda – Report of the Secretary-General' (2021)	*18-19, 272-274*
-	Wellbeing of Future Generations Bill [HL] (Explanatory Notes) <https://bills.parliament.uk/publications/41931/documents/417> accessed 27 April 2023	*91*

1 CHAPTER ONE: INTRODUCTION

> At a handful of moments in the past century, a few rare groups of people have held the world's fate in their hands, responsible for the tiny-but-real possibility of causing total catastrophe. Not just the end of their own lives, but the end of everything.[1]

1.1 BACKGROUND

Some risks are more serious than others. Some threaten to cause physical harm, and others to kill. A few may affect a single person, and others may affect hundreds or thousands of individuals. Some may impact millions, with the adverse effects stretching beyond national borders. Then there are the risks of pivotal seriousness; those that may transcend the *present* by threatening to destroy humanity's future. A risk of such potency is called an existential risk, and it has astronomically high stakes. While it may sound like sci-fi to those unfamiliar with the literature, existential risk is a topic of growing concern amongst academics and policymakers.

The precariousness of humanity's state of existence on this "pale blue dot"[2] is not a new theme. From non-human superintelligence, nuclear warfare and disastrous science experiments to asteroid impacts and the sun's death, scholars have long acknowledged that humanity's continued survival is not guaranteed. To safeguard the far future and humanity's potential, policymakers should do more to facilitate x-risk mitigation. In doing so, looking to the past can be indicative of future mitigation strategies.

This thesis will consider a controversy which involved particle accelerators and the alleged risk of humanity's premature extinction. Indeed, the thesis centres around considering how this past controversy—referred to as the LHC Controversy[3]—can serve as a drafting guide for the law's response to certain x-risks.

---

[1] Richard Fisher, 'The Moments that Could have Accidentally Ended Humanity' *BBC* (19 February 2021) <https://www.bbc.com/future/article/20210217-the-moments-that-we-could-have-destroyed-humanity> accessed 20 April 2023.

[2] This is inspired by the title of Sagan's book. Carl Sagan, *Pale Blue Dot: A Vision of the Human Future in Space* (Random House Publishing Group 1994).

[3] Section 1.2.1 introduces this controversy in further detail.

1.1.1 Aims of Chapter 1

This Chapter introduces the thesis and what it aims to achieve. To this effect, Chapter 1 includes a research overview, the research objective, questions, and parameters. The Chapter also outlines the main finding and argument of the thesis. In addition, it highlights how the thesis contributes to knowledge. Other elements include a Part on methodology and an overview of the remaining structure of the thesis.

1.2 RESEARCH OVERVIEW

This thesis explores how the *Large Hadron Collider Controversy* can inspire the design of a *broad legal intervention* for the mitigation of *existential risk*. It is possible to clarify this statement by introducing (i) the LHC Controversy, (ii) the concept of existential risk, and (iii) broad legal interventions for the mitigation of this risk category. Although this is done in turn below, the subsequent chapters will expand upon these concepts in more detail. In addition, Section 1.2.4 introduces the perspective of *longtermism*, which sits in the background of the thesis.

In reading the sections below, it is helpful to keep in mind that the thesis's overarching aim is to consider how the LHC controversy can serve as a drafting guide for the law's response to certain future risks within the x-risk landscape.[4]

1.2.1 The Large Hadron Collider Controversy

This Section defines the Large Hadron Collider Controversy. It is important to deal with this point early on because this Controversy is used as inspiration for the broad legal intervention advanced in this thesis. In other words, it is this Controversy—and the accompanying legal and x-risk literature—that will serve as a drafting guide for how the law might respond to certain future x-risks.

[4] Section 1.3.3 outlines the thesis's research parameters. In effect, it will shine light on the kind of x-risks relevant for the LHC-inspired intervention.

The Large Hadron Collider (the 'LHC') is an underground particle accelerator near Geneva, and physicists use it to pursue a deeper understanding of the universe.[5] While being one of the most ambitious endeavours ever, critics claimed that the machine's operation could destroy the Earth by, among other stipulated disaster scenarios, creating a black hole.[6] Essentially, it was suggested that the LHC posed – what this thesis and other x-risk scholars call – an existential risk.[7] Some even attempted to halt the LHC by taking their concerns to courts around the world.[8]

In contrast, virtually every proponent of the LHC argued the machine was safe, with one physicist saying: "Humanity is safe from the LHC…but is the LHC safe from humanity?"[9] The LHC is neither the only accelerator nor the first accelerator coming under fire as a potential doomsday machine. Another example (and past subject of litigation) is the Relativistic Heavy Ion Collider (RHIC) at the Brookhaven National Laboratory (BNL) in the US.[10]

It is perplexing for an external party – such as a court or regulator – to address the disagreement and evaluate whether particle accelerators truly pose an x-risk. There are several 'review barriers' capable of frustrating different modes of review.[11] For instance, the reviewer may be invited to validate science only comprehensible by a small community of experts. The external party is then, most likely, at the mercy of the same people it is asked to control. This barrier will disrupt a mode of external review where the scientific merits can reasonably dominate the dispute. As another example, the probability of the risk might be unquantified and resistant to quantification. This barrier can disrupt a review mode where the external party can consider the risk of harm or use cost-benefit analysis. The external reviewer may then be compelled to dismiss the claim or application of worried opponents as too speculative.

---

[5] CERN, 'The Large Hadron Collider'<https://home.cern/science/accelerators/large-hadron-collider> accessed 20 April 2023; CERN, 'Facts and Figures About the LHC' <https://home.cern/resources/faqs/facts-and-figures-about-lhc> accessed 20 April 2023.

[6] See Chapter 3.

[7] Section 1.2.2 introduces x-risk. This is taken further in Chapter 2 which considers the x-risk landscape.

[8] See Chapter 3.

[9] Johnathan Ellis, 'The LHC is Safe' (*CERN Colloquium*, 14 August 2008, [00:44:11]) <https://cdsweb.cern.ch/record/1120625> accessed 24 April 2023.

[10] See Chapter 3.

[11] See Chapter 3.

With the above in mind, the thesis uses the 'LHC Controversy' as a collective term to mean the disagreements and uncertainty about whether high-energy physics experiments facilitated by particle accelerators (i) pose any existential risk to humanity, (ii) pose an unacceptable existential risk to humanity, or (iii) should be delayed or suspended. The thesis uses the term 'LHC' rather than, for instance, the 'Particle Accelerator' Controversy because Chapter 3 primarily refers to the LHC. However, as mentioned above, the LHC is not the first and only accelerator associated with end-of the-world-scenarios.

Significantly, the value of this thesis is not contingent on the correctness of the assertion that running the LHC, the RHIC, or any other accelerator can cause an x-catastrophe. In other words, whether the existential catastrophes associated with high-energy physics experiments are physically possible or not is immaterial in the context of this thesis. The point is that the LHC Controversy can provide valuable lessons in terms of x-risk mitigation regardless of whether the risk is zero. Similarly, Bostrom and Ćirković suggest it can be meaningful to analyse "…scenarios which are almost certainly physically impossible. The hypothetical risk from particle collider experiments is a case in point."[12] Recognising the "…objective risk is probably zero, as believed by most experts…"[13] they also highlight the possibility that (i) the safety argument provided by the experts might be flawed in a yet unrealised way, (ii) the experts might be biased, and (iii) the experts might be part of the same community that wants the experiment to go ahead.[14]

1.2.2 Existential Risk

A preliminary understanding of x-risk is necessary to comprehend the essence of the LHC Controversy and the idea of a broad legal intervention aiming to mitigate this risk category. It is nearly impossible to write about existential risk without referring to Oxford Professor Nick Bostrom. He defines an x-risk as a risk "…that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic

[12] Nick Bostrom and Milan Ćirković, 'Introduction' in Nick Bostrom and Milan Ćirković (eds), *Global Catastrophic Risks* (OUP 2008) 5.
[13] ibid.
[14] ibid 19.

destruction of its potential for desirable future development."[15] Bostrom's definition is rather wide. In effect, this thesis defines an x-risk as one that threatens the premature extinction of humanity. Chapter 2 will deal with this and other relevant definitions in further detail.

For the sake of exposition, the most obvious example of an x-catastrophe is the literal destruction of Earth. Such a tragedy will currently qualify as an existential catastrophe because it will almost certainly result in humanity's premature extinction.

An existential risk threatens to destroy humanity's past, present and future.[16] It is the threatened destruction of the future that makes this risk category uniquely serious and the associated stakes astronomically high. On this note, the thesis will introduce a concept referred to as the 'astronomical stakes proposition' ('the ASP').[17] The ASP encapsulates the notions that (i) existential risks have astronomically high stakes, and (ii) existential catastrophes are uniquely bad relative to *all* non-existential catastrophes.[18] This recognition, if accepted, is significant for how policymakers should deal with risks threatening to destroy humanity's future. Indeed, the ASP will be utilised to empower the LHC-inspired legal intervention.

Existential risk is speculatively associated with hazards including artificial general intelligence, geoengineering, biotechnology and scientific endeavours such as experimental particle physics.[19] The very first anthropogenic x-risk arguably came with

[15] Nick Bostrom, 'Existential Risk Prevention as Global Priority' [2013] Global Policy 15, 15; Nick Bostrom, 'Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards' (2002) 9 Journal of Evolution and Technology. The definition in his 2002 paper is formulated differently where an existential risk is described as one 'where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential.' The 2013 article has replaced many themes from the 2002 article.

[16] Reference to humanity's past, present and future is borrowed from William Leiss as well as Toby Ord. See William Leiss, *The Doom Loop in the Financial Sector, and Other Black Holes of Risk* (University of Ottawa Press 2010) 17; Toby Ord, *The Precipice* (Bloomsbury Publishing 2020) 35.

[17] See Chapter 5. The phrase 'astronomical stakes proposition' is inspired by Bostrom. Nick Bostrom, 'Astronomical Waste: The Opportunity Cost of Delayed Technological Development' [2003] Utilitas 308; See also the "astronomical value thesis" in Phil Torres, *Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks* (Pitchstone Publishing 2017) 41.

[18] See Chapter 5.

[19] See Chapter 2.

the Manhattan Project and the detonation of the first A-bomb.[20] One may think that the risk referred to is the possibility of nuclear warfare. While such a scenario is relevant to the study of x-risk,[21] the risk alluded to here is the old speculation that a single nuclear explosion could ignite the atmosphere and end humanity.[22] On that note, a now declassified scientific report ('LA-602')[23] dealing with this possibility may represent the first risk assessment of human extinction.[24]

Efforts to understand existential risk have grown over the years.[25] For instance, Schneier has compiled a list with over 120 resources dealing with x-risk from various perspectives.[26] Furthermore, established in 2005, a core aspect of the Future of Humanity Institute (FHI) at Oxford University entails studying "…events that endanger the survival of Earth-originating, intelligent life or that threaten to drastically and permanently destroy our potential for realising a valuable future."[27] This formulation closely mirrors Bostrom's seminal definition, which is not surprising given that he is the Director of the FHI.[28] Another notable institution is the Centre for the Study of Existential Risk (CSER) at Cambridge University. The CSER works to understand extreme risks associated with emerging technologies and human activity.[29]

---

[20] See Chapter 6.

[21] See Chapter 2.

[22] Leiss, *The Doom Loop in the Financial Sector, and Other Black Holes of Risk* (n 16) 25-28; Ord, *Precipice* (n 16) 90-94; Olle Häggström, *Here Be Dragons: Science, Technology and the Future of Humanity* (OUP 2016) 198-199; Seth Baum, 'The Great Downside Dilemma for Risky Emerging Technologies' (2014) 89(12) Physica Scripta <https://iopscience.iop.org/article/10.1088/0031-8949/89/12/128004> accessed 20 April 2023; James Miller and Debbie Felton, 'The Fermi Paradox, Bayes' Rule, and Existential Risk Management' [2017] Futures 44.

[23] Emil Konopinski, Cloyd Marvin, and Edward Teller, 'Ignition of the Atmosphere with Nuclear Bombs' (1946) <https://sgp.fas.org/othergov/doe/lanl/docs1/00329010.pdf> accessed 20 April 2023.

[24] Bostrom, 'Existential Risk Prevention' (n 15) 29 (note 35); Eliezer Yudkowsky, 'LA-602 vs. RHIC Review' (*LessWrong*, 19 June 2008) <https://www.lesswrong.com/posts/f3W7QbLBA2B7hk84y/la-602-vs-rhic-review> accessed 20 April 2023.

[25] This is recognised by Alexey Turchin and David Denkenberger, 'Global Catastrophic and Existential Risks Communication Scale' [2018] Futures 27, 27.

[26] Bruce Schneier, 'Resources on Existential Risk' (Catastrophic Risk: Technologies and Policies Berkman Center for Internet and Society) <https://futureoflife.org/data/documents/Existential%20Risk%20Resources%20(2015-08-24).pdf> accessed 20 April 2023.

[27] Future of Humanity Institute, 'Research Areas' <https://www.fhi.ox.ac.uk/research/research-areas/> accessed 20 April 2023.

[28] Future of Humanity Institute, 'Team' <https://www.fhi.ox.ac.uk/the-team/> accessed 20 April 2023.

[29] Centre for the Study of Existential Risk, 'About us' <https://www.cser.ac.uk/about-us/> accessed 20 April 2023.

Workshops and academic courses have also taken place beyond Oxbridge in Scandinavia and the US.[30] The Legal Priorities Project, established in 2020 by researchers at Harvard University, also undertake legal research that "…mitigates existential risk and promotes the flourishing of future generations."[31]

Beyond the academic crowd, the topic has reached mainstream media[32] and policymakers. For instance, the Chief Scientific Adviser's annual report in 2014 included a chapter entitled *Managing Existential Risk from Emerging Technologies*.[33] Additionally, *Our Common Agenda*, published by the United Nations in 2021, acknowledges x-risk as deserving more attention.[34]

Despite the growing interest, the area of existential risk remains relatively understudied. As written by Torres, the "…subfield of 'existential risk mitigation strategies' is woefully underdeveloped, and consequently there aren't many shoulders upon which to stand."[35] Taking this further, the sub-subfield of *legal* strategies and mechanisms for x-risk mitigation is even less developed.[36] It is within this sub-subfield that this thesis finds its relevance. As mentioned, the thesis will consider the LHC Controversy from a legal perspective and advance a legal intervention that may mitigate certain future x-risk. In this way, the LHC controversy is used a rough drafting guide for x-risk mitigation.

[30] Adrian Currie and Seán Ó hÉigeartaigh, 'Working Together to face Humanity's Greatest Threats: Introduction to the Future of Research on Catastrophic and Existential Risk' [2018] Futures 1, 1-2.
[31] Legal Priorities Project, <https://www.legalpriorities.org/> accessed 20 April 2023.
[32] Simon Beard and Lauren Holt, 'What are the Biggest Threats to Humanity?' *BBC* (15 February 2019) <https://www.bbc.co.uk/news/world-47030233> accessed 20 April 2023; Furthermore, The *Deep Civilization* series by the BBC covers many aspects relevant to the future of humanity. See <https://www.bbc.com/future/columns/deep-civilisation> accessed 20 April 2023.
[33] The Government Office for Science, 'Innovation: Managing Risk, Not Avoiding It' (Annual Report of the Government Chief Scientific Adviser 2014) chp 10.
[34] United Nations, 'Our Common Agenda – Report of the Secretary-General' (2021) 65.
[35] Phil Torres, *The End: What Science and Religion Tell Us about the Apocalypse* (Pitchstone Publishing 2016) 228.
[36] Nevertheless, some legal works have contributed to this context.

1.2.3 Broad Legal Interventions

The general concept of a legal intervention should be clarified before distinguishing between *broad* and *targeted* interventions. This thesis defines a legal intervention as any legal mechanism, tool or approach seeking to reduce the likelihood that an x-risk materialises.[37] The focus is on preventing (or reducing the probability) that an x-catastrophe occurs at all. In contrast, a legal intervention can also aim to reduce the probability that a materialised catastrophe becomes severe enough to cause the extinction of humanity.

A legal intervention may reduce the likelihood of x-risk in different ways. For example, it may seek to elevate and protect the status of future generations. Another such intervention may centre around empowering a new or existing body at the UN to deal with different existential threats. An intervention may, alternatively, regulate individual (or associated) hazards linked to x-risk through a novel or revamped international treaty. Criminalising behaviour giving rise to or amplifies an x-risk via "error or terror"[38] is another legal mechanism that may reduce the probability of x-catastrophe. As a final example at this stage, a legal intervention may "…place long-term analysis, planning and thinking at the heart of national governance and the multilateral system…[and] expand our thinking and institutions across time."[39]

Having clarified the notion of a legal intervention, the next aspect to consider is the distinction between a broad and targeted legal intervention. This distinction matters since this thesis explores how the LHC Controversy can inspire the design of a **broad** intervention for x-risk mitigation.

[37] Cotton-Barratt, Daniel and Sandberg distinguish between mitigation strategies reducing the likelihood that (i) a catastrophe materialises in the first place, (ii) a materialised catastrophe becomes a sever global catastrophe, and (iii) humanity goes extinct following a severe global catastrophe. Owen Cotton-Barratt, Max Daniel and Anders Sandberg, 'diplomacypth Against Human Extinction: Prevention, Response, Resilience, and Why They All Matter' [2020] Global Policy 271, 272.

[38] This expression is borrowed from Rees who writes "…we're entering an era when a few individuals could, via error or terror, trigger societal breakdown." Martin Rees, 'Denial of Catastrophic Risks' (2013) 339(6124) Science <https://www.science.org/doi/10.1126/science.1236756> accessed 20 April 2023.

[39] UN, 'Our Common Agenda' (n 34) 45.

Put simply, broad interventions are designed to reduce the likelihood of x-risk from two or more hazards. An example may include the creation of science courts dealing with x-risk from different technologies and scientific endeavours. In contrast, a targeted legal intervention is hazard-specific and designed to mitigate x-risk from one hazard. An example would be a treaty dealing with experimental particle physics controlling, for instance, the acceptable energies of the collisions.

As mentioned above, this thesis concentrates on formulating a broad legal mechanism. Significantly, this thesis applies three conditions that the intervention should satisfy.

1. The LHC-inspired intervention should respond to the 'review barriers'[40] that may complicate different modes of external review in the LHC Controversy context. These barriers are examined in Chapter 3.

2. The broad legal intervention should utilise the ASP.[41] Chapter 5 deals with the ASP.

3. The legal mechanisms should, with some degree of plausibility, be able to find utility beyond the LHC Controversy. Chapter 2 surveys the x-risk landscape. In addition, Chapter 6 outlines examples which indicate that the LHC-inspired intervention may apply beyond the context of the LHC Controversy.

These conditions are used for different reasons. The first one is applied because the review barriers form a vital part of what can be learnt from the LHC Controversy. In effect, if the intervention cannot apply to the LHC Controversy, its theoretical foundation and future operation deserves little faith. The second condition is imposed because this thesis subscribes to the belief that safeguarding humanity's future is overwhelmingly important.[42] It views extinction as *uniquely bad* and suggests this recognition should underpin legal interventions for x-risk mitigation. The third condition is applied because the thesis concerns a *broad* intervention. This condition is

[40] These barriers were introduced in Section 1.2.1.
[41] The astronomical stakes proposition was introduced in Section 1.2.2.
[42] Chapter 5 will consider this further.

axiomatic given the aim to advance a mechanism that may reduce the likelihood of x-risk from different hazards.

Notably, the thesis will not formalise the legal intervention by outlining the configuration of a hypothetical treaty or statute. Similarly, it is not the aim to consider the institutional design within which the intervention may best function. Instead, what matters are the theoretical building blocks, the overarching function and the logic behind the intervention that draws inspiration from the LHC Controversy. That being said, Chapter 6 will briefly consider the institutional design and propose that the intervention might function best within some kind of agency tasked with x-risk mitigation.

The sections above are relevant for understanding this thesis's research objective, questions and parameters. These aspects are dealt with in more detail in Part 1.3. Before that, Section 1.2.4 outlines the view of longtermism which underpins much of this thesis.

1.2.4 Philosophical and Legal Longtermism

Whereas the thesis will not often refer expressly to longtermism going forward, this stance underlies the thesis's spirit and decision to focus on *existential* risk mitigation. This Section will clarify why longtermism sits in the background of this thesis.

The term was coined by Oxford scholars Toby Ord and William MacAskill, and the perspective of longtermism intertwines with ideas popular within the x-risk circle.[43] As summarised by MacAskill:

> Longtermism is the view that positively influencing the long-term future is a key moral priority of our time. It's about taking seriously the sheer scale of the future, and how high the stakes might be in shaping it. It means thinking about the challenges we might face in our lifetimes that could impact civilisation's whole trajectory, and taking action to benefit not just the present generation, but all generations to come.[44]

[43] Ord, *Precipice* (n 16) 306; William MacAskill, *What We Owe The Future: A Million-Year View* (Oneworld Publications 2022).
[44] William MacAskill, 'What is longtermism?' *BBC Futures* (8 August 2022).

Ord explains that longtermism "…takes seriously the fact that our own generation is but one page in a much longer story, and that our most important role may be how we shape—or fail to shape—that story."[45] In essence, longtermism centres around the perception that it is overwhelmingly important to ensure that the far future goes well.[46]

An x-risk is particularly troublesome because it threatens to destroy humanity's future and, in effect, foreclose humanity's opportunity to shape the future for the better. One may appreciate that x-risk mitigation is one – but not the only – way to positively influence the future. As written by Ord, "…a longtermist ethic is…especially well suited to grappling with existential risk. For longtermism is animated by a moral re-orientation towards the vast future that existential risks threaten to foreclose."[47]

Whereas this thesis subscribes to longtermism, there are various objections to this view, including that it is not possible to influence the future, that the future matters less than the present,[48] and that the future may, all things considered, contain more misery than positive things. Chapter 5 examines themes relating to and supporting the view of longtermism.

Scholars have used the premises associated with philosophical longtermism to consider the value of legal mechanisms.[49] Legal longtermism refers to "[t]he set of views associated with the claim that law and legal institutions ought to protect those in the far future…"[50] On that note, Martínez and Winter suggest that the legal system

<https://www.bbc.com/future/article/20220805-what-is-longtermism-and-why-does-it-matter> accessed 20 April 2023.

[45] Ord, *Precipice* (n 16) 46.

[46] Eric Martínez and Christoph Winter, 'Foundations of Experimental Longtermist Jurisprudence' in Stefan Magen and Karolina Prochownik (eds), *Advances in Experimental Philosophy of Law* (forthcoming).

[47] Ord, *Precipice* (n 16) 46.

[48] Legal Priorities Project, 'Legal Priorities Research: A Research Agenda' (2021) <https://www.legalpriorities.org/research_agenda.pdf> accessed 20 April 2023. See pages 17-22 for an overview to some objections to longtermism.

[49] Martínez and Winter, 'Foundations of Experimental Longtermist Jurisprudence' in *Advances in Experimental Philosophy of Law* (n 46).

[50] ibid.

can protect future generations and the far future: "…[L]egal interventions could play a significant role in mitigating…" inter alia x-risk.[51]

1.2.5 Summary

The opening sentence of Part 1.2 explained that this thesis concerns how the LHC Controversy can inspire the design of a broad legal intervention for x-risk mitigation. The sections above unpacked this statement. As will be displayed below, this examination has resulted in a finding that the LHC Controversy displays the value of a broad legal mechanism focusing on the underlying reliability of x-risk assessments.

1.3 RESEARCH OBJECTIVE, QUESTIONS AND PARAMETERS

1.3.1 Research Objective

The main research objective is to examine how the LHC Controversy can inspire the design of a broad legal intervention for the mitigation of existential risk. In effect, the LHC Controversy is used as a rough drafting guide for the law's response to some, though not all, risks within the x-risk landscape.

It is useful to highlight that the thesis considers the law's response in general. The LHC-inspired intervention can, in theory, be applied within different legal systems. However, references will be made to the English and US legal systems. Reference will also be made to international law. This is because legal scholars considering the LHC Controversy, x-risk or human extinction tend to approach these topics from the perspective of one of these legal regimes.

1.3.2 Research Questions

The main objective above can be approached by asking the correlative question: How can the LHC Controversy inspire the design of a broad legal intervention to mitigate

[51] ibid.

existential risk? It is helpful to outline some sub-questions to facilitate an answer to the central question. In effect, the thesis will consider the following sub-questions:

1. What makes existential risk unique, and why is it difficult to study and mitigate this risk category? This question is addressed in Chapter 2.

2. What makes it complicated for an external party to address the disagreements and uncertainty about whether particle accelerator experiments pose an x-risk? While addressed throughout the thesis, Chapter 3 is a cornerstone in considering this sub-question.

3. What is the justification for the ASP, and what is the relevance of this concept when it comes to legal interventions for x-risk mitigation? Chapter 5 is devoted to these questions.

4. How does the LHC-inspired intervention (i) operate in theory, (ii) align with existential risk studies and, (iii) utilise the astronomical stakes proposition? Chapter 6 will draw on the previous chapters and answer this sub-question.

1.3.3 Research Parameters

This Section outlines the research parameters of the thesis. It is helpful to start with a note about the width of the field of existential risk. This field of study can be approached from many perspectives, some of which are, to a lesser or greater extent, isolated from legal considerations. For instance, it is possible to explore (i) whether it is morally worthwhile to reduce this risk category, (ii) how to reduce specific or connected risks within this category, (iii) their sources and the science behind them, and (vi) their history.

As mentioned before, this thesis falls within the sub-field of x-risk mitigation, and it approaches this matter from a legal perspective. It is also possible to deal with this sub-field from non-legal perspectives.[52] The process of expanding humanity's

[52] This is explored in Chapter 5.

presence in space is one strategy that can largely be viewed as a non-legal intervention tackling x-risk by reducing the likelihood that a materialised catastrophe on Earth leads to humanity's premature extinction.

In short, the focus of this thesis is solidified by dealing with the sub-field of x-risk mitigation from a legal perspective. However, the focus can be sharpened further by working within additional research parameters. The point is that the sub-subfield of law and x-risk mitigation remains wide. For example, one can consider the regulation and governance of any given *hazard* – such as AGI, geoengineering, biotechnology and asteroid impacts – linked to x-risk. It is also possible to consider whether a legal mechanism can address *multiple* risks sharing common features.[53] As another example, it is possible to investigate a specific organisation and its ability to manage x-risk from different or specific hazards.[54] Given the residual width, additional research parameters have been employed to make the thesis more streamlined. These will be outlined momentarily. A number of these parameters have been selected on account of the LHC Controversy. Since the aim is to use the LHC Controversy as a drafting guide, it is practical to focus on x-risks that share certain commonalities with the postulated risk(s) in the LHC Controversy context. In particular, the following research parameters should be noted:

(i) No Universal Mitigation Strategy

The thesis does not aim to advance a broad legal intervention that mitigates x-risk from all hazards associated with the x-risk landscape. Due to the nature of existential risk, such an aim would possibly be too wide. The parameters below will further highlight the type of existential risks falling outside the scope of the thesis.

[53] Luke Kemp and Catherine Rhodes, 'The Cartography of Global Catastrophic Governance' (Report of Centre for the Study of Existential Risk 2020) 31.
[54] Reinmar Nindler, 'The United Nation's Capability to Manage Existential Risks with a Focus on Artificial Intelligence' [2019] International Community Law Review 5. Nindler explores whether the UN can address x-risks associated with different forms of AI.

(ii) Anthropogenic Risk > Natural Risk

A distinction is often drawn between human-made (anthropogenic) and natural x-risks.[55] The LHC-inspired intervention is concerned with human-made risks. However, drawing an absolute distinction between anthropogenic and natural x-risks is not always easy. For example, bearing in mind the interconnectedness of the human species, the power to spread (mis)information around the world within seconds, and the ability to learn from tragedies like the Spanish flu and Covid-19, one may wonder whether the threat of pandemics should be classified as a natural or human-made risk. Perhaps a more compromising answer is to view the risk as a collaboration between nature and humanity. Despite this observation, the distinction is arguably more apparent in other contexts, such as in relation to the LHC Controversy.

(iii) Error Risk > Terror Risk

The thesis concerns anthropogenic risks deriving from 'error' rather than 'terror'. The distinguishing quality between error and terror risks is arguably the state of mind of the actors creating the risk.[56] In effect, the LHC-inspired intervention does not aim to reduce the likelihood of x-risk from actors who *maliciously* and *intentionally* want to bring about humanity's premature end.

(iv) Bangs > Cascading Failures

This thesis focuses on events which will cause a *sudden* existential catastrophe should the alleged risk materialise. Bostrom outlines four categories of x-risk.[57] Under one of these, "…Earth-originating intelligent life goes extinct in relatively sudden disaster resulting from either an accident or a deliberate act of destruction."[58] Bostrom calls this category *Bangs*, and under it, one can discuss the risk associated with, among other things, the Manhattan Project and experimental particle physics.[59] Bostrom's

[55] This distinction is examined in Chapter 2.
[56] Cotton-Barratt, Daniel and Sandberg, 'Defence' (n 37) 272-274.
[57] Bostrom, 'Existential Risks' (n 15); He uses a different classification scheme in Bostrom, 'Existential Risk Prevention' (n 15) 19.
[58] Bostrom, 'Existential Risks' (n 15).
[59] ibid (n 15).

categorisation captures the thesis's focus on 'knock out events'[60] rather than cascading failures. However, it is to be appreciated that there are many "…slower and more intertwined ways in which the world might collapse, without being hit by spectacular hazards."[61]

(v) Human Extinction > Other Classes of Risk

Some x-risk classes do not involve the threat of extinction. Bostrom identifies the class of *permanent stagnation*, *flawed realisation,* and *subsequent ruination*.[62] The key point is that x-risk mitigation needs not merely be about humanity's physical survival. Despite this recognition, the thesis only deals with the class of *human extinction*,[63]

---

[60] Term borrowed from Currie and Ó hÉigeartaigh, 'Working Together to face Humanity's Greatest Threats' (n 30).

[61] Hin-Yan Liu, Kristian Cedervall Lauta and Matthijs Michiel Maas, 'Governing Boring Apocalypses: A New Typology of Existential Vulnerabilities and Exposures for Existential Risk Research' [2018] Futures 6, 10; Similarly, in providing an overview of certain x-risks, Farquhar and others write that less severe catastrophes might bring about extinction indirectly and over a longer timeframe. Sebastian Farquhar and others, 'Existential Risk: Diplomacy and Governance' (Global Priorities Project 2017) <https://www.fhi.ox.ac.uk/wp-content/uploads/Existential-Risks-2017-01-23.pdf> accessed 24 April 2023, 6; As further clarified by Matus: "Existential risks arise not merely from one-off large incidents, such as earthquakes, tsunamis, nuclear meltdowns or, indeed, asteroid hits. Rather, existential risks are about complex, inter-related processes that result in cascading effects that move across social systems." Kira Matus, 'Existential Risk: Challenges for Risk Regulation' (2014) <http://www.lse.ac.uk/accounting/assets/CARR/documents/R-R/2014-Winter/Risk-and-Regulation-28-existential-risk.pdf> accessed 24 April 2023; Currie and Ó hÉigeartaigh examine how research often examines distinct 'knock out' events. Examples of such an approach, they write, include looking at asteroid impacts or the potential consequences of physics experiments. Focusing on knock outs can potentially miss important sources of existential catastrophe. Yet, they note: "…[C]onsideration of different types of risk needn't be a zero-sum game. A thriving ecosystem of existential and catastrophic risk research can encompass slow-moving threats and cascading failures, as well as abrupt catastrophic events." Currie and Ó hÉigeartaigh, 'Working Together to face Humanity's Greatest Threats' (n 30) 4.

[62] Consider *permanent stagnation*. This x-risk class means humanity survives but fails to reach 'technological maturity'. By technological maturity, Bostrom refers to the "attainment of capabilities affording a level of economic productivity and control over nature close to the maximum that could feasibly be achieved." For example, a technologically mature civilisation could possibly colonise space and enhance human biology. Thus, destroying the opportunity to reach such maturity is an enormous loss of value. Bostrom explores scenarios of permanent stagnation. One such scenario is *plateauing* where progress flattens at a level far below technological maturity. It is possible to question the plausibility of *plateauing* due to the modern trend of rapid social and technological change. Nevertheless, Bostrom discusses a specific example which involves a global regime blocking all technological change. Such a scenario can contribute to *plateauing*. See Bostrom, 'Existential Risk Prevention' (n 15) 19-22.

[63] This class means *humanity* goes extinct before reaching *technological maturity*. ibid.

and the LHC-inspired intervention is designed to prevent or reduce the likelihood of a risk threatening the premature extinction of humanity.

It is helpful to summarise the research parameters outlined so far. In essence, the thesis focuses on *anthropogenic* risks originating from human *error*, which, if they materialise, will cause a *sudden* catastrophic event leading to the premature *extinction* of humanity. As mentioned, these parameters are used because the LHC Controversy involved postulated x-risk(s) sharing these hallmarks.

Beyond the parameters above, the thesis imposes two further research limits that pertain to the legal intervention itself. These are outlined next.

(vi) Theoretical Foundation > Practical Implementation

The thesis focuses on the intervention's theoretical building blocks, logic, and function instead of its practical implementation. Despite this focus, Chapter 6 discusses some implementation routes and the institutional framework within which the intervention may be employed.

(vii) Prevention > Reaction or Resistance

As already mentioned, the legal mechanism advanced in response to the LHC Controversy aims to prevent or reduce the likelihood of x-risk. The focus on prevention limits the scope of the thesis because it limits the type of interventions that can be advanced. For instance, it means that the thesis will not propose an intervention which attempts to lessen the impact of an already materialised catastrophe. Given the nature of the LHC Controversy, where the alleged catastrophes would be sudden and recovery virtually impossible, aiming to reduce the likelihood of the catastrophes is the obvious approach for an intervention inspired by the Controversy.

1.4 MAIN ARGUMENTS

The thesis has now clarified some threshold concepts, including the definition of the LHC Controversy, x-risk, and broad legal interventions. These concepts are important because they clarify what this thesis is about and what it aims to achieve.

In addition, it was explained that the main research objective and question concerns how the LHC Controversy can inspire the design of a broad legal intervention to mitigate certain risks within the x-risk landscape. With these aspects in mind, this Part will summarise the main arguments of the thesis.

Argument One

The first argument is simply that the LHC Controversy can be used as a drafting guide for a broad legal intervention for x-risk mitigation. It is argued that the past can inspire future legal mechanisms aiming to safeguard humanity from certain risks within the x-risk landscape.

Argument Two

The second argument is that the LHC Controversy shows the value of a legal intervention which focuses on the underlying *reliability* of scientific work assessing existential risk(s) associated with certain endeavours. More specifically, the thesis argues for a broad legal intervention giving an external reviewer[64] a principled basis upon which to base a finding that some endeavours should not *presently* be allowed to go ahead since, and notwithstanding the review barriers,[65] there are reasonable doubts as to the *reliability* of the scientific work assessing an existential risk linked to these endeavours.

[64] An *external reviewer* herein means any (independent) third party tasked with considering the underlying reliability of the x-risk assessment.
[65] Section 1.2.1 introduced these barriers, and they are further examined in Chapter 3.

It is crucial to note that the LHC-inspired intervention does not concern the scientific merits of the relevant safety assessment. In other words, the intervention does not involve the external reviewer validating conflicting scientific theories about the safety of any given endeavour. Instead, the intervention involves an analysis of interconnected deficiency factors, which can give credence to a concern that scientific work assessing existential risk(s) is not, at present, as reliable as it should be.

These deficiency factors require consideration of (i) the humans and organisations who consider the x-risk and (ii) the possibility that they have based their conclusion on flawed information or an incomplete state of knowledge or understanding.

For example, various unconscious biases may affect the reliability of the assessors' work. Likewise, the prevalence of a conflict of interest amongst the assessors can undermine the reliability of the risk assessment work. In addition, analysing the deficiency factors can give credence to an argument that the assessors' conclusion is based on flawed information or an incomplete state of knowledge. For example, the safety assessment may be based on relatively new theories or theories subject to much debate. As written in Chapter 6:

> The overall point is that the newness of the underlying scientific theory, the lack of historical robustness of relevant safety arguments, the gravity of past upsets within the field, the pace at which new ideas are introduced, and the frequency by which old ones are challenged can all give weight to assertations that the assessment excluding risk is potentially based upon defective scientific theory.

These and other aspects are dealt with in Chapter 6. The crucial point for now is that the *reliability* of the x-risk assessment can reasonably be questioned by analysing a set of deficiency factors which do not involve consideration of the accuracy of the scientific theories themselves. It is useful to repeat Bostrom and Ćirković observation that (i) relevant safety arguments might be flawed in a yet unrealised way, (ii) the risk assessors might be biased, and (iii) that they might be part of the same community that wants the relevant endeavour to go ahead.[66]

[66] Bostrom and Ćirković, 'Introduction' (n 12) 19.

Argument Three

The third argument is that the ASP can underpin legal interventions for x-risk mitigation. As mentioned earlier, the ASP encapsulates the notions that (i) x-risks have astronomically high stakes, and (ii) x-catastrophes are uniquely bad relative to all non-existential catastrophes.

Perhaps the most obvious example would involve using the ASP to justify onerous obligations being imposed on actors whose activities might give rise to x-risk. However, this is not the only way by which the ASP can underpin legal mitigation mechanisms. Chapters 5 and 6 will consider this in more detail.

In line with the third argument, the ASP underpins the LHC-inspired intervention. Indeed, this intervention, insists upon expansive thinking across time, and the ASP is the underlying tool which may justify this line of thinking. As written in Chapter 6:

> For example, a 100-year period is relatively brief when giving weight to humanity's long-term future. So too is 1,000 years. The claim that a 100-year or 1,000-year period is brief is most certainly **relative**. It is relative to the vast time scales that we *should* give weight to if we accept the ASP and the ideas that sway existential risk research and mitigation.

1.5 ORIGINALITY

It was mentioned above that the "…subfield of 'existential risk mitigation strategies' is woefully underdeveloped… ."[67] The sub-subfield of x-risk mitigation through legal interventions is even more underexplored. It is within the remits of this unchartered field that this thesis is situated.

There is practically no legal work – drawing on legal and x-risk literature – considering how the LHC Controversy can serve as a drafting guide for a broad legal intervention for x-risk mitigation. This means that a significant contribution comes from simply investigating how the LHC Controversy can inspire the design of a legal mechanism

[67] Torres, *The End* (n 35) 228.

sitting within and operating across the x-risk landscape. Whereas some legal scholars have dealt with the LHC Controversy, few have situated their discussion within the context of wider x-risk mitigation. As such, these works have not drawn on wisdom accumulated within the field of x-risk studies.

In particular, the thesis will unite observations within the x-risk circle with the works of Law Professor Eric Johnson[68] to formalise the LHC-inspired intervention. Johnson's works largely concern how a generalist judge might have handled a preliminary injunction request in the context of the litigation associated with the LHC. Johnson does not situate his work within the broader existential x-risk landscape. However, his arguments combine well with x-risk research and vice versa. Expanding upon Johnson's work by bringing it into the x-risk landscape, thinking about how his observations align with x-risk research and vice versa, and transposing Johnson's analysis beyond the context of preliminary injunction requests involves a significant element of originality. As seen in Chapters 4 and 6, Johnson promotes a 'meta-analysis', and the thesis utilises this analysis to compartmentalise the deficiency factors[69] that might be analysed as part of evaluating the reliability of the x-risk assessment. In sum, the thesis makes a contribution by formalising and justifying a broad legal intervention in light of the LHC Controversy and legal and x-risk literature.

1.6 METHODOLOGY

1.6.1 Methods & Methodology

It is practical to distinguish the concepts of method and methodology. Methods can be understood as the techniques used to answer the research enquiries. The methods

[68] Eric Johnson, 'The Black Hole Case: The Injunction Against the end of the World' [2009] Tennessee Law Review 819; Eric Johnson, 'Agencies and Science-Experiment Risk' [2016] University of Illinois Law Review 527; Eric Johnson, 'Judicial Review of Uncertain Risks in Scientific Research' in Gilles Motet and Corinne Bieder (eds), *The Illusion of Risk Control: What does it Take to Live with Uncertainty?* (SpringerOpen 2017); Eric Johnson, 'Uncertain Risk, Science Experiments, and the Courts' [2021] Florida State University Law Review 333.

[69] Part 1.4 introduced these deficiency factors.

entail what the researcher *does* to answer the questions.[70] This means that the researcher's questions will dictate which methods should be employ. For example, consider this hypothetical inquiry: 'Is the National Environmental Policy Act suitable to protect humanity from man-made existential risks?' In addressing this, it appears sensible for the researcher to examine the legislation, which, in turn, means that a doctrinal approach is suitable. In other words, what the researcher may *do* is read the statute and NEPA cases and draw analogous conclusions. Given how the hypothetical question is formulated, it also appears suitable for the researcher to consult x-risk literature, which means that an interdisciplinary approach is suitable.

Methodology is sometimes described as the system of methods applicable to the research.[71] It is a wide concept informing the overall approach the researcher takes and rationalises the various selected methods.[72] The methodology is relevant in explaining *why* the researcher does what she does. For instance, what is the overall aim of the research? Why and how does question or investigation X contribute towards this overarching research aim? Why is method Y (e.g., a doctrinal or interdisciplinary approach) appropriate to answer question X? On this understanding of methodology, it is impractical to confine it to one Part of the thesis since it describes the entire research process.

As such, the thesis will explain selected approaches as and when necessary. For instance, this Chapter has already justified some of the decisions. Section 1.2.3 explained that the LHC-inspired intervention should satisfy three self-imposed conditions. It also explained why these conditions have been set.

[70] Tamara Hervey, Rob Cryer and Bal Sokhi-Bulley, 'Legal Research Methodologies in European Union & International Law: Research Notes (Part 1)' (2007) 3 Journal of Contemporary European Research 161, 162.
[71] ibid.
[72] ibid.

1.6.2 Assumptions

Scholars often start their research by accepting different assumptions.[73] They may begin from obvious and less obvious assumptions, and it has been suggested that those which are less obvious should be explained but not necessarily justified.[74] A project is sometimes only relevant to the extent that the assumptions are accepted.[75] For example, suppose a researcher aims to consider how the judiciary can protect future generations. A sensible assumption here would perhaps be that there is value in protecting future generations in the first place. If this is not accepted, such a hypothetical project may lose some of its relevance. Before dealing with the methods employed in this thesis, it is necessary to outline some of its underlying assumptions.

Firstly, it is assumed that the LHC Controversy can inform the configuration of legal interventions seeking to reduce the likelihood of an x-catastrophe. As such, the thesis focuses on *how* this Controversy can be used as a drafting guide rather than whether it *should* be used as such.

Secondly, it is assumed that the LHC Controversy can valuably be used as a source of inspiration for the law's response to certain x-risks regardless of whether the objective risk of x-catastrophe from particle accelerator experiments is zero.

Thirdly, the thesis supposes that examining the theoretical substance of a broad legal intervention is valuable even though the practical implementation will likely prove difficult. In effect, the thesis is less concerned with practical concerns – e.g., effective monitoring, political willingness, jurisdiction, and the consent-based nature of international law – that will someday need to be addressed to implement and efficiently apply anything resembling the LHC-inspired intervention.

[73] Mark Van Hoecke, 'Legal Doctrine: Which Method(s) for What Kind of Discipline?' in Mark Van Hoecke (ed), *Methodologies of Legal Research: Which Kind of Method for What Kind of Discipline?* (Hart Publishing 2011) preface.
[74] ibid.
[75] ibid.

Fourthly, the thesis assumes that *broad* interventions can be useful even if the x-risk landscape comprises a range of hazards. For this reason, the thesis does not aim to explore whether 'broad' interventions are likely to prove more effective than 'targeted' mechanisms or vice versa.

Fifthly, the thesis assumes that protecting humanity's long-term future is valuable. Associated presumptions include that (i) those not yet born or in existence matter, (ii) humanity's future can, all things considered, contain more good things than bad things, and (iii) humanity's actions can affect the length of its future. Whereas Chapter 5 deals with humanity's future, the aim is not to navigate the sea of philosophical and other views feeding into whether someone views the future as overwhelmingly important.

1.6.3 Mix of Methods

It has been observed that academic lawyers often omit to include anything like a methodology section.[76] Traditional legal scholarship is sometimes said to be about commenting on the law.[77] Legislation, cases and doctrinal literature constitute the main source of information within the remits of such scholarship.[78] Academic lawyers build on this information to analyse and re-present the information to give persuasive arguments on, for example, how to improve the law.[79]

It is possible to critique this approach as being too descriptive and on the basis that the 'empirical data' is limited to legal texts and court decisions.[80] This Section will not engage with the debate on the methodological challenges for some forms of legal scholarship. It suffices to say that a purely doctrinal method is not always recommended and that such an approach is unsuitable for this thesis, given its aim and questions. For example, it is not possible to consider the unique nature of x-risks by referring to black letter law. Literature on x-risk must be considered for such an

[76] Philip Langbroek and others, 'Methodology of Legal Research: Challenges and Opportunities' [2017] 13 Utrecht Law Review 1, 2.
[77] ibid 1.
[78] ibid 2.
[79] ibid.
[80] Hoecke, 'Legal Doctrine: Which Method(s) for What Kind of Discipline?' in *Methodologies of Legal Research: Which Kind of Method for What Kind of Discipline?* (n 73) 2-4.

investigation. On the other hand, when examining lessons that can be learnt from the LHC Controversy, it is sensible to resort to the doctrinal method by consulting legal literature intersecting with the said Controversy. The crucial point is that this thesis requires a mix of methods to address the research aim and questions.

It is not uncommon for legal scholars to mix different methods to address their research questions. For instance, it may be necessary to draw on quantitative and qualitative empirical non-legal data to address a research issue. Schrama gives the following example: Dutch family law presumes that it is in a child's best interest that her adoptive parents are not too young or old. Checking whether this legal stance and reality mirrors the reality, a researcher must draw insight from other disciplines like psychology.[81] Similarly, this thesis's aim—i.e., to formulate a legal intervention in the light of the LHC Controversy siting and operating across the x-risk landscape—means that it is necessary to look beyond legal resources. In fact, it would arguably be a mistake to pursue this aim with a mere rudimentary appreciation of the intricacies of the field of x-risk and associated research insights.

1.6.4 Doctrinal & Interdisciplinary

As observed above, a scholar's aim and questions should serve as a guide when selecting suitable research methods.[82] It is to be recalled that Part 1.3 outlined the thesis's main objectives and questions.[83] These questions can be answered by consulting primary and – to a greater extent – secondary legal and non-legal sources. In effect, this thesis resorts to a doctrinal and interdisciplinary approach. This is certainly not a novel mix of methods.[84]

Doctrinal: The thesis employs a method grounded in the doctrinal approach. A feature of this approach is that arguments are derived from authority, such as cases,

[81] Wendy Schrama, 'How to Carry out Interdisciplinary Legal Research: Some Experiences with an Interdisciplinary Research Method' [2011] Utrecht Law Review 147, 149.
[82] ibid.
[83] See Part 1.3.
[84] Hutchinson notes that interdisciplinary research can have a doctrinal starting point and vice-versa. Terry Hutchinson, 'The Doctrinal Method: Incorporating Interdisciplinary Methods in Reforming the Law' [2015] Erasmus Law Review 130, 133.

legislation, and legal scholarly work. The thesis has consulted secondary legal sources – which, in turn, rely on a doctrinal method – to demarcate what the LHC-inspired intervention should and should not be about. Nevertheless, given the lack of primary legal sources that apply directly to x-risk mitigation, one can question whether the thesis is truly doctrinal. Firstly, it is possible to respond by noting that one can draw analogies as part of employing a doctrinal method. Secondly, it is possible to pinpoint specific examinations in the thesis that are doctrinal in nature. These include the exploration of the *Sancho case* and other legal aspects of the LHC Controversy, such as the legal status of the European Organization for Nuclear Research ('CERN')[85] (Chapter 3); the examination of injunctions and Johnson's meta-analysis (Chapter 4); and the treatment of legal articles and books which directly or indirectly deal with x-risk and/or the LHC Controversy (Chapter 5). Many other parts in the thesis are also appropriately classified as doctrinal. Examples include the examination of x-risk mitigation through international law and the problems flowing from the need to cooperate internationally to mitigate many risks within the x-risk landscape (Chapter 2); and the comparison between courts and agencies for the purpose of applying the LHC-inspired intervention (Chapter 6).

Interdisciplinary: An interdisciplinary approach is also embraced to address the research questions. The term 'interdisciplinary' is taken to mean that the thesis combines insights from legal and non-legal disciplines. It is fair to say that the thesis relies heavily on x-risk-related research as part of its aim to consider the LHC Controversy as a drafting guide for a broad legal intervention for x-risk mitigation. As suggested above, "it would…be a mistake to pursue this aim with a mere rudimentary appreciation of the intricacies of the field of x-risk and associated research insights." The interdisciplinary approach also helps to set this thesis apart from other legal works that consider x-risk or the LHC Controversy. Many parts in the thesis use the interdisciplinary method. For example, Chapter 2 considers the wider risk landscape by resorting to x-risk-related literature. Chapter 5 will draw on such literature to justify the ASP. Chapter 6 combines legal and x-risk literature in outlining the LHC-inspired intervention's theoretical foundation. X-risk mitigation is a multifaceted topic requiring consideration of many features of the risk landscape, and an essential aim has always

[85] CERN is a research centre and an IO. Chapter 3 deals with CERN further.

been to formulate an intervention which is adequately informed by insights, observations, and wisdom from the field of x-risk.

Addressing this thesis's aim and questions is a matter of accessing and addressing legal and non-legal literature, readily available via the University's online library. In addition, many crucial resources, such as official information from CERN and reports on x-risk and the alleged risk(s) associated with particle accelerators, are freely available online. However, it is worthwhile to flag here that Chapter 3 will refer to scholarly work dealing with court cases concerning the LHC. Finding the primary sources for some of these cases has been challenging, and so the thesis has referred to the secondary legal sources instead. This is not an issue because Chapter 3 mainly focuses on the *Sancho case*, for which the official court documents have been accessed and reviewed.

1.6.5 Approach

Part 1.6 will be concluded with a few words on the approach to finding and weeding out relevant from less relevant resources. The field of x-risk research is, as indicated above, exceptionally wide. This is the case even when the field is approached from a legal point of view and within the sub-field of x-risk mitigation. Because of this, the thesis has used the research parameters outlined in Section 1.3.3. Significantly, these parameters have been relevant when distinguishing relevant from less relevant resources. For example, Section 1.3.3 noted that the thesis concentrates on *preventing* x-catastrophes rather than enhancing humanity's *resilience* to materialised catastrophes. This parameter means the thesis is not too concerned with works exploring how humanity's resilience can be improved. Another example is that the thesis, which focuses on *anthropogenic* x-risks, is not overly concerned with works on natural extinction risks from, say, potential comet or asteroid impacts. Section 1.3.3 outlined further research parameters – such as the focus on *sudden* catastrophes and those flowing from human *error* – which have informed the selection of resources.

Beyond the parameters in Section 1.3.3, the thesis has resorted to additional strategies to weed out irrelevant resources. Indeed, the thesis's aim and questions intersect with a range of topics, such as AI, biotechnology, experimental particle

physics, injunctions, science courts, the length and value of humanity's future, physical eschatology, cost-benefit analysis, international organisations, various unconscious biases, climate litigation and much more. Most of these topics can, in and of themselves, constitute the basis for multiple PhDs. In effect, some overarching filters have been required to (i) keep the thesis focused and streamlined, (ii) avoid indiscriminate use of resources, and (iii) avoid falling down a rabbit hole with the effect of consulting too many resources, in too much detail, which are too distant from the thesis's essential investigation.

(i) Legal Resources

The following question has been used to select relevant *legal* resources: Does this legal work concern x-risk, the risk of human extinction, and/or some aspects of the LHC Controversy? If the answer to this question is no, that scholarly piece has not generally been a foundational work for this thesis. Of course, this overarching approach has not been absolute. It has been necessary to use 'other' legal resources to offer a contextual and authoritative starting point for certain discussions. For example, Chapter 3 deals with international organisations ('IOs') and why it might be difficult to challenge the activities of IOs in domestic courts. This is examined because CERN – the primary risk originator in the context of the LHC experiments – is an IO. As part of this discussion, the thesis has referred to legal resources dealing with the status and nature of IOs. As another example, Chapter 4 considers injunctions from a general perspective to provide a background to Johnson's so-called meta-analysis.

The overarching filter—in the form of the question above—for distilling legal resources has been crucial in keeping the thesis feasible and streamlined. Without this guiding approach, it would have been tricky for one PhD project to deal with and devote equal attention to every facet that may feed into the aim of examining the LHC Controversy as a drafting guide for x-risk mitigation.

(ii) X-risk-related Resources

Navigating x-risk-related resources has been a matter of sticking to the parameters in Section 1.3.3. However, an additional sorting technique has been used to distinguish

relevant from less relevant x-risk-related resources. More specifically, the thesis has prioritised resources within either one of the following two overarching categories:

- Category (i): The x-risk-related resource can plausibly assist in mapping the x-risk landscape.

- Category (ii): The x-risk-related resource concerns the alleged risks associated with particle accelerators such as the LHC or RHIC.

In respect of Category (i), the thesis has assumed that literature dealing with the following aspects has been particularly relevant in mapping the x-risk landscape: (a) Definitions and threshold concepts to understand the notion of x-risk; (b) alleged hazards—beyond particle accelerator experiments—within the x-risk landscape; (c) obscuring features making it difficult to study and mitigate existential risk; and (d) the significance of humanity's long-term future.

Regarding Category (ii), the thesis has focused on scholarly works which explore the alleged extinction risks associated with the LHC or RHIC. Significantly, it has not excluded works that do not specifically use the terminology 'existential' risk. This is because some foundational pieces on the LHC Controversy talk about 'extinction' and 'the end of the world' rather than existential risk.

To summarise, the research parameters in Section 1.3.3 and the focus on x-risk-related resources falling within Category (i) or Category (ii) have helped the thesis to navigate the wealth of literature.

1.7 THE REMAINING STRUCTURE OF THE THESIS

Introduction and Conclusion aside, the thesis comprises five substantive chapters. Below is an overview of each of the remaining chapters and how they fit together.

Chapter 2: The X-risk Landscape

Chapter 2 explores the x-risk landscape. By definition, a *broad* legal intervention will, to a greater or lesser extent, sit and operate within and across the x-risk landscape. As such, it is necessary to survey the landscape and identify issues and trends that may affect broad legal interventions. Chapter 2 will explore the notion of x-risk in more depth, some of the key hazards within the risk landscape and features making it difficult to study and mitigate x-risk.

Chapter 2 will facilitate an answer to one of the thesis's sub-questions. Namely, what makes x-risk unique, and why is it difficult to study and mitigate this risk category? It is also to be recalled that the LHC-inspired intervention should satisfy three conditions.[86] The next Chapter is vital in thinking about the condition that the intervention should find utility beyond the LHC Controversy.

Chapter 3: Legal Aspects of the LHC controversy

Chapter 3 considers the LHC Controversy in greater detail. This is important since this Controversy is used as a source of inspiration for the LHC-inspired intervention. While most experts agree that the risk is zero, there have been concerns that high-energy physics experiments, such as those facilitated by the LHC and RHIC, may trigger an x-catastrophe. In particular, Chapter 3 considers *Sancho v US Department of Energy*[87] which arose from the LHC Controversy. In doing so, the Chapter also considers the 'review barriers' that can complicate different modes of external review when addressing the disagreements on whether particle accelerators pose an x-risk.

Chapter 3 is important in addressing the following sub-question: What makes it complicated for an external party to address the disagreements about whether particle accelerator experiments pose an x-risk? This Chapter is also fundamental in addressing the self-imposed conditions that the LHC-inspired intervention should respond to the review barriers.

[86] Firstly, the intervention should respond to the 'review barriers' that may thwart distinct modes of external review in the LHC Controversy. Secondly, it should utilise the ASP. Finally**,** with some degree of plausibility, the mechanisms should find utility beyond the LHC Controversy.
[87] 578 F Supp 2d 1258 (D Haw 2008) 1259 [hereinafter *Sancho case*].

Chapter 4: The Ultimate Injunction

Chapter 4 examines Johnson's meta-analysis. Johnson suggests that a generalist judge in the US could have employed a kind of meta-analysis when dealing with the preliminary injunction request in the *Sancho case*.

It is to be recalled that the LHC-inspired intervention aims to target the reliability of the x-risk assessment. More specifically, the intervention can be levied to critically evaluate the reliability of the assessment by inviting an external reviewer to analyse a set of interconnected deficiency factors. Significantly, the thesis turns to Johnson's meta-analysis as an instrument for compartmentalising these deficiency factors. Engaging with Johnson's meta-analysis will feed into Chapter 6, which aims to answer the thesis's central research question.

Chapter 5: The Astronomical Stakes Proposition and Legal Interventions

Chapter 5 deals with the ASP. The Chapter also considers a mix of hypothetical *broad* and *targeted* legal interventions for x-risk mitigation and argues that the ASP may influence and empower such interventions. Chapter 5 will address the following sub-question of this thesis: What is the justification for the astronomical stakes proposition, and what is the relevance of this concept vis-à-vis legal interventions for x-risk mitigation? Finally, one of the self-imposed conditions is that the LHC-inspired intervention should utilise the ASP. Chapter 5 is central in justifying this condition.

Chapter 6: A Broad Legal Intervention in Response to the LHC Controversy

Chapter 6 will formalise the theoretical substance for the LHC-inspired intervention. In essence, the Chapter will argue for an intervention that centres around giving an external reviewer a principled basis upon which to base a finding that some endeavours should not presently be allowed to go ahead since, and notwithstanding the review barriers or a combination thereof, there are reasonable doubts as to the reliability of the scientific work assessing existential risk linked to these endeavours.

As mentioned above, the reliability of the x-risk assessment can reasonably be questioned by analysing interconnected deficiency factors that do not involve

consideration of the accuracy of the scientific theories or the quantification of the risk of catastrophe. These factors require consideration of (i) the actors who consider the relevant x-risk and (ii) the possibility their conclusions are based on flawed information or an incomplete state of knowledge or understanding.

Chapter 6 is central in addressing the thesis's main research question, which, as might be recalled, is as follows: How can the LHC Controversy inspire the design of a broad legal intervention to mitigate x-risk? In addition, Chapter 6 addresses the final interlinked sub-question: How does the LHC-inspired intervention (i) operate in theory, (ii) align with x-risk studies, and (iii) utilise the ASP?

1.8 CONCLUSION

This thesis considers how the LHC Controversy can inspire the design of a broad legal intervention for x-risk mitigation. In other words, this Controversy is used as a drafting guide in exploring the law's response to some, though not all, risks within the x-risk landscape.

One crucial point deserves repetition here: "[T]he value of this thesis is not contingent on the correctness of the assertion that running the LHC, the RHIC, or any other accelerator can cause an existential catastrophe. In other words, whether the x-catastrophes associated with high-energy physics experiments are physically possible or not is immaterial in the context of this thesis."[88] This thesis does not attempt to convince anyone that the LHC is dangerous. Instead, it delivers the simple message that the LHC Controversy reveals that law might play a role in controlling the underlying reliability of certain x-risk assessments going forward.

The main finding is that the LHC Controversy (and accompanying legal and x-risk literature) reveal the value of a legal mechanism targeting the reliability of the scientific work assessing and potentially excluding x-risk. More specifically, the thesis argues for a legal response that centres around giving an external reviewer a principled basis upon which to base a finding that some endeavours should not presently be allowed

[88] See Section 1.2.1.

to go ahead since, and notwithstanding the review barriers, there are reasonable doubts as to the reliability of the work assessing x-risk(s) linked to these endeavours.

The reliability of the relevant assessment can reasonably be questioned by inviting an external reviewer to analyse interconnected deficiency factors that do not involve validation of conflicting scientific theories or the quantification of the risk of catastrophe. Instead, the reviewer can, broadly speaking, analyse and consider (i) the actors who assessed the risk and (ii) the possibility that their conclusions are based on flawed information or an incomplete state of knowledge or understanding.

2 CHAPTER TWO: THE X-RISK LANDSCAPE

2.1 BACKGROUND

The Milky Way is one out of an astonishing number of galaxies in the observable universe.[89] Judge Richard Posner writes that there are probably "…billions, maybe trillions, of planets, on some of which intelligent life almost certainly evolved long before it evolved on earth."[90] It is surprising, then, suggests Posner, that no beings have developed the technological means to contact us.[91]

This paradox is not new, and in 1950, Enrico Fermi famously asked: "Where are they?"[92] The answer might reside in the fact that we are all alone in the universe. It is also possible that any beings are too far away despite their technological maturity. But Posner also highlights a less cheerful hypothesis. It is conceivable that whenever a race "…reaches the level of technological sophistication…at which it would be possible to make contact with intelligent life elsewhere in the universe, it destroys itself."[93] It unleashes forces which it cannot control.[94]

This thesis does not explore the merit of the different hypotheses to the 'Fermi paradox'. Instead, whatever the fate of other intelligent beings in the universe (if any), scholars on our planet speculate that humanity has "...reached the threshold where

[89] Christopher Conselice, Aaron Wilkinson, Kenneth Duncan and Alice Mortlock, 'The Evolution of Galaxy Number Density at z < 8 and its Implications' (2016) 830(83) The Astrophysical Journal <https://iopscience.iop.org/article/10.3847/0004-637X/830/2/83> accessed 24 April 2023. The study notes that there are "…at least 2×10^{12} (two trillion) galaxies in the currently visible universe, the vast majority of which cannot be observed with present-day technology as they are too faint."

[90] Richard Posner, *Catastrophe: Risk and Response* (Oxford University Press 2004) 20; In contrast, Sandberg, Drexler and Ord suggest it is a mistake to be highly confident that the universe contains other civilizations based on size. Anders Sandberg, Eric Drexler and Toby Ord, 'Dissolving the Fermi Paradox' (Future of Humanity Institute 2018) <https://arxiv.org/pdf/1806.02404.pdf> accessed 24 April 2023, 1.

[91] Posner, *Catastrophe* (n 90) 20.

[92] Eric Jones, 'Where is Everybody? An Account of Fermi's Question' (Technical Report LA-10311-MS, Los Alamos National Laboratory 1985), in Sandberg, Drexler and Ord, 'Dissolving the Fermi Paradox' (n 90).

[93] Posner, *Catastrophe* (n 90) 20; This hypothesis is not novel as highlighted in Sandberg, Drexler and Ord, 'Dissolving the Fermi Paradox' (n 90) 2.

[94] Posner, *Catastrophe* (n 90) 20; See also Harry Lehmann, *No Canary in the Quanta: Who Gets to Decide if the Large Hadron Collider is Worth Gambling Our Planet?* (Green Swan 2009), 2, 11.

we might be able to destroy ourselves."[95] If humanity's actions were to bring about a localised x-catastrophe, our species might just come to perpetuate an identical paradox at some other distant place in the universe.

2.1.1 Aims of Chapter 2

This Chapter examines aspects relating to the x-risk landscape. It will provide a bird's-eye view of the x-risk landscape. Doing so is significant for several reasons. Firstly, the Chapter will provide useful background context, given that this thesis deals with x-risk as a distinct risk category. Secondly, the thesis aims to promote a *broad* legal intervention in response to the LHC Controversy. Broad interventions will sit within and operate across the x-risk landscape. Ultimately, the bird's-eye view in this Chapter will reveal factors that may affect any intervention that strives to reduce the likelihood of existential risk from two or more hazards. Finally, Chapter 2 will help to answer the following sub-question: "What makes existential risk unique, and why is it difficult to study and mitigate this risk category?"

2.2 EXISTENTIAL RISK: DEFINITION AND RELATIVE SERIOUSNESS

This Part will clarify the thesis's definition of *existential risk* and *humanity*. It also deals with risk and the elements of consequence and probability. Finally, this Part ends with a note on classifying a risk's seriousness.

2.2.1 Existential Risk and Humanity

Chapter 1 briefly introduced the thesis's definition of existential risk. This Section offers a more comprehensive explanation. There are different definitions of the term existential risk.[96] As noted by Torres, these have different advantages and

[95] Ord, *Precipice* (n 16) 22.

[96] Torres provides an overview of five distinct definitions of 'existential risk' and argues that one should employ the definition that best fits the specific context and audience. Phil Torres, 'Existential Risks: A Philosophical Analysis' [2019] Inquiry 614.

disadvantages and may suit distinct audiences.[97] With that being said, this thesis will use the following definitions:

Table 2.1

Existential Risk	A risk that threatens the premature extinction of humanity.
Existential Catastrophe	The premature extinction of humanity.
Humanity	Earth-originating intelligent life (inclusive of, but not limited to, *Homo sapiens*.

The terminology in Table 2.1 largely mirrors some parts of Bostrom's seminal definition, which is repeated here for convenience: "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."[98]

Before pushing on, it is useful to note that x-risk-related studies existed long before Bostrom's seminal definition. Indeed, while Bostrom formalised the concept of existential risk in 2002[99] and popularised its use,[100] some earlier works have been material for the growth of the x-risk field. As one example, it is possible to note Leslie's book from 1996.[101] While the terminology differs, his book deals with various natural and anthropogenic (existential) risks.[102] It also considers other aspects relevant to the study of x-risk, including Brandon Carter's doomsday argument, selection effects, space colonisation, Fermi's Paradox, as well as philosophical doctrines which "…cast doubt on any real ethical need to keep the human race in existence."[103]

[97] ibid.
[98] Bostrom, 'Existential Risk Prevention' (n 15) 15. This definition can be criticised. Torres highlights that it can be problematic to define and agree on the meaning of inter alia 'intelligent life' and 'potential'. Phil Torres, 'Problems with Defining an Existential Risk' (IEET 2015) <https://ieet.org/index.php/IEET2/more/torres20150121> accessed 24 April 2023; Cotton-Barratt and Ord also notes that Bostrom's definition lacks clarity. Owen Cotton-Barratt and Toby Ord, 'Existential Risk and Existential Hope: Definitions', (Future of Humanity Institute – Technical Report #2015-1) <http://www.fhi.ox.ac.uk/Existential-risk-and-existential-hope.pdf> accessed 24 April 2023, 3.
[99] Bostrom, 'Existential Risks' (n 15).
[100] This is recognised in Torres, *The End* (n 35) 27.
[101] John Leslie, *The End of the World: The Science and Ethics of Human Extinction* (Routledge 1996).
[102] ibid 25-131.
[103] ibid 155.

Getting back on track, Bostrom's definition speaks of two x-catastrophe tracks: There is (i) humanity's premature extinction and (ii) the permanent and drastic destruction of its potential. Notably, the definitions of x-risk and catastrophe in Table 2.1 encapsulate both prongs of Bostrom's definition.[104] This is because humanity's premature extinction axiomatically destroys all potential for future development.

Ord also acknowledges that extinction will result in a permanent and drastic curtailment of all future potential.[105] He defines an x-risk as one which threatens the destruction of humanity's long-term potential.[106] In other words, Ord only echoes Bostrom's second definitional prong.[107] He suggests that the focus on the destruction of humanity's long-term potential draws attention to what matters when dealing with existential risk.[108]

This thesis also considers that humanity's potential across the far future is what truly matters. Nevertheless, it resorts to the definitions in Table 2.1 because it encapsulates **(i)** the sort of x-risk relevant to this thesis[109] and **(ii)** both facets of Bostrom's definition. The same is not true of Ord's description: Whereas premature extinction inevitably destroys humanity's long-term potential, the destruction of humanity's long-term potential does not inevitably flow from extinction.[110] Admittedly, the definition embraced in this thesis is open to protest for being too narrow[111] and simplistic.[112] Nonetheless, using the term 'humanity' as opposed to 'human' and appreciating that the loss of all future potential is inherent in extinction can offset the simplicity. In any

[104] It is possible to talk about the second prong in isolation from the first prong. I.e., the destruction of humanity's potential will not necessarily need to flow from humanity's extinction.
[105] Ord, *Precipice* (n 16) 300.
[106] ibid 37.
[107] ibid 300.
[108] ibid 37.
[109] See the research parameters in Section 1.3.3.
[110] Farquhar and others write that a catastrophe leaving some survivors alive can qualify as existential on Bostrom's definition if they are unable to rebuild society. The distinguishing factor is the permanent curtailment of recovery. Farquhar and others, 'Existential Risk' (n 61) 6.
[111] Cotton-Barratt and Ord prove an example where a totalitarian regime takes perpetual control. They show that the definition of x-catastrophe should encompass this situation. Cotton-Barratt and Ord, 'Existential Risk and Existential Hope' (n 98).
[112] Torres, 'Existential Risks: A Philosophical Analysis' (n 96).

event, the vital commonality is the seriousness of existential risk. An existential catastrophe will destroy the future and, with it, humanity's long-term potential.[113]

Before moving on, the term 'humanity' should be clarified, given the term's frequent use herein. As outlined in Table 2.1, humanity means "Earth-originating intelligent life" as opposed to *Homo sapiens*. The main point to appreciate is that humanity does not only refer to the human species. The term is much broader than that. There are justifications for employing a definition of humanity stretching beyond *Homo sapiens* when dealing with x-risk and subscribing to longtermism.

As suggested by Torres, *Homo sapiens* may evolve into a species of posthumans, and such pseudo-extinction may be regarded as desirable if these posthumans are free from many of the predicaments affecting the life of many today.[114] Bostrom discusses similar notions explaining that we have "…no reason to suppose that the biological species concept tracks what we have reason to value."[115] He suggests that humans may evolve or self-modify "…to such an extent that it no longer satisfied the biological criteria for species identity…with contemporary *Homo sapiens*…"[116] This transformation need not count as an x-catastrophe.[117]

An article co-authored by fourteen individuals makes a similar case for why more extensive definitions of what it means to be "human" might be suitable when dealing with longer time scales.[118] On that note, legal interventions working across vast time scales may also need to account for a more expansive definition of humanity. To

[113] Ord, *Precipice* (n 16) 37; Bostrom, 'Existential Risk Prevention' (n 15) 17.
[114] Torres, 'Existential Risks: A Philosophical Analysis' (n 96).
[115] Bostrom, 'Existential Risk Prevention' (n 15) 20.
[116] Ibid.
[117] ibid.
[118] To get a sense of a 'longer time scale', Baum and others start their paper with this question: "What will human civilization look like in one million, one billion or one trillion years?" They continue by explaining that their "…paper uses the term "human civilization" to refer to any civilization that traces to the current human population. This definition includes civilizations led by genetic descendants of *Homo sapiens sapiens*, as well as civilizations led by biological or non-biological beings that are engineered by *Homo sapiens sapiens* or its genetic descendants. This definition permits a study of long-term trajectories that does not need to constantly account for whether the civilization is still in some sense "human." Seth Baum and others, 'Long-term Trajectories of Human Civilization' [2019] Foresight 53, 53-54.

summarise, the definition of humanity in Table 2.1 is appropriate because it accounts for what might transpire in the far future.

2.2.2 Risk: Consequence and Probability

Risk is a big topic, and back in the 80s, Kaplan and Garrick recognised that the literature had grown very large.[119] They highlighted that one could find references to business risk, economic risk, investment risk, safety risk and political risk.[120] Beyond looking at qualitative aspects of risk,[121] their paper noted that one way of defining risk is to express it as the probability and consequence of a particular scenario.[122]

While there are different ways to express and define risk, this thesis will take the term risk to mean "an unwanted event which may or may not occur."[123] It is possible to find the elements of *probability* and *consequence* within this definition. Firstly, in terms of the consequence, when the unwanted event is the premature extinction of humanity, the risk is classified as an x-risk.

Secondly, in terms of probability, what is required is that the unwanted event may – even though it may not – occur. In other words, no risk can exist absent subjective or objective probability.[124] The distinction between the objective and subjective notions of probability is noteworthy within the x-risk landscape because the likelihood of existential risk cannot be derived from keeping a track record of how frequently a

[119] Stanley Kaplan and John Garrick, 'On The Quantitative Definition of Risk' [1981] Risk Analysis 11.
[120] ibid 11.
[121] ibid 12.
[122] More specifically, they defined risk as a set of triplets: (s_i, p_i, x_i), where: s_i is the *i*th *scenario*; p_i, is the *probability* of that scenario; and, x_i is the *consequences* of that scenario. ibid 13.
[123] Sven Hansson, 'Risk' (*The Stanford Encyclopedia of Philosophy* 2018) <https://plato.stanford.edu/archives/fall2018/entries/risk/> accessed 24 April 2023.
[124] Beard, Rowe and Fox explain the objective notion of probability like this: "According to this approach, probabilities are fundamentally related to the frequencies of events based on past observations." In contrast, "[t]he second notion of probability is the Bayesian, or subjective, account, according to which probabilities represent our level of belief that a phenomenon will occur. One begins with a subjective prior belief about the probability of an event and then updates this via Bayes' Theorem (or Bayes' rule), which specifies how additional information affects the probability of an event." Simon Beard, Thomas Rowe and James Fox, 'An Analysis and Evaluation of Methods Currently Used to Quantify the Likelihood of Existential Hazards' (2020) 115 Futures <https://doi.org/10.1016/j.futures.2019.102469> accessed 24 April 2023.

particular x-catastrophe has occurred. The reason for this is that an x-catastrophe will prevent anyone from making and recording such observations in the first place.

Beard, Rowe and Fox have found that subjective probability generally dominates the field of x-risk.[125] Part 2.3 below will show that assigning precise probabilities to x-risk scenarios is often difficult.[126] The LHC Controversy is no exception. In fact, with statements such as "…it is impossible for microscopic black holes to be produced at the LHC…"[127] proponents circumvented the element of probability and the concept of existential risk. Thus, a legal intervention targeting high-energy physics experiments as worthy of regulation faces the obstacle of proving that the experts' assertions are either wrong, or that there is a probability that they might be wrong.

2.2.3 Seriousness of Risk

The sections above have introduced the idea of risk and *existential* risk. It should be appreciated that the latter is a *serious* risk owing to its consequences.[128] Before examining the hazards associated with existential risk, this section will briefly summarise Bostrom's approach to characterising a risk's seriousness.[129] Bostrom admits that his approach is not necessarily universally accepted. Indeed, some scholars within the x-risk circle propose a refined risk typology.[130] Nonetheless, his approach provides a starting point for appreciating the relative seriousness of risks.

---

[125] ibid; See also Seth Baum, 'Quantifying the Probability of Existential Catastrophe: A reply to Beard et al.' (2020) 123 Futures <https://doi.org/10.1016/j.futures.2020.102608> accessed 24 April 2023.
[126] See also Bostrom, 'Existential Risk Prevention' (n 15) 16; Ord, *Precipice* (n 16) 39.
[127] CERN, 'The Safety of the LHC' <https://home.cern/science/accelerators/large-hadron-collider/safety-lhc> accessed 28 April 2023.
[128] Bostrom, 'Existential Risks' (n 15).
[129] Bostrom, 'Existential Risk Prevention' (n 15) 16-17.
[130] Liu, Lauta and Maas suggest that Bostrom's typology provides little guidance on differentiating the diverse risks within the x-risk category. This is because the risks are only distinguished according to their *severity* and *scope* and not according to their *source*, *characteristics*, or *complexity*. Liu, Lauta and Maas, 'Governing Boring Apocalypses' (n 61) 9; Avin and others also propose a refined classification scheme, albeit when it comes to GCRs. The scheme breaks down the analysis of GCR scenarios into three components. (1) a *critical system* whose safety boundaries are breached by a potential threat, (2) the *global spread mechanisms* by which such threat may spread globally and impact the human population, and (3) the *prevention and mitigation failures* which refers to the manner in which we may fail to prevent components (1) and (2). See Shahar Avin and others, 'Classifying Global Catastrophic Risks' [2018] Futures 20, 21-24.

As described by Bostrom, a risk is evaluated negatively, and to establish its seriousness, it is necessary to employ a standard of evaluation by which to measure the negative value of a loss scenario.[131] Bostrom's framework depends on the assumptions that (i) persons' lives have positive value which is aggregative in the sense that two lives have twice the value of one life, (ii) the quality and duration of lives are constant, and (iii) the value of lives do not depend on when it occurs, whether it already exists or is yet to begin.[132] Bostrom characterises a risk's seriousness within this framework using the variables of *scope*, *severity* and *probability*.[133]

Table 2.2 below explain what these variables mean. The variables are sometimes followed by other scholars within the existential risk circle to determine the seriousness of risk.[134] However, there are other ways to evaluate seriousness. For example, Posner quantifies the 'gravity' (which seemingly means the same thing as seriousness) of a risk by using the functions of *probability* that it will materialise and the *awfulness* of the consequences if it materialises.[135]

Table 2.2

Scope	Relates to relates to the size of the population at risk. The scope can be *personal*,[136] *local*,[137] *global*,[138] *trans-generational*[139] and *pan-generational.*[140] Existential risks are characterised by their pan-generational scope which means that they will affect humanity over all future generations.[141]
Severity	Relates to how badly the size of the population would be affected. One can classify the severity as being *imperceptible*,[142] *endurable*[143] and

[131] Bostrom, 'Existential Risk Prevention' (n 15) 16.
[132] ibid.
[133] ibid.
[134] Farquhar and others, 'Existential Risk'(n 61) 6.
[135] Posner, *Catastrophe* (n 90) 8.
[136] Affecting one person.
[137] Affecting a distinct group or some geographical region.
[138] Affecting a large part of or the entire human population.
[139] Affecting humanity for generations.
[140] Affecting humanity over all future generations.
[141] Bostrom, 'Existential Risk Prevention' (n 15) 17.
[142] Barely noticeable.
[143] Causing harm which is significant without completely ruining quality of life.

	crushing.[144] Existential risks are characterised by their crushing severity which means that they cause death or a drastic and permanent reduction of quality of life.[145] (The focus of this thesis on premature extinction means that 'crushing severity' is taken to mean death).[146]
Probability	Relates to the likelihood of the disaster. [147]

(i) Existential Risk v Global Catastrophic Risk

Existential risk can be viewed as a subset of global catastrophic risks (GCRs).[148] However, it is pertinent to distinguish between GCRs and x-risks when examining a risk's seriousness. This is especially true when subscribing to the view of longtermism.

As mentioned above, this thesis defines an x-risk as one that "threatens the premature extinction of humanity." In contrast, a GCR does not involve the threat of extinction. Bostrom and Ćirković write that the term GCR lacks a sharp definition.[149] That being said, they employ the term to refer to a risk with the potential to inflict serious damage to human well-being on a global scale. They suggest that a catastrophe causing 10 million fatalities would count as a global catastrophe.

A global catastrophe has also been defined by other scholars as "…a possible event or process that, were it to occur, would end the lives of approximately 10% or more of

[144] Causing death or a permanent and drastic reduction of quality of life.
[145] Bostrom, 'Existential Risk Prevention' (n 15) 17.
[146] One can argue that eternal torture is worse than death. While not explored in any detail herein, Chapter 5 briefly considers risks of astronomical suffering ('s-risks'). See also David Althaus and Lukas Gloor, 'Reducing Risks of Astronomical Suffering: A Neglected Priority' (Center on Long-Term Risk 2016) <https://longtermrisk.org/reducing-risks-of-astronomical-suffering-a-neglected-priority/#link_ajs-fn-id_4-3131> accessed 24 April 2023; See also Max Daniel, 'S-risks: Why they are the Worst Existential Risks, and how to Prevent them' (Center on Long-Term Risk 2017) <https://longtermrisk.org/s-risks-talk-eag-boston-2017/> accessed 24 April 2023.
[147] Bostrom, 'Existential Risk Prevention' (n 15) 16.
[148] Bostrom and Ćirković, 'Introduction' (n 12) 4.
[149] ibid 1-3.

the global population, or do comparable damage."[150] This is the definition of a 'global catastrophe' that this thesis will embrace.

Distinguishing between GCRs and x-risks is vital in the context of this thesis because the two are sometimes amalgamated for the purpose of legal mitigation strategies.[151] In other words, some scholars examine the mitigation of x-risks and GCRs under the same proposed framework. There is nothing inherently wrong with such an approach. However, it is important to remember that only an x-catastrophe will destroy humanity's entire future. In contrast, humanity has already 'survived' the materialisation of GCRs, including the two world wars and certain pandemics.

As examined in Chapter 5, this thesis subscribes to the view that there is, in fact, a massive difference between x-risks and GCRs and that legal interventions aiming to reduce the former category can utilise the ASP.

2.3 SOURCES OF RISK

We have largely considered existential risk in the abstract up until this point. Thus, this Part examines how x-risk is sometimes associated with, inter alia, nuclear war, geoengineering, pandemics and biotechnology, AGI, unknown unknowns and high-energy physics experiments. It is possible to refer to these as hazards – a term taken to mean the *source* of potential danger.[152] One could spend a great deal of time on

[150] Owen Cotton-Barratt and others, 'Global Catastrophic Risks' (Global Challenges Foundation 2016) <https://globalchallenges.org/wp-content/uploads/2019/07/Global-Catastrophic-Risk-Annual-Report-2016.pdf> accessed 24 April 2023, 22.

[151] See Grant Wilson, 'Minimizing Global Catastrophic and Existential Risks from Emerging Technologies Through International Law' [2013] Virginia Environmental Law Journal 307.

[152] Kaplan and Garrick explain that hazard is the source of danger, whereas risk is the likelihood of conversion of that source into loss or damage. Kaplan and Garrick, 'On The Quantitative Definition of Risk' (n 119) 12; Similarly, Kemp and Rhodes define hazards as "direct threats that could cause global calamity." They explore hazards including AI, Pandemics and Solar Geoengineering. Kemp and Rhodes, 'The Cartography of Global Catastrophic Governance' (n 53) 4.

any of the hazards below.[153] Nevertheless, this Part is comparatively brief and provides an overview of some key hazards within the x-risk landscape.[154]

Before considering the selected hazards, Section 2.3.1 will introduce a method to estimate the total *natural* extinction risk per century. It was mentioned in Chapter 1 that scholars often distinguish between natural and human-made risks, even though the distinction is not always easy to make.[155] It is legitimate to wonder why this thesis deals with the potential total *natural* extinction risk per century, given that the LHC-inspired intervention is concerned with human-made risks. Two reasons can justify a brief discussion on *natural* x-risks.

Firstly, human-made x-risks can generally be regarded as more pressing than natural ones. A brief discussion on natural risks which reinforces this argument can, in effect, guide mitigation efforts. For instance, it supports the claim that legal interventions aiming to mitigate *anthropogenic* x-risk, such as the LHC-inspired intervention, may be more urgently needed than those concerning *natural* x-risks.

Secondly, the discussion on natural x-risks is interesting in light of what Cambridge Professor Adrian Kent terms "…the argument of dominant risk."[156] This argument revolves around the view that a new risk may be acceptable if its probability is lower than that of some existing risks.[157] As written by Kent, "[a] refinement of this view is that a new artificial risk is acceptable only if smaller than presently unavoidable *natural* risks."[158] Although Kent is sceptical of such a view, it is not inconceivable that proponents of a given endeavour will attempt to justify the x-risk(s) associated with their endeavour by resorting to some version of the argument of dominant risk. Likewise, it is not inconceivable that a future legal intervention concerning risk

[153] Ord, *Precipice* (n 16) 67-162; Torres, *The End* (n 35) 43-159; Leslie, *The End of the World* (n 101) 25-131; Posner, *Catastrophe* (n 90) 21-89.
[154] For another brief overview, see Häggström, *Here Be Dragons* (n 22) 188-201; Matt Boyd and Nick Wilson, 'Existential Risks to Humanity Should Concern International Policymakers and More Could Be Done in Considering Them at the International Governance Level' [2020] Risk Analysis 2303, 2304.
[155] See e.g., Bostrom and Ćirković, 'Introduction' (n 12) 7; Cotton-Barratt, Daniel and Sandberg, 'Defence' (n 37) 273-275; Posner, *Catastrophe* (n 90) 21-30.
[156] Adrian Kent, 'A Critical Look at Risk Assessments for Global Catastrophes' [2004] Risk Analysis 157, 166.
[157] ibid.
[158] ibid.

acceptability uses this credo as some kind of threshold. For example, an argument might be made that a human-made existential risk is acceptable if the estimated probability is lower than the estimate of total *natural* extinction risk per century. Of course, there are difficulties with resorting to the dominant risk argument because it hinges upon being able to assign probabilities to the new artificial x-risk.

2.3.1 Natural Risks - Total Risk

The risk of extinction from natural processes, including asteroid impacts and super-volcanic eruptions, has been in the background for thousands of years.[159] The existential threat from these hazards is said to come from the ensuing 'darkened sky' – i.e., the *impact* or *eruption* will create thick dust rising to the atmosphere blocking the sunlight.[160] What may follow is global cooling, darkness and mass starvation.

It has been observed that some notable natural x-risks were unknown until relatively recently.[161] Thus, one may reason that it would be a mistake to think that humanity has now discovered every natural extinction risk.[162] There might be other unknown natural extinction risks out there. Nevertheless, it seems unlikely that some mysterious and unknown natural risk is lurking around the corner, with Ord suggesting that there is a way of estimating the *total natural extinction* risk per century.

Ord resorts to the fossil record for gauging how long species like us survived.[163] It is reasoned that *if* the total natural extinction risk is exceptionally high, say 20% per century, a species will not survive for very long – only five centuries, on average. But humanity has survived for hundreds of thousands of years.[164] Ord explores different ways of using the fossil record along these lines.

---

[159] Bostrom, 'Existential Risk Prevention' (n 15) 15; Farquhar and others, 'Existential Risk' (n 61) 6; Toby Ord, 'Will We Cause Our Own Extinction? (Public Lecture, Future of Humanity Institute 2015) <https://www.cser.ac.uk/events/extinction/> accessed 24 April 2023. See from [00.07.41] for natural risks and from [00:30:35] for anthropogenic risks.

[160] Ord, *Precipice* (n 16) 68, 74; Farquhar and others, 'Existential Risk' (n 61) 10.

[161] As explained by Ord, it was not before 1960 that proof emerge that Earth had been hit by a large asteroid or comet. Furthermore, gamma ray bursts were unknown until 1989. See Ord, *Precipice* (n 16) 81.

[162] ibid.

[163] ibid.

[164] See Farquhar and others, 'Existential Risk' (n 61) 6.

Let's consider one of his techniques.[165] He estimates that *Homo sapiens* has survived for around 200,000 years (i.e., 2,000 centuries). With no recorded extinctions, the risk per century is *lower* than one in 2,000. On the other hand, the risk should not be zero in 2,000 because that would mean that extinction is impossible. Ord highlights that there is debate about what probability to assign in such cases: He recognises that all suggested methods produce numbers between *zero* and *one* in 2,000; that is, 0 to 0.05%. He writes that this range (i.e., 0-0.05%) can be treated as a rough best-guess estimate of total natural extinction risk per century.[166] The above can be represented by following equation:

Equation 2.1

$$\text{Estimate of total natural extinction risk per century } (\%) = \left(\frac{1}{2000}\right)(100)$$
$$= 0.05\%$$

By using Equation 2.1, the estimates can be adjusted by substituting the number of centuries a species has survived. In essence, using the fossil record, Ord makes a somewhat compelling case that risks traceable to anthropogenic processes are more likely to cause extinction than natural ones.[167] He suggests that "…we face about a thousand times more anthropogenic risk over the next century than natural risk…"[168]

In contrast to the estimate of the total natural extinction risk, it is not feasible to use the past when it comes to anthropogenic risks.[169] The first human-made existential risk arguably came with the development of the first atomic bomb.[170] For instance, Ord dates the beginning of *the Precipice*[171] to 16 July 1945: The very moment of the Trinity Test.[172] Torres suggests that this conclusion is probably incorrect because two

[165] Ord, *Precipice* (n 16) 81-82.
[166] ibid 82.
[167] Ord, 'Will We Cause Our Own Extinction? (n 159); Andrew Snyder-Beattie, Toby Ord and Michael Bonsall, 'An Upper Bound for the Background Rate of Human Extinction' (2019) 9 Scientific Reports <https://doi.org/10.1038/s41598-019-47540-7> accessed 24 April 2023.
[168] Ord, *Precipice* (n 16) 87.
[169] ibid; This is also recognised in Bostrom, 'Existential Risk Prevention' (n 15) 15-16.
[170] ibid 29.
[171] See Ord, *Precipice* (n 16). On page 92, Ord describes this as the age of heightened risk. On page 33, he explains that the Precipice is the time (in human history) when humanity is at high risk of destroying itself.
[172] ibid 92.

anthropogenic risks began to unfold before the Trinity explosion.[173] This debate aside, it is presumed herein that humanity introduced the first non-natural extinction risk aligning with this thesis's focus less than 100 years ago.[174]

With this in mind, Equation 2.1 above exemplifies why the past cannot be used when estimating the total anthropogenic extinction risk per century. This time, it is necessary to replace 2,000 centuries with one century to (roughly) account for how long *Homo sapiens* has survived since the beginning of *the Precipice*. Using Equation 2.1 – adjusting only how many centuries *Homo sapiens* have survived so far – returns an estimate that the total risk per century is 100%. Of course, this does not mean that the total extinction risk per century is 100%. But the bottom line is that the same approach used to consider the total natural extinction risk will not work for anthropogenic risks.

Two final points should be noted before moving on. Firstly, having survived a range of postulated natural x-risks for 2,000 centuries does not guarantee humanity's continued survival in a revamped x-risk landscape where human-made risks now seem to dominate. In effect, *past survival* does not guarantee *future survival*. The discussion above merely relates to how the past can possibly be utilised to estimate total *natural* extinction risk per century.

Secondly, the influence of different observation selection effects (such as survivorship bias) should also be noted when it comes to the estimates above.[175] A detailed discussion is beyond the scope of Chapter 2. However, the following example encapsulates the idea of survivorship bias: Imagine that there are 1,000 identical planets in the universe. Within 2,000 centuries, humanity has gone extinct on 999 of these. The survivors on the fortunate planet begin to estimate how likely it is that they are to go extinct. Unaware of humanity's fate on the other 999 planets, they take great

[173] E.g., Torres writes that global warming began prior to the Atomic Age. Torres, *The End* (n 35) 141.
[174] See the research parameters in Section 1.3.3.
[175] Leslie, *The End of the World* (n 101) 139-141 on observation selection; Ord, *Precipice* (n 16) 84 on survivorship bias; Ćirković, Sandberg and Bostrom write: "…[S]ome ER probabilities derived from past records are unreliable due to the presence of observation selection effects. Anthropic bias…can lead to underestimation of the probability of a range of catastrophic events." Milan Ćirković, Anders Sandberg and Nick Bostrom, 'Anthropic Shadow: Observation Selection Effects and Human Extinction Risks' [2010] Risk Analysis 1495, 1496.

comfort in their survival and longevity.[176] They conclude that the past demonstrates that future survival is far more likely than not.

2.3.2 Anthropogenic Risks

Anthropogenic risks are those arising from the activities of humanity. It is said that many of the most pressing x-risks can be traced to anthropogenic processes.[177] "[M]ost of the biggest existential risks…" writes Bostrom, "…seem to be linked to potential future technological breakthroughs that may radically expand our ability to manipulate the external world or our own biology."[178] Bostrom reached the same conclusion in his earlier article. Having considered numerous existential risks, he said "…the top risks are engendered by our activities."[179] The upshot of this recognition is that humanity itself can play a decisive role in addressing and regulating some of the anthropogenic existential risks that scholars have identified.[180] To cite the 2020 Human Development Report:

> [T]he rise in anthropogenic risk means that most of the existential risk we face likely arises from our own actions. While this is a disturbing trend, there is a flip side that should give us hope: Humanity's future is largely within humanity's control…[T]he risks from nuclear war, climate change and engineered pandemics arise from activities that humans perform—and thus that humans can stop.[181]

Various scenarios and hazards can be explored within the boundaries of anthropogenic existential risk. Posner, for example, divides human-made

[176] Example inspired by Ord, *Precipice* (n 16) 84.
[177] Kristian Cedervall Lauta, 'Dysfunction and Disruption: Disaster Risks and the Law' [2017] European Journal of Risk Regulation 83, 87; Bostrom, 'Existential Risk Prevention' (n 15) 15; Posner also suggests many catastrophic risks are the product of science and its technological application. Posner, *Catastrophe* (n 90) 8.
[178] Bostrom, 'Existential Risk Prevention' (n 15) 16.
[179] Bostrom, 'Existential Risks' (n 15).
[180] McKinnon writes: "…[I]t is within our control (at least in theory) to address dangers of extinction with anthropogenic as opposed to non-anthropogenic causes, in large part because we are able to create these dangers through our own conduct…" Catriona McKinnon, 'Endangering Humanity: An International Crime?' [2017] Canadian Journal of Philosophy 395, 396.
[181] United Nations Development Programme, 'Human Development Report' (2020) 110.

catastrophes into three groups; Scientific *accidents*,[182] other *unintended* man-made catastrophes[183] and *intentional* catastrophes.[184] Posner's categorisation underscores the fact that individuals can bring about catastrophe both via error and terror.[185]

Similarly, in classifying risks by origin, Cotton-Barratt, Daniel and Sandberg, differentiate *accident* and *malicious* risks.[186] A distinguishing factors is intentionality. They explain that the prospect of a rogue group creating and releasing a deadly virus is a malicious risk. The prospect that an engineered pathogen escapes a laboratory despite safety precautions is an accident risk.[187] The error or terror distinction may be useful when formulating legal interventions for x-risk mitigation. Chapter 1 made it clear that the LHC-inspired intervention is concerned with error-type risks.

The upcoming Sub-sections will consider some anthropogenic x-risks and how these are associated with different hazards. Outlining some of the more common hazards is integral in mapping out the x-risk landscape within which a *broad* legal intervention will sit and operate. Of course, broad legal interventions can be more or less wide. In other words, intervention X might be 'broader' than intervention Y.[188]

(i) Nuclear War

It has been said that humanity owe much gratitude to Vasili Arkhipov, a Russian officer who reportedly "…single-handedly prevented nuclear war during the height of the Cuban Missile Crisis."[189] Nuclear war is a dreadful prospect. The explosions will be devastating and end many lives directly,[190] but they are not the immediate cause of a

[182] An example discussed here includes 'the strangelet scenario' whereby the collision of particles in particle accelerators could lead to an end-of-the-world scenario. Posner, *Catastrophe* (n 90) 30-43.
[183] For instance, Posner discusses global warming here. ibid 43-58.
[184] Examples discussed under this heading include a 'nuclear winter' and bioterrorism. ibid 71-84.
[185] See also Rees, 'Denial of Catastrophic Risks' (n 38).
[186] Cotton-Barratt, Daniel and Sandberg, 'Defence' (n 37) 272-274.
[187] ibid.
[188] This is explored in Chapter 5.
[189] Tucker Davey, '55 Years After Preventing Nuclear Attack, Arkhipov Honored With Inaugural Future of Life Award' (Future of Life Institute 2017) <https://futureoflife.org/2017/10/27/55-years-preventing-nuclear-attack-arkhipov-honored-inaugural-future-life-award/> accessed 24 April 2023; See also the discussion in Ord, *Precipice* (n 16) 4-6.
[190] Westin and others estimate that 150,000 deaths were caused by the by the nuclear bomb in Hiroshima in 1945. They write that out of those individuals, around 70,000 died within a day of the

potential extinction scenario.[191] As summarised by Rowe and Beard, the *existential* risk stems from a potential *nuclear winter*.[192] They further explain that only a subset of nuclear wars may be able to bring about such a scenario.[193] The most severe consequence of the war is said to be the dust, smoke and soot from, among other things, burning cities which would block the sunlight.[194] The resulting cold and darkness could interrupt much of the world's vegetation and animal life, and billions of people may die owing to starvation.[195] As recognised by Kemp and Rhodes, the consequences that humanity needs to prepare for are similar when it comes to asteroid impacts, super-volcanic eruptions and nuclear winter.[196]

While nuclear war will have grave consequences, it is not possible to conclude that nuclear winter scenarios will inevitably trigger an *existential* catastrophe.[197] In fact, Ord highlights contemporary research, all of which indicate that human extinction is unlikely to flow from nuclear winter scenarios.[198]

(ii) Geoengineering

Geoengineering can be defined as the "…deliberate large-scale intervention in the Earth's climate system, in order to moderate global warming."[199] Geoengineering can, thus, help to address anthropogenic climate change which is another potential x-risk.

---

explosion. See Ulrika Westin and others, 'Global Catastrophic Risks 2020' (Global Challenges Foundation Annual Report 2020) <https://globalchallenges.org/wp-content/uploads/Global-Catastrophic-Risks-2020-Annual-Report.pdf> accessed 24 April 2023, 7.

[191] One can divide the damage from nuclear war into the damage from (i) the blast, fire and radiation, and (ii) a potential nuclear winter scenario. See Cotton-Barratt and others, 'Global Catastrophic Risks' (n 150) 36.

[192] Thomas Rowe and Simon Beard, 'Probabilities, Methodologies and the Evidence Base in Existential Risk Assessments' (Centre for the Study of Existential Risk 2018) <http://eprints.lse.ac.uk/89506/> accessed 25 April 2023, 27.

[193] ibid.

[194] Ord, *Precipice* (n 16) 98; Westin and others, 'Global Catastrophic Risks 2020' (n 190) 8.

[195] ibid; Ord, *Precipice* (n 16) 98.

[196] Kemp and Rhodes, 'The Cartography of Global Catastrophic Governance' (n 53) 7.

[197] See Farquhar and others, 'Existential Risk' (n 61) 7.

[198] Ord, *Precipice* (n 16) 99, 336.

[199] The Royal Society, 'Geoengineering the Climate: Science, Governance and Uncertainty' (2009) <https://royalsociety.org/topics-policy/publications/%202009/geoengineering-climate/> accessed 25 April 2023, ix and 1; After discussing the scientific background of how we are changing the climate, Häggström considers means under the heading 'geoengineering' for solving and preventing dangerous climate change. Häggström, *Here Be Dragons* (n 22) 25-33.

However, some have asked whether geoengineering poses a greater x-risk than climate change.[200]

Geoengineering methods can be divided into two broad classes: (a) Carbon Dioxide Removal (CDR) techniques and (b) Solar Radiation Management (SRM) techniques.[201] CDR and SRM methods have the same aim, namely, to reduce global temperatures.[202] Researchers have considered SRM methods ranging from brightening the Earth's surface to injecting, for instance, sulphate aerosols into the stratosphere.[203] The latter method, 'Stratospheric Aerosol Injection' (SAI), is according to Halstead the most discussed form of SRM.[204] Halstead explains that "[o]ne of the main risks of SAI stems from the fact that it could be terminated suddenly causing rapid and damaging warming."[205] He refers to this as 'termination shock'. Baum, Maher and Haqq-Misra explore this notion and suggest that a worst-case scenario could plausibly involve human extinction.[206] Nevertheless, Halstead highlights that the risk of termination shock is possibly overstated and relatively small.[207] As with nuclear war, it is premature to conclude that some deliberate large-scale interventions will cause an x-catastrophe.

(iii) Pandemics and Biotechnology

The Black Death is singled out as one of the deadliest catastrophes in human history.[208] Kilbourne writes that the total number of deaths is estimated at 75 million

---

[200] Ord, *Precipice* (n 16)112-113.

[201] CDR techniques seek to remove CO2 from the atmosphere; SRM techniques seek to reflect a percentage of the sun's light and heat back into space. The Royal Society, 'Geoengineering the Climate' (n 199).

[202] ibid.

[203] ibid 23-31.

[204] John Halstead, 'Stratospheric Aerosol Injection Research and Existential Risk' [2018] Futures 63, 63.

[205] ibid 68.

[206] Seth Baum, Timothy Maher and Jacob Haqq-Misra, 'Double Catastrophe: Intermittent Stratospheric Geoengineering Induced by Societal Collapse,' [2013] Environment Systems & Decisions 168.

[207] According to Halstead, *if* SAI were abruptly stopped, other countries and actors would likely be willing and able to resume SAI to avoid the consequences. Halstead, 'Stratospheric Aerosol Injection Research and Existential Risk' (n 204) 68.

[208] Muehlhauser estimates that "…the deadliest event before the industrial revolution (the Black Death) killed ~9.7% of world population, and the deadliest event after the industrial revolution (the 1918 flu pandemic) killed 3.3% of world population.". Luke Muehlhauser, 'How big a deal was the Industrial

people.[209] Nevertheless, the estimates vary and depend on whether one sums the number of deaths across different pandemics all caused by *Yersinia pestis.*[210] Muehlhauser is using the term Black Death to refer to the first wave in the 14th century. With this in mind, his *lowest* and *highest* plausible estimates add up to ~21 million and 68 million deaths, respectively.[211] Relative to the world population in the 14th century, (443 million[212]) the Black Death is one of the worst human catastrophes of all times. Another example is the influenza pandemic of 1918 (the Spanish flu), which is estimated to have killed 20 – 50 million people.[213] A present-day example is Covid-19, reportedly responsible for millions of deaths around the globe.[214] On the flipside, some scholars suggest that humanity can learn a range of lessons from Covid-19 when it comes to the governance of existential risks.[215] Humanity is, no doubt, vulnerable to natural pandemics. Scholars even regard the Black Death as an embodiment of a global catastrophe.[216]

According to Posner, "…the fact that *Homo sapiens* has managed to survive every disease to assail it in the 200,000 years or so of its existence is a source of genuine comfort…" when considering extinction events.[217] Posner's statement is reminiscence of the fossil record discussion for estimating the total natural extinction risk per century.

Revolution?' <https://lukemuehlhauser.com/industrial-revolution/> accessed 25 April 2023; Cotton-Barratt and others rely on sources to make an estimate that the Black Death plague in the 14th century killed between 11% - 17% of the world population. See Cotton-Barratt and others, 'Global Catastrophic Risks' (n 150) 42.

[209] Edwin Kilbourne, 'Plagues and Pandemics: Past, Present, and Future' in Nick Bostrom and Milan Ćirković (eds), *Global Catastrophic Risks* (OUP 2008) 295.

[210] Muehlhauser, 'How big a deal was the Industrial Revolution?' (n 208).

[211] ibid; Furthermore, Ord's *lower* and *upper* bound assumptions add up to 24 million and 63 million deaths, respectively. Ord, *Precipice* (n 16) 350.

[212] While there are different estimates, this thesis has opted to cite the United States Census Bureau. United States Census Bureau, 'Historical Estimates of World Population' (Last Revised: 5 December 2022) <https://www.census.gov/data/tables/time-series/demo/international-programs/historical-est-worldpop.html> accessed 25 April 2023.

[213] Kilbourne, 'Plagues and Pandemics' (n 209) 290; One should note, however, that estimates vary here too.

[214] World Health Organisation, 'WHO Coronavirus Disease (COVID-19) Dashboard' <https://covid19.who.int/> accessed 25 April 2023.

[215] See Hin-Yan Liu, Kristian Lauta and Matthijs Maas, 'Apocalypse Now? Initial Lessons from the Covid-19 Pandemic for the Governance of Existential and Global Catastrophic Risks' [2020] Journal of International Humanitarian Legal Studies 295.

[216] Cotton-Barratt and others recognise that the Black Death is a global catastrophe. See Cotton-Barratt and others, 'Global Catastrophic Risks' (n 150) 22.

[217] Posner, *Catastrophe* (n 90) 23.

However, it is not flawless. While the fossil record provides a case against a high extinction risk from natural pandemics, it has to be recognised that the risk landscape is different in modern times.[218] Ord and other scholars suggest that humanity can amplify the overall risk.[219] The world is more interconnected (feeding into the spread of pathogens), has a larger population, and there is an increased interaction between humans and animals (feeding into the origin of new diseases).[220] On the other hand, there is now better sanitation and hygiene, scientific understanding, medicine, and tools of communication.[221] It is unclear whether the net effects are positive or negative, but it cast doubt on using the longevity of *Homo sapiens* as a source of comfort when dealing with pandemics in the 21st Century. As written by Leslie in 1996, "[t]he fact that the human species has survived past diseases might…be unimpressive in view of today's new conditions…"[222]

This thesis addresses pandemics in the Section on 'anthropogenic' risks due to the developments in biotechnology,[223] allowing for the design and creation of pathogens. As explained by Cotton-Barratt and others, there is often an inverse relationship between a pathogen's lethality and transmissibility—a highly lethal pathogen will fail to spread far before killing its host.[224] They write that biotechnology can offset this correlation.[225] Scholars suggest that the risk when it comes to biotechnology stems from a potential laboratory accident (biosafety issue) or the intended misuse

[218] Ord, *Precipice* (n 16) n126-127.
[219] ibid; Bostrom and Ćirković, 'Introduction' (n 12) 16; Kilbourne, 'Plagues and Pandemics' (n 209) 293, 302-303.
[220] Ord, *Precipice* (n 16) 126-127; Bostrom and Ćirković, 'Introduction' (n 12) 16; Cotton-Barratt and others, 'Global Catastrophic Risks' (n 150) 42.
[221] Ord, *Precipice* (n 16) 127; There are also bodies like the WHO. The website states "Our primary role is to direct and coordinate international health within the United Nations system." See WHO, 'About WHO' <https://www.who.int/about> accessed 25 April 2023.
[222] Leslie, *The End of the World* (n 101) 141.
[223] See e.g., Ali Nouri and Christopher Chyba, 'Biotechnology and Biosecurity' in Nick Bostrom and Milan Ćirković (eds), *Global Catastrophic Risks* (Oxford University Press 2008).
[224] Cotton-Barratt and others, 'Global Catastrophic Risks' (n 150) 52.
[225] Ibid.

(bioterrorism).[226] Regarding the former, there have reportedly been past mishaps which highlight the inadequacy of laboratory safety for dangerous pathogens.[227]

It is not clear whether an engineered pathogen could cause an existential catastrophe. Farquhar and others write that making a pathogen which could derail humanity's long-term future is close to impossible at present. Yet, they recognise that humans may "…eventually gain the capacity to create pathogens which could deliberately or accidentally cause an existential catastrophe."[228]

(iv) Artificial General Intelligence

Scholars often talk about AI as a potential hazard in the x-risk landscape.[229] Popular figures including Stephen Hawking, Elon Musk and Bill Gates have expressed concerns in the media about the risks associated with AI.[230]

However, the media occasionally misplace the primary concern. As explained by Tegmark, many AI researchers roll their eyes when seeing extravagant headlines about evil AI systems and killer-robots.[231] The concern when it comes to AI relates to *competence*, not *malevolence*.[232] Whereas today's AI is often seen as *narrow* (i.e.,

---

[226] Wilson explores the risks arising from the (1) accidental release of harmful organisms, and (2) malicious release of such organisms. Wilson, 'Minimizing' (n 151) 317; Martin Rees, *Our Final Century: Will Civilisation Survive the Twenty-First Century?* (Arrow Books 2003) 57.

[227] See Ord, *Precipice* (n 16) 129-131. On page 131, Ord lists five examples of 'notable laboratory escapes' including an accidental smallpox release in 1971 and 1978; Wilson, 'Minimizing' (n 151) 318-320.

[228] Farquhar and others, 'Existential Risk' (n 61) 9.

[229] Ord, *Precipice* (n 16) 138; Farquhar and others, 'Existential Risk' (n 61) 9; Bostrom, 'Existential Risks' (n 15); Wilson, 'Minimizing' (n 151) 329; Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (OUP 2014) ch 8; Allan Dafoe and Stuart Russell, 'Yes, We are Worried About the Existential Risk of Artificial Intelligence' (2016) MIT Technology Review <https://www.technologyreview.com/s/602776/yes-we-are-worried-about-the-existential-risk-of-artificial-intelligence/> accessed 25 April 2023.

[230] See Stuart Dredge, 'Artificial Intelligence will Become Strong Enough to be a Concern, says Bill Gates', *The Guardian* (29 January 2015) <https://www.theguardian.com/technology/2015/jan/29/artificial-intelligence-strong-concern-bill-gates> accessed 25 April 2023.

[231] Max Tegmark, 'Benefits & Risks of Artificial Intelligence' (Future of Life Institute 2015) <https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/> accessed 25 April 2023.

[232] ibid.

designed to perform narrow tasks like face recognition and driving cars),[233] *existential* risk is commonly associated with future *general* AI ('AGI').

According to Ord, the most plausible x-risk stems from creating agents with a general intelligence surpassing human-level intelligence.[234] In turn, AGI can possibly lead to superintelligence,[235] defined by Bostrom as "…any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest."[236]

A specific risk scenario involves the alignment problem.[237] Hawking explains that "[a] super-intelligent AI will be extremely good at accomplishing its goals, and if those goals aren't aligned with ours we're in trouble."[238] Consider the following example provided by Bostrom: "[W]e could mistakenly elevate a subgoal to the status of a supergoal. We tell it to solve a mathematical problem, and it complies by turning all the matter in the solar system into a giant calculating device..."[239] Whereas the exact practical processes by which this scenario could materialise are unclear, the point is that the alignment problem is a significant issue to address.

Ord examines an alternative risk-scenario which involves an AI system seizing control of humanity's future: Contrary to possible misconceptions, such an undertaking does not require robots.[240] Instead, the AI system may exploit the internet, financial resources and its superior intelligence to manipulate people to do its bidding.[241] He justify the theoretical possibility by noting that Hitler and Stalin exploited their (human-

[233] ibid.
[234] Ord, *Precipice* (n 16) 141.
[235] If humans could create AGI, such AGI could create higher intelligence which could create yet higher intelligence and so on. See Vincent Müller and Nick Bostrom, 'Future Progress in Artificial Intelligence: A Survey of Expert Opinion' in Vincent Müller (eds), *Fundamental Issues of Artificial Intelligence* (Springer 2016).
[236] Bostrom, *Superintelligence* (n 229).
[237] As summarised by Liu, Lauta and Maas, the problem is often said to be one of 'value alignment' where the discrepancy between humanity's interest and that of the superintelligence lead to the demise of humanity. Liu, Lauta and Maas, 'Governing Boring Apocalypses' (n 61) 8.
[238] Stephen Hawking, *Brief Answers to the Big Questions* (John Murray Publishers 2018) 188.
[239] Bostrom, 'Existential Risks' (n 15); See the paperclip maximiser argument in Nick Bostrom, 'Ethical Issues in Advanced Artificial Intelligence' <https://www.fhi.ox.ac.uk/wp-content/uploads/ethical-issues-in-advanced-ai.pdf> accessed 25 April 2023; Nindler, 'The United Nation's Capability' (n 54) 16.
[240] Ord, *Precipice* (n 16) 146.
[241] ibid 146-147.

level) intelligence to control millions of people to win physical conflicts.[242] Thus, a *super-intelligent* entity may, in ways currently unforeseeable, come to orchestrate large parts of our story and ultimate fate. Of course, the outcome of such a scenario needn't, all things considered, be undesirable.[243]

The x-risk scenarios associated with AI are highly speculative. Ord recons that the case for existential risk from AI is the most speculative one in his book.[244] For instance, there is disagreement about *if* and *when* AGI and superintelligence may become a reality.[245] It is also unclear whether humanity should fear or welcome the advent of such intelligence.[246]

(v) Unknown Unknowns

In 2002, Secretary of Defense Donald Rumsfeld made the following statement:

> [T]here are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns – the ones we don't know we don't know.[247]

While the statement has been criticised as nonsense by some critics, others suggest that it makes sense and have even shown that research sometimes uncover unknown unknowns.[248] More to the point, the idea of unknown unknowns is relevant in charting the x-risk landscape.[249] It is unwise to think that humanity has identified all future hazards that can be linked to anthropogenic x-risks.

---

[242] ibid.
[243] ibid.
[244] ibid 149.
[245] Häggström, *Here Be Dragons* (n 22) 106.
[246] Tegmark, 'Benefits & Risks of Artificial Intelligence' (n 231).
[247] Donald Rumsfeld (DoD News Briefing, 12 February 2002) <https://archive.ph/20180320091111/http://archive.defense.gov/Transcripts/Transcript.aspx?TranscriptID=2636> accessed 25 April 2023.
[248] David Logan, 'Known Knowns, Known Unknowns, Unknown Unknowns and the Propagation of Scientific Enquiry' [2009] 712, 712.
[249] See Torres, *The End* (n 35) ch 10.

Scholars have noted that future technological or scientific developments may reveal new ways of destroying the world.[250] As explained by Farquhar and others, it was not anticipated by many at the beginning of the 20th century that nuclear weapons, engineered pandemics and AI would be among the most severe existential risks.[251] Rees explores a similar theme in his book. He demonstrates the difficulties in forecasting future discoveries, developments and transformations.[252] Chapter 5 in Ord's book also deals with this notion to some extent.[253] In essence, it is unwise to assert that new technologies are impossible.[254] In *The End: What Science and Religion Tell Us about the Apocalypse*, Torres considers the relevance of unknown unknowns and writes: "Perhaps if this book were written in 2100, it would be ten times as long and contain a whole different and completely novel set of existential risks."[255] Unknown unknowns may or may not take the appearance of scientific endeavours. As noted by Rees in 2003:

> Over an entire century, we cannot set limits on what science can achieve, so we should leave our minds open, or at least ajar, to concepts that now seem on the wilder shores of speculative…[A]stonishing advances could eventually stem from fundamentally new concepts in basic science that haven't yet even been envisioned and which we as yet have no vocabulary to describe.[256]

(vi) Scientific Endeavours & Physics Disasters

Chapter 3 examines the LHC Controversy in detail. For now, it is sufficient to appreciate that scientific undertakings, like the Manhattan Project and high-energy physics experiments, can be considered alleged hazards within the x-risk landscape. Another notable example is the CUORE experiment, whereby scientists cooled a copper vessel to -273.144 degrees Celsius, making it, for 15 days, the coldest cubic

---

[250] Bostrom, 'Existential Risks' (n 15); Torres, *The End* (n 35) 143.
[251] Farquhar and others, 'Existential Risk' (n 61) 10.
[252] Rees highlights an example involving the US National Academy of Sciences. In 1937, NAS organised a study which sought to predict breakthrough. The study did not predict some remarkable breakthroughs including nuclear energy, antibiotics, computers, or jet aircrafts. Rees, *Our Final Century* (n 226) 12-15.
[253] Ord, *Precipice* (n 16) 121-123.
[254] ibid 122.
[255] Torres, *The End* (n 35) 143-144.
[256] Rees, *Our Final Century* (n 226) 15-16.

meter in the universe.[257] Furthermore, as noted above, future scientific endeavours that may be associated with x-risk(s) may currently reside in unknown unknowns.

Bostrom considers the Manhattan Project and high-energy particle accelerator experiments under the heading "physics disasters", whereby the world (and humanity's future) could end in a bang.[258] However, most experts now believe that neither the Manhattan Project nor experiments facilitated by particle accelerators, including the LHC, posed any x-risk. Despite this, Chapter 1 explained that the LHC Controversy can offer valuable lessons for x-risk mitigation irrespective of whether particle accelerator experiments pose any risk. Bostrom also writes:

> The main reason for concern in the "physics disasters" category is the meta−level observation that discoveries of all sorts of weird physical phenomena are made all the time, so even if right now all the particular physics disasters we have conceived of were absurdly improbable or impossible, there could be other more realistic failure−modes waiting to be uncovered.[259]

The Sub-sections above have considered some of the hazards commonly associated with anthropogenic x-risk. These Sections have helped this thesis to build and convey a more holistic picture of the risk landscape within which and across a broad legal intervention will operate. In addition, having considered the literature dealing with the hazards, the thesis can document some general observations that pertain to interventions aiming to operate across the x-risk landscape. These observations are outlined further below.

2.3.3 General Observations

(i) Benefit & Risk

So far, the thesis has emphasised the risk aspect of the 'hazards' above. However, it is untenable to overlook the benefits which can flow from scientific and technological progress. This thesis is in favour of such progress, but it aims not to sway the reader

[257] Interactions, 'CUORE: The Coldest Heart in the Known Universe' (2014) <https://www.interactions.org/press-release/infn-cuore> accessed 25 April 2023.
[258] Bostrom, 'Existential Risks' (n 15).
[259] ibid.

that the benefits generally outshine the risks or vice versa. The point is that many of the hazards are associated with great benefits.[260] For instance, biotechnology finds application in human health to name one example.[261] It can combat infectious diseases,[262] and as explained by Davey, when Jenner invented vaccines, he used the power of biotechnology: So too did Fleming when discovering antibiotics.[263] Davey further highlights that biotechnology is behind hit drugs which treat cancer and heart disease.[264]

Several benefits are also associated with particle accelerators like the LHC. The most obvious benefit may be that they allow humanity to gain a deeper understanding of the universe. It is also possible to highlight more concrete benefits associated with particle accelerators. For one, their creation has led to various spinoff technologies.[265] As written by Larsen: "Meeting the technological challenges posed by building high-energy accelerators…requires innovation and creativity that have applications to manufacturing, planning, and communications processes."[266]

CERN also boasts that its "…engineers, technicians and scientists develop novel technologies and expertise contributing to applications in fields beyond high-energy physics, addressing global societal challenges in areas such as health and environment."[267] Furthermore, Piccirillo considers how science is a valuable

[260] The main exception to this statement is possibly that of nuclear *war.* This thesis will not discuss the notion of nuclear deterrence, and whether it helps to ensure global peace. For a discussion see, Westin and others, 'Global Catastrophic Risks 2020' (n 190) 10-11.
[261] Tucker Davey, 'Benefits & Risks of Biotechnology' (Future of Life Institute 2018) <https://futureoflife.org/background/benefits-risks-biotechnology/> accessed 25 April 2023.
[262] Organisation for Economic Co-Operation and Development, 'Biotechnology and Sustainability: The Fight Against Infectious Disease' (2003) <https://www.oecd.org/sti/emerging-tech/2508407.pdf> accessed 25 April 2023.
[263] Davey, 'Benefits & Risks of Biotechnology' (n 261).
[264] ibid.
[265] See e.g., Kristine Larsen, *Particle Panic!: How Popular Media and Popularized Science Feed Public Fears of Particle Accelerator Experiments* (Springer 2019) 173-177; Carsten Welsch, 'What Have Particle Accelerators Ever Done For Us?' *Physics World* (20 August 2019) <https://physicsworld.com/a/what-have-particle-accelerators-ever-done-for-us/> accessed 26 April 2023; Graeme Burt, 'Five Ways Particle Accelerators Have Changed the World (Without a Higgs Boson in Sight)' *The Conversation* (8 February 2016) <https://theconversation.com/five-ways-particle-accelerators-have-changed-the-world-without-a-higgs-boson-in-sight-54187> accessed 26 April 2023.
[266] Larsen, *Particle Panic!* (n 265) 174.
[267] CERN, 'Contribute to Society' <https://home.cern/about/what-we-do/our-impact> accessed 26 April 2023.

enterprise; using the LHC as an example of big science, he explains its importance and added value to society.[268] A notable achievement of the LHC is the detection of the Higgs boson – aka the God Particle.[269] Through this accelerator, says Piccirillo, we have expanded our knowledge of fundamental physics and effectively expanded human horizons.[270] Another way in which the LHC undertaking can expand horizons is by improving collaboration among states, overcome ideological and cultural boundaries.[271] CERN states that, as of 2017, over 17,500 people from around the world are working together in an attempt to push the limits of knowledge.[272]

AI is another example which can allegedly improve efficiency in most or all industry sectors.[273] Self-driving vehicles will reportedly "…eliminate one of the biggest causes of accidental death and injury in [the] United States…"[274] Other benefits include AI helping to detect cancer, prevent aeroplane collisions[275] and reduce risks in the private security industry.[276] If humanity get things right, the potential advent of AGI and superintelligence may perhaps allow humanity to 'survive' and safely navigate the x-risk landscape. While mindful of the potential downsides of AGI, Tardif writes: "The

[268] Lucio Piccirillo, 'Big Science and Small Science: Reflections on the Relationship Between Science and Society from the Perspective of Physics' in Simona Giordano (ed), *The Freedom of Scientific Research: Bridging the Gap between Science and Society* (Manchester University Press 2019).
[269] Michael Greshko, 'Elusive Elementary Particle Finally Caught Decaying—Get the Facts' *National Geographic* (29 August 2018) <https://www.nationalgeographic.com/science/2018/08/higgs-boson-decay-quarks-lhc-standard-model-physics/> accessed 26 April 2023.
[270] Piccirillo, 'Big Science and Small Science' (n 268) 133-135.
[271] ibid 135.
[272] CERN, 'Our People' <https://home.cern/about/who-we-are/our-people> accessed 26 April 2023.
[273] Dame Wendy Hall and Jérôme Pesenti, 'Growing the Artificial Intelligence Industry in the UK' (Independent report, the Department for Digital, Culture, Media & Sport and the Department for Business, Energy & Industrial Strategy 2017) <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf> accessed 26 April 2023.
[274] Peter Stone and others, 'Artificial Intelligence and Life in 2030' (One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, Stanford University 2016) <https://ai100.stanford.edu/sites/default/files/ai100report10032016fnl_singles.pdf> accessed 06 November 2020, 2; See also Sven Beiker, 'Legal Aspects of Autonomous Driving' [2012] Santa Clara Law Review 1145.
[275] Amitai Etzioni and Oren Etzioni, 'Should Artificial Intelligence be Regulated?' (2017) 33(4) *Issues in Science and Technology* <http://issues.org/33-4/perspective-should-artificial-intelligence-be-regulated/> accessed 06 November 2020.
[276] Royal Academy of Engineering, 'Innovation in Autonomous Systems' (Summary of an event held at the Royal Academy of Engineering, 2015) 7 <https://www.raeng.org.uk/publications/reports/innovation-in-autonomous-systems> accessed 06 November 2020.

potential benefits of releasing an AGI are astronomical."[277] Indeed, AGI may lead to superintelligence which "…may be the last invention humans ever need to make."[278] Bostrom further notes that "[i]t is hard to think of any problem that a superintelligence could not either solve or at least help us solve."[279]

The risk/benefit trade-off is true more broadly when it comes to scientific and technological progress.[280] While the benefits outlined above are not exhaustive, they back the claim that legal interventions may need to navigate risk and benefit.

For example, Cass Sunstein writes about "miracles" in his book: Miracles, he writes, can be "… understood as extreme upsides, making human life immeasurably better."[281] He notes that regulation may reduce the probability of miracles,[282] and highlights that there may be "catastrophe-miracle" tradeoffs.[283] He writes:

> Reasonable regulators might want to prevent a possible catastrophe, even if the price is to prevent a possible miracle. The downside risk of (say) extinction might reasonably be seen to deserve more attention that the upside potential of (say) immortality.[284]

This thesis will not use the term "catastrophe-miracle" tradeoffs. It resorts to the term 'existential trade-off' which describes a situation where the taking of x-risk (A) may help humanity mitigate x-risk (B) that is potentially more likely than risk (A).

The crucial observation is that broad legal interventions will sit within a risk landscape where the hazards are associated with astronomical downsides and possibly astronomical benefits, such as potential existential trade-offs. Nevertheless,

---

[277] Antoine Tardif, 'How we can Benefit from Advancing Artificial General Intelligence (AGI)' *Unite.AI* (27 September 2020) <https://www.unite.ai/artificial-general-intelligence-agi/> accessed 26 April 2023.
[278] Bostrom, 'Ethical Issues in Advanced Artificial Intelligence' (n 239).
[279] ibid.
[280] Posner recognises that science and technology can help society to avoid man-made and natural catastrophes. See Posner, *Catastrophe* (n 90) 15; Lauta echoes a similar point and writes that the expected accelerating knowledge and technology development may provide new possibilities to mitigate certain x-risks. See Lauta, 'Dysfunction and Disruption' (n 177) 87.
[281] Cass Sunstein, *Averting Catastrophe: Decision Theory for COVID-19, Climate Change, and Potential Disasters of All Kinds* (NYU Press 2021) 2.
[282] ibid 13.
[283] ibid 26.
[284] ibid 58.

comparing the potential benefits associated with, say, AGI and particle accelerators may lead someone to say that only the former may produce astronomical benefits. For example, it seems more reasonable to believe that actualising AGI will facilitate an existential trade-off than the continued use of particle accelerators.

(ii) Probability of the Initial Catastrophe

It is often difficult to estimate the probability of the relevant unwanted events, such as the probability of a nuclear war, an AI system seizing control of humanity's future, or a physics experiment producing a synthetic and dangerous black hole.

It is fair to say that the risk-scenarios are often speculative. It is essential not to misuse speculative scenarios as an excuse to curb progress. Nouri and Chyba capture this observation in relation to biotechnology: "Any approach…which substantially curtails the utility of biotechnology to treat and counter disease, runs the risk of sacrificing large numbers of lives to head off hypothetical risks."[285] On the other hand, the speculative nature of the field should not discourage x-risk mitigation.

(iii) Probability of Extinction Following the Initial Catastrophe

It is not always clear whether a specific unwanted event – should it actually happen – would qualify as an *existential* catastrophe. For example, predicting the precise outcome should an AI system seize control of humanity's future is difficult. Furthermore, it is unclear whether nuclear winter would cause the premature extinction of humanity or permanently destroy its future potential. Many of the hazards are, in effect, linked to less serious risk.

The following question should briefly be addressed at this stage. Does it really matter whether a risk threatens the premature extinction of humanity as opposed to 99% of everyone alive? Chapter 5 will argue that the distinction matters greatly (as long as it is presumed that the remaining 1% could eventually recover). The distinction between

[285] Nouri and Chyba, 'Biotechnology and Biosecurity' (n 223) 451; Davey, 'Benefits & Risks of Biotechnology' (n 261).

the *recoverable* and *unrecoverable*, underlies this thesis. The distinction is also at the heart of the ASP, which, it is argued, can empower legal interventions aiming to mitigate x-risk.

Getting back to the fundamental observation in this Section: It is not always possible to conclusively conclude that a given unwanted event, should it happen, will mark the beginning of humanity's premature end.

(iv) Baum's Equation

The two observations above can be linked to the quantification of the probability of existential catastrophes as outlined by Baum.[286] He explains that an x-catastrophe is an event defined by its *severity* – i.e., it is defined in terms of a minimum severity threshold.[287] To avoid confusion with Bostrom's typology of a risk's seriousness, which also uses the term 'severity',[288] this thesis uses the term 'harms' threshold. The applicable harms threshold in this thesis is the premature extinction of humanity given that it defines an existential catastrophe as the premature extinction of humanity.

To determine the probability of an x-catastrophe, it is necessary to look at two things: (1) The probability of the primary catastrophic (unwanted) event and, (2) the probability that this event will cause harm exceeding the applicable harms threshold. Baum uses the following equation to explain this notion:

$$P_{EC}=P_1 * P_2$$

Baum explains that P_{EC} is the probability of x-catastrophe; P_1 is the probability of the initial catastrophe event. For instance, the probability of a nuclear war, an asteroid hitting Earth, an AI system taking control of humanity's future, or a dangerous black hole being produced by a physics experiment.

[286] Baum, 'Quantifying the Probability' (n 125).
[287] ibid.
[288] See Table 2.2.

P_2 is the probability that the severity of that event will exceed the harms threshold. For instance, the probability that the nuclear war, asteroid impact, the influence of the AI system, or the synthetic black hole will lead to humanity's premature extinction.[289] Thus, when calculating P_{EC} via Baum's equation, it is necessary to employ two distinct analyses – one for P_1 and another one for P_2.[290] Baum's explanation encapsulates the observations above. The observation in Sub-section 2.3.3 (ii) can be linked to P_1. The observation in Sub-section 2.3.3 (iii) can be linked to P_2.

(v) The Probabilities Differ Across the Landscape

With Baum's equation in mind, the thesis can document another observation related to probabilities. Namely, the risks within the landscape often differ significantly in their probability estimates. That some risks are more or less likely than others is not surprising.[291] However, within the x-risk landscape, this feature is particularly noteworthy. As written by Ord, "[o]ne of the most striking features of [the] risk landscape is how widely the probabilities vary between different [existential] risks. Some are a million times more likely than others…"[292]

The above can be put into context using Baum's equation. For example, consider AGI and particle accelerators in terms of P_1. This may involve analysis of, for instance, the probability of an AI system taking control of humanity's future[293] or a black-hole disaster owing to high-energy physics experiments. After such analysis, a reasonable conclusion is that these two disaster events are not equally probable. Or, at least, it is clear that they are not seen as equally probable by others.

The probabilities also vary regarding the likelihood that the initial catastrophe will exceed the severity threshold. Compare climate change and a black hole disaster as

[289] Baum, 'Quantifying the Probability' (n 125).
[290] ibid.
[291] Beard, Rowe and Fox, 'An Analysis and Evaluation of Methods Currently Used to Quantify the Likelihood of Existential Hazards' (n 124) Appendix A; Ord, *Precipice* (n 16) 167; Anders Sandberg and Nick Bostrom, 'Global Catastrophic Risks Survey', (Future of Humanity Institute – Technical Report #2008-1) <https://www.fhi.ox.ac.uk/reports/2008-1.pdf> accessed 26 April 2023.
[292] Ord, *Precipice* (n 16) 168.
[293] Ord focuses on this because he finds it "…the most plausible existential risk from AI." ibid 148.

an example. As Baum explains, it is virtually certain that climate change is occurring.[294] Thus, $P_1 \approx 1$. The main uncertainty in this context revolves around P_2 and "…how severe the impacts will end up being…"[295] The inverse pattern arises when considering P_1 and P_2 and high-energy physics experiments. This time, the uncertainty concerns P_1 since it is said that the experiments facilitated by machines, including the LHC and RHIC, do not pose *any* existential risk. On the other hand, it appears virtually certain that a black hole disaster, if it should, would amount to an existential catastrophe. Thus, $P_2 \approx 1$.[296]

(vi) Governance Regime – A Selective Approach

The existing governance regimes for the hazards vary significantly. Kemp and Rhodes map the international governance regime for various hazards, including many of those considered above.[297] Their cartography provides a useful starting point when seeking an overview of the related governance regimes. According to Kemp and Rhodes, the legal landscape is sometimes "…littered more with gaps than effective policy…"[298] Together with an overview of each governance regime, they indicate the relative significance of the gaps.[299] An important takeaway is that they are not equally significant and pressing. For example, the authors demonstrate that the governance gap is more significant in respect of AI than when it comes to nuclear warfare. The cartography also demonstrates that the hazards can have 'internal' gaps. For example, a current treaty may, at least in theory, encompass LAWs; in contrast, the development of AGI is ungoverned.[300]

Boyd and Wilson also demonstrate how some x-risks have been discussed more than others at the international governance level.[301] They emphasise this selective neglect

[294] Baum, 'Quantifying the Probability' (n 125).
[295] ibid.
[296] Baum also notes that it might not be possible to survive a physics catastrophe. ibid.
[297] Kemp and Rhodes, 'The Cartography of Global Catastrophic Governance' (n 53).
[298] ibid 2.
[299] ibid 30-31.
[300] E.g., when it comes to LAWs, Kemp and Rhodes highlight that the 'Current Coverage in International Law' includes the Convention on Certain Conventional Weapons. In contrast, for AGI, the 'Current Coverage in International Law' is zero. ibid 4-6, 35.
[301] Boyd and Wilson, 'Existential Risks to Humanity' (n 154).

by exploring how x-risk has been discussed in documents in the UN Digital Library.[302] For example, they report that nuclear war is the most frequently discussed existential threat to humanity. In contrast, their search indicates that there are "key gaps" when it comes to, for example, AI and biotechnology. Furthermore, they found no mention of the risk from high-energy physics experiments.[303]

A question to consider is whether the perceptions of P_1 impact the governance regimes within the x-risk landscape. In other words, is there a correlation between P_1 and potential governance gaps so that the higher the perceived probability of the initial catastrophe, the more attention the relevant hazard will receive? This does not seem unlikely. For example, Posner writes that if the "…danger of the new accelerators [were] perceived to be serious, there probably would be no great difficulty in negotiating an international agreement to rescind or delay projects such as the LHC or RHIC–II."[304] This thesis will not assess whether the effectiveness of the governance regimes corresponds with the estimated probability of initial catastrophe. However, it appears reasonable to think that the wider perceptions of P_1 can impact the scope of future governance regimes and gaps within the x-risk landscape. Of course, it is necessary to prioritise mitigation efforts, and one method to filter the lesser deserving hazards is to consider P_1.

The basic message in this Sub-section is that some hazards receive (and will perhaps continue to receive) more attention than other. The Sub-section below will consider the lack of a *specific* international governance regime in the context of x-risk from scientific undertakings such as the collisions facilitated by the LHC.

(vii) International Law and Scientific Undertakings

According to Voeneky there is no current international treaty on x-risks and scientific research.[305] However, it does not follow that international law is irrelevant. There are

[302] Boyd and Wilson searched for eight possible existential threats: (1) Nuclear War; (2) AI; (3) Synthetic Biology; (4) Geoengineering; (5) Nanotechnology; (6) Asteroid/Comet Impacts; (7) Supervolcanic Eruption; (8) Experimental Physics Disaster. ibid
[303] ibid 2305-2307.
[304] Posner, *Catastrophe* (n 90)130.
[305] Silja Voeneky, 'Human Rights and Legitimate Governance of Existential and Global Catastrophic Risks' in Silja Voeneky and Gerald Neuman (eds), *Human Rights, Democracy, and Legitimacy in a*

treaties applicable to activities and areas of science that scholars have linked to x-risk.[306] Examples of such areas include biotechnology and geoengineering.[307] Regarding the former, Rhodes, explains that there are around 40 international regulations which apply to biotechnology in general.[308]

A destructive force from the "darker side"[309] of physics is also on the radar of international governance. That force is nuclear weapons, with physicists having played a vital role in their creation.[310] The Comprehensive Nuclear-Test-Ban Treaty prohibits "…any nuclear weapon test explosion or any other nuclear explosion…"[311] In effect, this Treaty aims to prohibit a category of experiments that many deem dangerous and undesirable.[312] By their very nature, however, some novel experiments will almost certainly fall outside the scope of existing regulation.

On this note, there is no international governance regime that qualifies the collisions carried out in research accelerators such as the RHIC and LHC.[313] As written by Johnson, if there had been a Worldwide Accelerator Safety Administration and CERN had violated one of its safety standards having the force of law, the LHC Controversy

World of Disorder (Cambridge University Press 2018) 146; Separately, Voeneky explains that there is no coherent body of international law obligating states and private entities to evaluate technological risks. See Silja Vöneky, 'The Public International Law Perspective on Evaluating Existential Risks' (*Cambridge Conference on Catastrophic Risk 2018, The Centre for the Study of Existential Risk*, 17 April 2019, [00:02:08] – [00:02:59]) <https://www.youtube.com/watch?v=O6ydGbBv2m0> accessed 26 April 2023; Farquhar and others also highlight the difficulty in applying existing governance instruments to existential risks. Farquhar and others, 'Existential Risk' (n 61) 12.

[306] Voeneky, 'Human Rights and Legitimate Governance' (n 305) 146.

[307] Examples when it comes to biotechnology include the Biological Weapons Convention, Convention on Biological Diversity, Cartagena Protocol on Biosafety. An example regarding geoengineering is the Vienna Convention for the Protection of the Ozone Layer. ibid 146.

[308] Catherine Rhodes, 'The Future Governance of Biotechnology' in Simon Whitby and others (eds), *Preventing Biological Threats: What You Can Do* (Bradford Disarmament Research Centre 2015) para 19.

[309] Samuel Adams, '"Honey I Blew Up the World!"? One Small Step Towards Filling the Regulatory "Black Hole" at the Intersection of High-Energy Particle Colliders and International Law' [2009] Georgia Journal of international and Comparative Law 131, 159.

[310] 'Physicists and the Bomb' (2015) 11(201) Nature Physics <https://doi.org/10.1038/nphys3287> accessed 26 April 2023.

[311] Article 1.

[312] Adams, 'Honey' (n 309) 159.

[313] As such, Adams explores the design of such a regime. ibid 160-164.

would have been straightforward in the sense that the machine could have been temporarily stopped.[314]

While there are treaties that apply to areas of science and hazards linked to the x-risk landscape, it is uncertain whether they can govern novel scientific activities.[315] It is said that many x-risks pose an unprecedented and uncertain future threat.[316] As mentioned by Voeneky, the drafters can neither anticipate the future nor the radical scientific developments that will take place.[317] Each treaty must also be interpreted according to its objective and purpose.[318] Thus, it may be difficult to control new risks by relying on old instruments not tailored for such purposes.

Wilson demonstrates this point when it comes to biotechnology.[319] An example from his article involves the Convention on Biological Diversity and the Cartagena Protocol on Biosafety. Wilson explains that these instruments are unable to regulate the x-risk associated with an accidental release of living modified organisms; the Convention lacks an effective enforcement mechanism; it also fails to address what the Parties must do to "...regulate, manage or control the risks associated with the use and release of living modified organisms..."[320] For instance, measures on laboratory safety are absent. While the Protocol expands on the Convention, it is said to focus too much on trade and transboundary activities. As such, it is of limited relevance when a laboratory operates within a single state. Wilson also describes the Protocol's requirement on risk assessment and risk management as too discretionary to help in mitigating the risk.[321]

[314] Johnson, The Black Hole Case' (n 68) 861.
[315] See Voeneky, 'Human Rights and Legitimate Governance' (n 305) 147.
[316] Farquhar and others, 'Existential Risk' (n 61) 12.
[317] Voeneky, 'Human Rights and Legitimate Governance' (n 305) 147.
[318] Vöneky, 'The Public International Law Perspective on Evaluating Existential Risks' (n 305) [00:02:08] – [00:02:59].
[319] Wilson, 'Minimizing' (n 151) 339.
[320] See Article 8(g).
[321] Wilson, 'Minimizing' (n 151) 339-344.

(viii) Multidisciplinary

The hazards explored indicate that the study of x-risk is multidisciplinary. In fact, the illustrative hazards above are not exhaustive. X-risk is also associated with climate change, nanotechnology, global totalitarianism, SETI[322] and the claim that humanity may live in a computer simulation.[323] Once again, this list remains incomplete.

In essence, one can be sceptical of a broad legal intervention aiming to tackle all hazards head-on. Whereas the design of such intervention is perhaps not impossible, it is ambitious. One possible approach is to tackle a 'group' of related hazards. For example, Wilson offers a solution aiming to address global catastrophic risks (GCRs) and existential risks (ERs) from emerging technologies, including AI, biotechnology and nanotechnology.[324] He suggests that his solution may also come to cover future technologies.[325] In effect, it is possible to characterise Wilson's legal approach as a *broad* legal intervention because it aims to reduce the likelihood of x-risk from two or more hazards. Wilson's proposal centres around a treaty made up of regulatory mechanisms, including the precautionary principle, a body of experts, reforms for scientists and laboratory safety measures.[326] His proposal relies on a broad structure to apply to various hazards. The advantage of this approach is scope and flexibility. Yet, these variables can come at the expense of a more tailored regime for different hazards in the form of targeted legal interventions.

Whereas Wilson's framework treaty provides an example of a broad legal intervention for x-risk mitigation, it is possible to question his decision to combine GCRs and ERs under the same treaty. In other words, is it desirable to pursue identical mechanisms when it comes to these distinct risk categories? This question is relevant more generally when considering legal interventions within the x-risk landscape. Should broad legal interventions focus solely on x-risk, or should they also focus on less serious risks such as GCRs? According to Wilson, "[s]everal GCRs are also ERs, such

[322] Search for Extra-terrestrial Intelligence by e.g., sending signals to attract attention. See Ord, *Precipice* (n 16) 160.
[323] The risk is that the computer simulation might be shut down. Bostrom, 'Existential Risks' (n 15).
[324] Wilson, 'Minimizing' (n 151) 348.
[325] ibid 350.
[326] ibid 351-363.

as nuclear war, certain experimental technologies,…" biotechnology and AI.[327] This argument may support his approach. However, it is not always clear whether this is true. Think back to the Sub-section above dealing with the probability that the materialised catastrophe exceeds the relevant harms threshold.

The more pertinent aspect to note is that the combination of GCRs and ERs may undermine the distinction between the *recoverable* and *unrecoverable*, between harm to the *present* and destruction of the *future*. An x-risk threatens the premature extinction of humanity. In contrast, a GCR threatens to cause damage to human well-being on a global scale.[328] Both are serious. But only one threatens to obliterate the future, and this makes an x-catastrophe uniquely bad.[329] The ASP is mainly relevant within the context of x-risk. In effect, if one believes that the ASP should underpin and empower legal interventions, it may be better to tackle x-risks separately.

Nevertheless, it might be that Wilson does not find the distinction between GCRs and ERs convincing. In other words, his decision to combine GCRs and ERs for his broad intervention might be grounded in a view that the two risk categories warrant the same response. Some other legal scholars can also be seen to undermine (deliberately or otherwise) the distinction between the relative seriousness of GCRs and ERs.[330]

Another example is Posner's *Catastrophe: Risk and Response*. This book is mentioned by other people who are invested in x-risk research.[331] Moynihan writes that Posner's book is one of the "…important precursors to the work of Bostrom et al…"[332] It is true that *Catastrophe: Risk and Response* contains many observations that apply to the field of x-risk research. However, it is ultimately about catastrophes

[327] ibid 312.
[328] This is the definition employed by Wilson. ibid 311.
[329] This is considered in Chapter 5.
[330] E.g., Voeneky lumps together *x-risks* and *GCRs* in her chapter. The decision to entwine these categories of risk leads to the following question: Is this approach an indication that they (i) disagree with the view that an x-catastrophe is uniquely bad, or (ii) that they accept the unique badness, but think that this conclusion does not matter for the mitigation of x-risk? Voeneky, 'Human Rights and Legitimate Governance' (n 305).
[331] Phil Torres, 'Facing Disaster: The Great Challenges Framework' [2019] Foresight 4, 7; Cotton-Barratt, Daniel and Sandberg, 'Defence' (n 37) 273; Baum, 'The Great Downside Dilemma' (n 22).
[332] Thomas Moynihan, *X-Risk: How Humanity Discovered its Own Extinction* (Urbanomic 2020) 21.

that "…threaten the survival of the human race."[333] The destruction of the future is not a dominant (or even particular important) concern in Posner's analysis. He writes:

> Some people think it important that the human race survive for millions, even billions or trillions, of years. Worried therefore about the expansion of the sun into the earth's orbit…they want us to begin thinking seriously about colonizing other planets…Most people who think along these lines do so not because they have too much imagination but because they have too little.[334]

Later on, and for the purpose of cost-benefit analysis, Posner estimates the cost of extinction of the human race at $600 trillion.[335] For his analysis, Posner values human lives by considering the value that people place on avoiding low-probability risks of death. From this perspective, he reaches a conclusion where he values an individual life at $50,000.[336] He then multiplies $50,000 by 6 billion (to account for the global population in 2004). In effect, the cost of extinction of the human race is estimated at $300 trillion. This estimate does not take the future into consideration. Indeed, the reason for summarising Posner's approach comes next. He writes: "The figure of 6 billion for the population at risk is undoubtedly too low. It ascribes no value to future humanity…"[337] In light of this recognition, he continues: "Suppose as a crude adjustment we simply double the figure for the current population and…multiply it by …$50,000."[338] Since $50,000 x 12 billion = $600 trillion, the cost of extinction of the human race, *accounting for the entire future*, is estimated at $600 trillion. Note, then, how little weight Posner attaches to the potential value of humanity's entire future. While putting a price tag on the future is very difficult, the 'crude adjustment' is highly incompatible with, for instance, the ASP. But to his defence, Posner recognises that his estimate is extremely conservative.[339]

Another example of a scholar that undermines the distinction between the relative seriousness of GCRs and ERs is Johnson. This thesis will return to Johnson's work in more detail in subsequent chapters. For now, we can note that Johnson considers the

[333] Posner, *Catastrophe* (n 90) 6.
[334] ibid 16.
[335] ibid 141, 165-170.
[336] ibid 168.
[337] ibid 169.
[338] ibid 169-170.
[339] ibid 141; Posner, *Catastrophe* (n 90) 190.

value of future lives for the purpose of cost-benefit analysis in the context of the LHC Controversy. At first glance, this decision suggests that Johnson is open to the ideas embedded in the ASP. However, on a closer examination, his analysis is suggestive of the opposite truth. Consider the following equation, which he uses to calculate L – human extinction loss – presuming that the human population on Earth would stay constant but for the extinction event:

$$L = V_{yr} \times H_{\oplus} \times Y_{\oplus}$$

In this equation, V_{yr} is the value of one human life per year, $H_{\oplus}$ is Earth's human population, and $Y_{\oplus}$ is the number of years Earth has left before the sun has its way with the planet. Multiplying the numbers, Johnson finds that L = $3.87 septillion, as seen in this footnote.[340]

It is possible to question the numbers that Johnson plug into his equation. More specifically, he presumes that Earth has ~ five billion years left before the sun renders our planet uninhabitable. However, the principal interest should not be how many years Earth has left. Instead, it should—as implied by his equation—be how many years the human population (or *humanity* more broadly defined) can survive on Earth. As will be observed in Chapter 5, there is a non-negligible subjective probability that humanity can survive on Earth until it becomes uninhabitable as well as colonise space.[341] What matters is that Johnson does not account for the possibility of space colonisation. As dealt with in Chapter 5, this is the prospect that will unlock the "…lion's share of the expected duration of our existence."[342]

Johnson's decision to omit the prospect of space colonisation might be viewed as indicative that he is not too concerned with the themes that underpin the ASP. However, it is the upcoming observations that truly suggests that Johnson approach can be seen as undermining the distinction between the relative seriousness of GCRs

[340] Johnson presumes that V_{yr} = $129,000, $H_{\oplus}$ = six billion, $Y_{\oplus}$ = five billion. Thus, $L = (\$1.29 \times 10^5)(6 \times 10^9 \text{ people})(5 \times 10^9 \text{ yr})$. See Johnson, 'The Black Hole Case' (n 68) 882-883.
[341] See Chapter 5.
[342] Nicholas Beckstead, 'On the Overwhelming Importance of Shaping the Far Future' (DPhil thesis, Rutgers University 2013) 57.

and ERs. Having concluded that *L* = $3.87 septillion, Johnson highlights a "big gap" in his analysis.[343] Noting that future people might be miserable, he writes that we cannot guarantee that life provides a net benefit in the grand scheme of things.[344] He continues by saying that if a black hole disaster unfolds, "…there would be no more pain, no more suffering, no more hunger, and no more grieving."[345] Therefore, "…there will be no soul left to shed a tear."[346] Johnson goes on to consider how the abovementioned conception of death is enshrined in the American common law of torts, after which he makes the remark that we highlighted above: Namely, that the downside of a particle-accelerator disaster might be nothing provided that it kills everyone quickly.[347] This outlook is in stark conflict with the standpoint in this thesis.

Interestingly, Ord writes that when he was younger, he also toyed with the idea that humanity's destruction might not be bad given that there "…would be no people to suffer or grieve."[348] However, he explains why he no longer holds this view. In fact, it is a view that conflicts with the notion of what makes an existential catastrophe uniquely bad. Thus, those interested may want to consider Ord's current take on the sort of idea that Johnson paints above.[349]

(ix) Summary

Part 2.3 has dealt with some of the sources of existential risk ('hazards'). Exploring the hazards above has helped the thesis to map out, to some extent, the x-risk landscape within which a broad legal intervention will operate. Moreover, examining the hazards and the literature dealing with them has enabled the thesis to document some general observations that may be relevant to broad legal interventions for x-risk mitigation. For example, legal interventions may need to balance existential risk with astronomical benefits. Broad legal interventions will also need to be mindful that not all existential risks are equally probable, neither in terms of how likely the risk is to materialise in the

[343] Johnson, 'The Black Hole Case' (n 68) 883.
[344] ibid.
[345] ibid.
[346] ibid.
[347] ibid.
[348] Ord, *Precipice* (n 16) 47.
[349] ibid 47-48.

first place nor how likely it is that a materialised catastrophe will amount to an *existential* catastrophe.

Part 2.4 below will deal with another aspect that is relevant in understanding the existential risk landscape. Namely, it will consider a set of features that can complicate the study and mitigation of these risk.

2.4 OBSCURING FEATURES

This Part highlights some obscuring features which can complicate the study and mitigation of existential risks. Distilling some of these features will contribute to a more holistic understanding of the existential risk landscape. These features can affect broad legal interventions, given that they operate to make x-risk mitigation more problematic across the board.

As articulated by Bostrom, "[m]any factors conspire against the study and mitigation of existential risks."[350] The field of x-risk is relatively neglected, and the obscuring features can seemingly contribute to the status quo. In 2002, Bostrom wrote that there was less scholarly work on x-risks than on the life−habits of the dung fly.[351] The general neglect is also emphasised more recently,[352] with Ord highlighting that the society invests much too little to reduce existential risks.[353]

Torres writes that Google Scholar returns millions of more results for the term "cancer" than "existential risk".[354] Having attempted a similar search, it is possible to affirm his

[350] Bostrom, 'Existential Risk Prevention' (n 15) 26.
[351] Bostrom, 'Existential Risks' (n 15).
[352] Bostrom, 'Existential Risk Prevention' (n 15); Phil Torres, 'Facing Disaster' (n 331); Boyd and Wilson, 'Existential Risks to Humanity' (n 154) 2305-2307.
[353] While recognising that it is difficult to measure the precise global spending on x-risk, Ord writes that "…we can state with confidence that humanity spends more on ice cream every year than on ensuring that the technologies we develop do no destroy us." See Ord, *Precipice* (n 16) 58.
[354] Torres: "…conducted some Google Scholar searches and found that, as of January 24, 2018, there were exactly 1,910 results for the word "existential risk." In comparison, there were…5,390,000 results for "cancer."" See Phil Torres, 'Why an Existential Risk Expert Finds Hope in Humanity's Certain Doom' (Medium, 16 December 2019) <https://onezero.medium.com/rebelling-against-extinction-d7e112979bed> accessed 27 April 2023.

conclusion at the time of writing this section.[355] Torres suggests that his finding is "…ironic given that curing cancer doesn't much matter if there's no one around to cure."[356] His statement will likely not resonate with a lot of people. One can wonder why Torres choose this emotionally charged example[357] to highlight the relative neglect of x-risk research. His example seems counter-intuitive if the aim is to inspire more people to question current research priorities.

2.4.1 The Hazards and Obscuring Features

Bostrom's article notes why it may be practical to look at existential risk as a distinct category, and in doing so, he also highlights some of the obscuring features that complicate the study and mitigation of existential risks:

> Existential risks have a cluster of features that make it useful to identify them as a special category: the extreme magnitude of the harm that would come from an existential disaster; the futility of the trial−and−error approach; the lack of evolved biological and cultural coping methods; the fact that existential risk dilution is a global public good; the shared stakeholdership of all future generations; the international nature of many of the required countermeasures; the necessarily highly speculative and multidisciplinary nature of the topic; the subtle and diverse methodological problems involved in assessing the probability of existential risks; and the comparative neglect of the whole area.[358]

The sections below will neither analyse all obscuring features nor the selected few in detail.[359] For those interested, Posner's book devotes a full chapter to examining why the society is doing so little to mitigate catastrophic risks.[360] He deals with this under three main headings: Cultural Factors, Psychological Factors and Economic Factors.

---

[355] On 02 December 2020, I used Google Scholar to search for articles from 'Any time'. I used *Advanced search* and selected the parameters to find articles 'with the exact phrase' and 'in the title of the article.' Within these parameters "Existential risk" returned ~ 112 results and "Cancer" returned ~ 1,450,000 results.

[356] Torres, 'Why an Existential Risk Expert Finds Hope in Humanity's Certain Doom' (n 354); In another piece, Torres recognises that cancer research is important. However, he writes that a significant portion of its value is predicated on humanity's continued existence. Thus, ensuring human survival should, in his opinion, take precedence over curing cancer. Phil Torres, 'Facing Disaster' (n 331) 20.

[357] However, he did, in fact, conduct searches of inter alia "French cheese", "Super Mario Brothers" and "Hospitality Management". ibid 20.

[358] Bostrom, 'Existential Risks' (n 15).

[359] For some overlooked 'complicating factors', see Phil Torres, 'Facing Disaster' (n 331) 18-25.

[360] Posner, *Catastrophe* (n 90) 92-138.

This terminology is borrowed to compartmentalise some of the sections below. The features examined below are selected because they align with this thesis's research focus and parameters outlined in Chapter 1.

(i) Irrevocable Nature

Of pivotal importance is the fact that an existential catastrophe is irrevocable.[361] It is not necessarily the case that all people instinctively and immediately appreciate that an existential catastrophe is final and permanent. Humanity has never experienced one because the event forecloses the tale from being retold.

In describing this thesis to others, some have asked whether an existential catastrophe has happened in the past. Some have also asked whether COVID-19 poses an x-risk or whether it is the manifestation of an x-catastrophe. These questions arguably demonstrate that those who do not commonly think about x-risk sometimes fail to instinctively appreciate the true finality of an existential catastrophe. The irrevocable nature is an obscuring feature, which has different implications.

Firstly, humanity cannot pursue a trial-and-error approach to the mitigation of x-risk.[362] Unfortunately, it is precisely such an approach, writes Bostrom, that has shaped our institutions.[363] On a related note, Wiener explains that "[r]are mega-catastrophes may be literally 'too big to handle' by the traditional legal system."[364] Some catastrophes will destroy the entire legal system. This is certainly the case when we are dealing with an existential catastrophe. Thus,

> [A]ny *ex post* legal remedy, such as liability, would be an empty prospect when the catastrophic risk itself would destroy the institutions meant to impose such an *ex post* remedy. The court system would not likely be operating after a large global catastrophe.[365]

[361] Wilson, 'Minimizing' (n 151) 312.
[362] Bostrom, 'Existential Risks' (n 15); Ord, *Precipice* (n 16); Wilson, 'Minimizing' (n 151) 312.
[363] Bostrom, 'Existential Risk Prevention' (n 15) 27.
[364] Jonathan Wiener, 'The Tragedy of the Uncommons: On the Politics of Apocalypse' [2016] Global Policy 67, 73.
[365] ibid.

Likewise, Annas writes that new technologies have made it possible for *some* humans to put *all* humans at risk of extinction.[366] Naturally, explains Annas, "…only prevention matters in this extreme sphere: an extinct…species cannot prosecute its destroyer. Only unsuccessful attempts to destroy the human species are prosecutable."[367] Secondly, the unprecedented nature will influence how society think about and estimate the probability of different existential catastrophes. Thirdly, the unprecedented nature can sway our perception of existential risks (see below on psychological factors).

Legal interventions aiming to reduce the probability of existential risk will need to account for the irrevocable nature of existential risk. The most apparent effect is that legal interventions that concern *sudden* existential catastrophes rather than cascading failures will need to prevent the harm from occurring in the first place.

(ii) Psychological Factors

The irrevocable nature means that an existential catastrophe cannot be experienced before it is too late: "If only seeing is believing, we will step blindly over the precipice."[368] Scholars including Bostrom,[369] Posner,[370] Ord,[371] Yudkowsky,[372] Farquhar and others,[373] and Wiener[374] recognise that psychological factors can affect a person's judgement of x-risks. Torres's book also lists several cognitive distortions that are relevant to the field of existential risk studies.[375] Consider the notion of

[366] George Annas, 'Crimes Against the Human Species ('Type II Crimes Against Humanity' Explained), in Britta van Beers, Luigi Corrias and Wouter Werner (eds), *Humanity across International Law and Biolaw* (Cambridge University Press 2014) 120.
[367] ibid 118-120. It is interesting to highlight what Annas is exploring in his work. He essentially promotes a new category of crimes against humanity (called 'type II') which include all intentional acts threatening humanity with extinction, and all species alterations endangering our species by radically changing what it means to be human. On page 115, he explains that species-endangering *experiments* are the typical type II crime against humanity.
[368] Ord, *Precipice* (n 16) 61.
[369] Bostrom, 'Existential Risks' (n 15); Bostrom, 'Existential Risk Prevention' (n 15) 26-27.
[370] Posner, *Catastrophe* (n 90) 119-123.
[371] Ord, *Precipice* (n 16) 60-61.
[372] Eliezer Yudkowsky, 'Cognitive Biases Potentially Affecting Judgment of Global Risks' in Nick Bostrom and Milan Ćirković (eds), *Global Catastrophic Risks* (OUP 2008).
[373] Farquhar and others, 'Existential Risk' (n 61) 12.
[374] Wiener, 'The Tragedy of the Uncommons' (n 364).
[375] Torres, *Morality, Foresight, and Human Flourishing* (n 17) 53-54.

availability heuristic. Yudkowsky explains that "[a]n example of a heuristic is to judge the frequency or probability of an event by its *availability*, the ease with which examples of the event come to mind."[376] Posner refers to this notion as the 'economy of attention'.[377] One reason that makes the public worry more about airline crashes than asteroid collisions, Posner writes, is that planes crash every few years. Such events are more readily available to the mind.[378] On a related note, it is not surprising that people may have hesitated to fly on the 737 MAX as opposed to other aircrafts.[379] The implication of the above is that humanity may perpetually neglect and underestimate x-risk.

Another bias is *scope neglect* which relates to a lack of sensitivity to the scale of a benefit or harm.[380] Given that "…existential risk derives its key moral importance from the size of what is at stake, scope neglect leads us to seriously underweight its importance."[381] Yudkowsky vividly summarises the issue:

> Human emotions take place within an analogous brain. The human brain cannot release enough neurotransmitters to feel emotions a 1000 times as strong as the grief of one funeral. A prospective risk going from 10,000,000 deaths to 100,000,000 deaths does not multiply by ten the strength of our determination to stop it. It adds one more zero on paper for our eyes to glaze over…[382]

[376] Yudkowsky, 'Cognitive Biases' (n 372) 92.
[377] Posner, *Catastrophe* (n 90) 122.
[378] ibid 121.
[379] Given the option, they might prefer another aircraft. A quick internet search reveal that some airlines anticipate that customers may be reluctant to fly on a 737 MAX. As written by American Airlines Newsroom: "If a customer doesn't want to fly on the 737 MAX, they won't have to. Our customers will be able to easily identify whether they are traveling on one even if schedules change. If a customer prefers to not fly on this aircraft, we'll provide flexibility to ensure they can be easily re-accommodated." American Airlines Newsroom, 'Return of the Boeing 737 MAX to Service' (18 November 2020) <http://news.aa.com/news/news-details/2020/Return-of-the-Boeing-737-MAX-to-service-OPS-DIS-11/default.aspx> accessed 27 April 2023.
[380] Yudkowsky, 'Cognitive Biases' (n 372) 105.
[381] Ord, *Precipice* (n 16) 61.
[382] Yudkowsky, 'Cognitive Biases' (n 372) 106-107.

Consider the statement that an existential catastrophe may deprive a *quadrillion*[383] people of a good future.[384] Does this number mean anything to most people? Does it convey what is at stakes? The point is perhaps that we sometimes feel for one what one cannot feel for many.

A similar point is expressed by Slovic. Examining the neglect of genocide (and the psychological obstacles behind such neglect), Slovic explains that we cannot feel the humanity behind large numbers.[385] Statistics fail to spark emotion and motivate actions.[386] He observes that the importance and value that we attach to saving lives diminishes as the number of lives increases.[387] He writes:

> As the psychophysical research indicates, constant increases in the magnitude of a stimulus typically evoke smaller and smaller changes in response. Applying this principle to the valuing of human life suggests that a form of *psychophysical numbing* may result from our inability to appreciate losses of life as they become larger…[388]

If the vast majority underestimate the seriousness of x-risk, it seems reasonable to think that a politician "…who wants to raise taxes today to minimize the risk of catastrophes a thousand years hence…"[389] will commit political suicide. This statement captures the more general idea that protecting the future – a key driver behind reducing existential risks – is not a top priority for politicians. Ultimately, the above may suggest that there may be a disconnect between what is at stake and legal interventions that aim to reduce existential risk.

---

[383] I.e., 1 x 10^{15} or 1,000,000,000,000,000.
[384] This is not just a random number. Bostrom, for instance, talks about the *quadrillions* of people that may come to exist in the future (absent an x-catastrophe). Bostrom, 'Existential Risk Prevention' (n 15) 18-19 and 26.
[385] Paul Slovic, '"If I look at the Mass I will Never act": Psychic Numbing and Genocide' [2007] Judgment and Decision Making 79, 84.
[386] ibid 80.
[387] ibid 85.
[388] ibid; Wiener, 'The Tragedy of the Uncommons' (n 364).
[389] Posner, *Catastrophe* (n 90) 118.

(iii) Political Factors

Posner and Ord suggest political factors contribute to the neglect of x-risks. Along with other scholar, they note that politicians will devote their energy to deal with short-term priorities.[390] Posner writes that politicians are unlikely to win points for preventing something that may occur in the future.[391] Policymakers also apply social discount rates,[392] which can undermine the ambition of investing today to protect the future.

Nonetheless, several countries including Sweden, Finland and Hungary have, with varying degree of success, established institutional structures that represent the interest of future generations.[393] As such, there is not a complete absence of political incentive to protect future generations with many constitutions now "…referencing future generations in some capacity…"[394]

In the UK, one can also note the Wellbeing of Future Generations Bill, which, at the time of writing this section,[395] has completed its House of Lords stages and is soon to have its second reading debate in the House of Commons.[396] The Bill emphasises the interests of future generations and the Explanatory Notes specifically mention x-risk.[397] Whereas the Bill (as brought from the Lords) does not mention x-risk, it contains clauses that bring these risks to the forefront. Section 16 requires the

---

[390] Jones, O'Brien and Ryan, rely on the premise that future generations are under-represented in contemporary political structures, one of the reasons for this being political 'short-termism'. Natalie Jones, Mark O'Brien, Thomas Ryan, 'Representation of Future Generations in United Kingdom Policymaking' [2018] Futures 153, 153-154.

[391] Posner, *Catastrophe* (n 90) 118; Ord, *Precipice* (n 16) 60.

[392] "A social discount rate is a technique that policy-makers use in their cost-benefit analyses to gauge whether to make investments with a long-term impact. It weighs the upsides for future people against costs borne in the present-day, and proposes that the calculated value of benefits to future economies and people should steadily decline over time." See Richard Fisher, 'The Perils Of Short-Termism: Civilisation's Greatest Threat' *BBC Future* (14 December 2020) <https://www.bbc.com/future/article/20190109-the-perils-of-short-termism-civilisations-greatest-threat> accessed 27 April 2023.

[393] Jones, O'Brien, Ryan, 'Representation of Future Generations' (n 390) 153-163.

[394] Eric Martinez and Christoph Winter, 'Protecting Future Generations: A Global Survey of Legal Academics', (LPP Working Paper Series No 1-2021) <https://www.legalpriorities.org/documents/1%20-%20Protecting%20Future%20Generations.pdf> accessed 27 April 2023 13.

[395] 14 May 2022.

[396] Wellbeing of Future Generations HL Bill (2021-22) 253.

[397] Wellbeing of Future Generations Bill [HL] (Explanatory Notes) <https://bills.parliament.uk/publications/41931/documents/417> accessed 27 April 2023.

Secretary of State to publish a "futures and forecasting report".[398] According to section 16 (1) (c), this report must include "…an assessment of risks, including high-impact, low-probability risks, environmental risks, global risks and risks that may emerge or grow in the future, for at least the next 25 years…"[399] What is more, Part 4 provides existing and new public bodies with different reporting and oversight roles.[400] Furthermore, Part 5 provides for the establishment of the "Future Generations Commission for the United Kingdom"[401] with a general duty to, among other things, "act as a guardian of the ability of future generations to meet their needs"[402] and "encourage public bodies to take greater account of the long-term impact of the things that they do"[403]

(iv) International Cooperation

An existential catastrophe will affect all jurisdictions. Thus, states should try to cooperate to mitigate risks within the x-risk landscape. For example, they can resort to international law to regulate certain hazards such as AGI and SRM techniques. Depending on a hazard's governance gap,[404] it may be suitable to either create a new treaty or amend an existing one. Recognising that AGI can pose an x-risk, Nindler, for example, considers the virtue of an international treaty ensuring safe AI research.[405] Wilson is another scholar who explores the option of regulating x-risks from emerging technologies by creating an international treaty.[406]

Nevertheless, creating an efficient multilateral treaty is easier said than done. Wittes and Blum highlights this in their book.[407] Firstly, international law is consent-

[398] Wellbeing of Future Generations HL Bill (n 398) 253.
[399] ibid.
[400] ibid.
[401] Section 24(1).
[402] Section 25(a)(i).
[403] Section 25(a)(ii).
[404] Kemp and Rhodes, 'The Cartography of Global Catastrophic Governance' (n 53).
[405] Nindler, 'The United Nation's Capability' (n 54) 29.
[406] Wilson, 'Minimizing' (n 151).
[407] Benjamin Wittes and Gabriella Blum, *The Future of Violence: Robots and Germs, Hackers and Drones: Confronting A New Age of Threat* (Basic Books 2015) 237-243.

based,[408] meaning that joining a treaty is voluntary.[409] States can refrain from joining a key treaty on the mitigation of existential risk. Consider the following statement:

> [E]ven if a majority of countries pass and enforce national laws against the creation of some specific destructive version of nanotechnology, will we really have gained safety if some less scrupulous countries decide to forge ahead regardless?[410]

It is possible to reframe this question: Has humanity gained any safety if most states join a treaty prohibiting the same? It is possible to use Baum's equation ($P_{EC}=P_1 * P_2$) to consider this question. Arguably, a treaty with 150 parties can sway P_1 by reducing the likelihood of a catastrophe event. If so, humanity has gained some additional safety. Still, Bostrom's core argument is valid: "Many existential risks simply cannot be substantially reduced by actions that are internal to one or even most countries."[411]

Secondly, Wittes and Blum explain that countries can ratify treaties but make reservations which free them from important substantive obligations.[412] This leeway can undermine the effectiveness of a treaty. The points above demonstrate a tension between national sovereignty and existential risks. To tackle this tension, extreme measures like the use of force, have been proposed.[413]

---

[408] "Since states are considered to be sovereign…law can only be made with their consent…Hence international law is often said to be a consent-based (or consensual) system." See Jan Klabbers, *International Law* (2nd edn, Cambridge University Press 2017) 24; "[I]nternational law works on the basis that the general consent or acceptance of states can create rules of general application." See James Crawford, *Brownlie's Principles of Public International Law* (8th edn, OUP 2012) 20.

[409] E.g., the US has reportedly not ratified certain weapons conventions. See Wittes and Blum, *The Future of Violence* (n 407) 238.

[410] Bostrom, 'Existential Risks' (n 15).

[411] ibid.

[412] Wittes and Blum, *The Future of Violence* (n 407) 239.

[413] Bostrom writes: "Respect for national sovereignty is not a legitimate excuse for failing to take countermeasures against a major existential risk." He further argues that, in some cases, "…the mere decision to go forward with development of the hazardous technology in the absence of sufficient regulation must be interpreted as an act of aggression…". See Bostrom, 'Existential Risks' (n 15). Legal scholars will likely criticise this proposal. Having noted Bostrom's argument, Nindler states: "…[I]t is clear that the use of force can be only legal under very narrow circumstances." He further explains that the mere decision to go forward with development of the hazardous technology in the absence of sufficient regulation will not align with the definition of aggression in international law. Nindler, 'The United Nation's Capability' (n 54) 31-32.

Suppose that all states sign up to a treaty on the mitigation of existential risk. This brings to light another issue that can weaken its efficiency. Namely, compliance gaps can remain even when treaties have achieved widespread ratification.[414] Some agreements lack effective enforcement measures.[415] Additionally, transnational monitoring, as a tool for encouraging compliance, come with no guarantees.[416] Wittes and Blum invite us to consider an agreement that regulates the worldwide use, stockpiling, and acquisition of knives. "[T]he prospects for effective monitoring and enforcement…" their book reads, "…are too slim to justify the endeavor."[417] This example is only illustrative of their main point that "…coordination and enforcement with respect to *technologies of mass empowerment* present profound difficulties."[418]

Rees makes a similar observation when it comes to biothreats.[419] He writes that "[t]housands of individuals, perhaps even millions, may someday acquire the capability to disseminate "weapons" that could cause widespread (even worldwide) epidemics."[420] It could, therefore, prove difficult to monitor everyone who can trigger a catastrophic event. This can be contrasted with, for example, detecting the production of nuclear weapons.[421] Similarly, Ord speaks of the 'democratisation' of biotechnology.[422] This thesis will not analyse how the scene painted by Rees and Ord may influence P_2. However, it can seemingly increase P_1 since more individuals can initiate a catastrophic event. On the other hand, it is not unthinkable that such a reality could reduce P_1 from the relevant hazard ('H_n'). For instance, states may attempt to monitor, regulate and control all individuals to respond to the growing threat. Powerful

[414] Wittes and Blum, *The Future of Violence* (n 407) 239.
[415] The Arms Trade Treaty might be used as example as it merely encourages states to cooperate and includes no international enforcement measures. ibid 240.
[416] Wittes and Blum note the Biological Weapons Convention as a 'disturbing' example where transnational monitoring is inadequate. ibid.
[417] ibid 243
[418] ibid. On pages 20-23, Wittes and Blum consider "Modern Technologies of Mass Empowerment". They continue by giving an overview of different mass-empower technologies including 'Networked Computers', 'Biotechnology' and 'Robotics'.
[419] Rees, *Our Final Century* (n 226) 48.
[420] ibid.
[421] Elaborate equipment is necessary to create nuclear weapons. ibid; Bostrom also recognises that nanotechnology (as an example) can be harder to regulate than nuclear weapons due to the more onerous production requirements of the latter. Bostrom, 'Existential Risks' (n 15).
[422] That is "…the speed at which cutting-edge techniques can be adopted by students and amateurs." Ord, *Precipice* (n 16) 133.

monitoring techniques can, in itself, contribute towards an x-catastrophe.[423] Paradoxically, reducing P_1 from H_1 (biotechnology) can perhaps increase P_1 from H_2 (a repressive totalitarian regime).

There are other factors too which can undermine the effectiveness of an international treaty and the possibility to conclude one in the first place. Examples include the difficulty in keeping up with the onrush of science and technology,[424] lack of funding,[425] political disagreement about which activities to supress or permit,[426] and the temptation of states to free-ride on the mitigation efforts of other nations.[427]

The issues identified above will likely affect most legal interventions, especially if they are grounded international law.

(v) Conceptual Complexity

This Section highlights a feature that can be referred to as conceptual complexity.[428] As the name indicates, it concerns the underlying *conceptual* complexities that can impede someone's ability to deal with x-risk as a unified field of study. Many things can be linked to this overarching feature, such as the world's *complexification*[429] and

---

[423] Global coordination strong enough to continually monitor the entire world population can create a risk of permanent stagnation or flawed realisation under a repressive totalitarian regime. See Bostrom, 'Existential Risk Prevention' (n 15) 29.

[424] "In the case of law, neglect of the catastrophic risks is part of a larger problem, that of the law's faltering struggle to cope with the onrush of science." See Posner, *Catastrophe* (n 90) 8; Relevant here is the pacing problem. As explained by Thierer, it concerns the problem that technological innovation outpaces the ability of regulations to keep up. Adam Thierer, 'The Pacing Problem and the Future of Technology Regulation' *The Mercatus Center* (8 August 2018) <https://www.mercatus.org/bridge/commentary/pacing-problem-and-future-technology-regulation> accessed 27 April 2023; There is much literature on this problem. See for instance Gary Marchant, Braden Allenby and Joseph Herkert (eds), *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem* (Springer 2011).

[425] Ord, *Precipice* (n 16) 135-6.

[426] "Unlike with trafficking in women and children, say, or the hijacking of airplanes, many states regard the Internet, biotechnology, and robotics as facilitating desirable activities…" See Wittes and Blum, *The Future of Violence* (n 407) 242.

[427] This issue flows from the fact that existential risk mitigation is a global public good. Farquhar and others, 'Existential Risk' (n 61) 11.

[428] This thesis has taken inspiration from Torres in labelling this obscuring feature. See Phil Torres, 'Facing Disaster' (n 331) 20.

[429] ibid 20-21.

the field's *speculative* and *interdisciplinary* nature.[430] The point is not that reading about existential risk is more complicated than reading about other areas. Instead, the essential idea is described by Torres:

> [T]he more interdisciplinary a field, the greater the impact of individual ignorance, given the "epistemic breadth-depth tradeoff" (i.e. crudely put, one can know a lot about a little, or a little about a lot). Indeed, a central aim of this nascent but important field is to determine which directions humanity should steer the ship, insofar as we have control over our trajectory through time. But acquiring robust knowledge about high-level, big-picture issues that span so many fields of human inquiry – from economics to biology, astrobiology to population ethics, computer science to sociology, technology studies to decision theory and so on – is…increasingly beyond the bounds of human capability.[431]

He explains that complexification makes it impossible for anyone to make truly wise decisions about how to navigate all the different existential hazards.[432] Ord and other scholars[433] paint a similar picture by outlining the multifaceted nature of the field of x-risk studies.[434] Firstly, understanding the risks (hazards), require insight into physics, biology and computer science to name only a few disciplines. Secondly, it is necessary to resort to moral philosophy and economics to judge how much is at stake. Thirdly, attempting to solve and mitigate the risks require international relations and political science. Ord writes that engaging properly with all relevant disciplines is not possible for any one individual.[435]

This thesis is affected by this obscuring feature. This Chapter is perhaps the primary victim because it attempts to map different aspects of the x-risk. However, anyone introducing different existential hazards will soon be out of her intellectual depth.[436] On the other hand, a researcher can limit her field of inquiry and focus on a smaller piece

---

[430] "Research is perhaps inhibited by the multidisciplinary nature of the problem, but also by deeper epistemological issues." See Bostrom, 'Existential Risk Prevention' (n 15) 26.
[431] Phil Torres, 'Facing Disaster' (n 331) 21-22.
[432] ibid 22.
[433] Nindler writes: "Research dealing with the management of existential risks originating from novel technologies on a global level inevitably has to be interdisciplinary." See Nindler, 'The United Nation's Capability' (n 54) 7.
[434] Ord, *Precipice* (n 16) 7.
[435] ibid.
[436] See e.g., the preface in Häggström, *Here Be Dragons* (n 22).

of the puzzle. For instance, this thesis has used several research parameters. It has also utilised specific approaches for weeding out relevant from irrelevant legal and x-risk-related resources. The Part on Methods and Methodology in Chapter 1 examined the applicable approaches. In addition, the upcoming Chapters will focus on more specific issues. While this approach makes this thesis more manageable, it may also diminish its capacity "…to provide precisely the sort of insights needed to ensure a good outcome for our lineage."[437]

(vi) Other Features

Beyond the obscuring features mentioned above, it is also possible to suggest that other factors such as scientific illiteracy,[438] scientific worship,[439] irresponsible doomsday predictions in the past[440] and the connection between sci-fi and illustrations of human extinction[441] may undermine x-risk mitigation efforts.

2.5 CONCLUSION

From high-energy physics experiments to engineered pandemics and a future superintelligence, many scholars believe that the most pressing x-risks are caused by humanity's activities.

Chapter 2 has considered various aspects relating to the x-risk landscape. Such investigation is crucial because it is within this landscape that broad legal interventions, including the LHC-inspired intervention, will sit and operate. This Chapter has identified factors that may affect and pertain to any broad legal mechanism aiming to reduce the probability of an x-catastrophe.

---

[437] Phil Torres, 'Facing Disaster' (n 331) 21-22.

[438] Posner suggests laypeople may not always fully understand the science underlying the potential hazards, and that this may lead to more neglect of risk than fear. Posner, *Catastrophe* (n 90) 93.

[439] ibid 97.

[440] ibid 110.

[441] Movies, video games and comics often illustrate extinction events, but no such catastrophe has been seen in real life. In Posner's view, this likely alleviate fears as opposed to exaggerating them. ibid 100-109.

The upcoming Chapter will focus on selected aspects of the LHC Controversy, which involves high-energy physics experiments and postulated x-risks. It to be remember that the LHC Controversy is explored to see what can be learnt from the past in terms of deriving a legal intervention that will situate itself within the broader x-risk landscape outlined in this Chapter.

3 CHAPTER THREE: LEGAL ASPECTS OF THE LHC CONTROVERSY

3.1 BACKGROUND

Martin Rees opens Chapter 9 of *Our Final Century* with the following statement: "Some experiments could conceivably threaten the entire Earth. How close to zero should the claimed risk be before such experiments are sanctioned?"[442] Physics experiments are sometimes designed to generate extreme conditions not occurring naturally on Earth[443] or even in the universe.[444] It is not always possible to know or predict the outcome of such scientific experiments in advance. For, as noted by Rees, if the results are entirely predictable, there is no point behind the experimentation.[445] Going beyond the current state of knowledge and finding something interesting is often the essence of the scientific enterprise.[446]

It is recognised that new conditions raise new prospects, including the possibility of a large-scale catastrophe.[447] As mentioned in the previous two chapters, concerns have been raised that high-energy physics experiments, such as those facilitated by the LHC and RHIC, may trigger an x-catastrophe. As noted by Beyleveld and Brownsword: "Some maintained that the LHC should not have been built, because, for all we know, its use might destroy the world."[448]

Fortunately, virtually every physicist agree that it is improbable that experimental particle physics will trigger an x-catastrophe.[449] The experts have even moved away from a quantitative approach and avoided the word *probability*. Perhaps reflecting unwarranted mistrust or a healthy dose of scepticism, not everyone agreed with the

[442] Rees, *Our Final Century* (n 226) 115.
[443] Frank Wilczek, 'Big Troubles, Imagined and Real' in Nick Bostrom and Milan Ćirković (eds), *Global Catastrophic Risks* (OUP 2008) 346; Rees, *Our Final Century* (n 226) 116.
[444] Interactions, 'CUORE' (n 257).
[445] Rees, *Our Final Century* (n 226) 116.
[446] Johnson, 'Judicial Review' (n 68) 67; Francesco Calogero, 'Might a Laboratory Experiment Destroy Planet Earth?' [2000] Interdisciplinary Science Reviews 191, 192.
[447] Wilczek, 'Big Troubles, Imagined and Real' (n 443) 346.
[448] Deryck Beyleveld and Roger Brownsword, 'Emerging Technologies, Extreme Uncertainty, and the Principle of Rational Precautionary Reasoning' [2012] Law, Innovation and Technology 35, 55.
[449] Rees, *Our Final Century* (n 226) 118; Bostrom and Ćirković, 'Introduction' (n 12) 18; Wilczek, 'Big Troubles, Imagined and Real' (n 443) 347.

experts' reassurances and instead continued to speculate that some experiments could, at least in theory, destroy humanity's future.[450]

As far as the LHC is concerned, some individuals took it upon themselves to challenge this machine in court, arguing that its operation could destroy Earth and, in effect, humanity's future. The injunction requests that followed makes for a fascinating case study about the difficulties involved in stopping or halting a legitimate and tremendously expensive scientific endeavour based on speculative disaster theories.

3.1.1 Aims of Chapter 3

This Chapter considers different legal aspects of the LHC Controversy. It focuses on one of the court cases that arose following the concerns that the LHC could allegedly cause an x-catastrophe. Furthermore, it will consider legal aspects relating to the primary 'risk originator' in the LHC context. The thesis uses the term risk originator to refer to the actor who creates the x-risk. The primary risk originator in the LHC context is the European Organization for Nuclear Research ('CERN'). Given that broad legal interventions will affect x-risk originators, it is interesting to consider, in some detail, the actor who created the risk in the LHC context.

A central aim of Chapter 3 is to consider the potential 'review barriers' that may complicate different modes of external review in relation to the LHC Controversy. It should be remembered that the LHC-inspired intervention should satisfy three self-imposed conditions. One of those is that the intervention should, to some extent, respond to these barriers. Ultimately, when exploring how the LHC Controversy can inspire the design of a broad legal intervention, it is practical to consider why it is tricky to resolve this Controversy via the legal system.

Chapter 3 will shine light on the following sub-question: What makes it complicated for an external party to address the disagreements and uncertainty about whether particle accelerator experiments pose an x-risk?

[450] Rees, *Our Final Century* (n 226) 116; Posner, *Catastrophe* (n 90) 30-32.

3.2 PRELIMINARY ASPECTS AND OVERVIEW

To uncover how the universe works, physicists have built complex machines – particle accelerators.[451] According to CERN, the LHC is currently the most powerful accelerator on the planet.[452]

It is located in an underground tunnel near Geneva and consists of a 27-km ring of superconducting magnets with accelerating structures to boost the energy of the particles along the way.[453] Inside the accelerator, two beams of particles travel in the opposite direction and are made to collide at enormous speed.[454] The LHC forms part of CERN's accelerator complex.[455] The machine will reportedly help to address many unanswered questions about the universe.[456] Supporters assert the LHC is part of the pursuit of understanding nature at its most fundamental level.[457] The quest for a more profound knowledge did not, however, come without warnings.

Some individuals feared that operating the LHC could destroy Earth.[458] These concerns crystallised into lawsuits which sought to delay the LHC from commencing its collisions. Without much success, scientists in this camp sought relief in domestic as well as international courts.[459] In *Sancho v. U.S Department of Energy*, (the *Sancho case*) two individuals sought to enjoin federal agencies in the US as well as CERN from operating the LHC.[460] They claimed that the LHC experiments could trigger a set

[451] ibid 119; CERN, 'Our Mission' <https://home.cern/about/who-we-are/our-mission> accessed 27 April 2023.
[452] CERN, 'The Large Hadron Collider' (n 5).
[453] CERN, 'Facts and Figures About the LHC' (n 5); CERN, 'The Large Hadron Collider' (n 5).
[454] CERN, 'The Large Hadron Collider' (n 5).
[455] The CERN accelerator complex is a succession of machines. Each machine accelerates a beam of particles to a given energy before injecting the beam into the next machine in the chain. The next one brings the beam to an even higher energy and so on. The LHC is the last element of this chain – it is where the beams reach their highest energies. See CERN, 'LHC Season 2: Facts & Figures' <https://run2-13tev.web.cern.ch/background/lhc-season-2-facts-figures> accessed 1 February 2020.
[456] CERN, 'Facts and Figures About the LHC' (n 5).
[457] Johnson, 'The Black Hole Case' (n 68) 827.
[458] Other accelerators have been the subject of similar concerns. See Rees, *Our Final Century* (n 226) 120.
[459] Wilson, 'Minimizing' (n 151) 335.
[460] See Complaint for Temporary Restraining Order, Preliminary Injunction, and Permanent Injunction (No. 00136-HG-KSC Civ. 08) [hereinafter *Sancho complaint*]; *Sancho case* (n 87); See also *Sancho v US Department of Energy* 392 Fed Appx 610 (9th Cir 2010) [hereinafter *Sancho appeal*].

of irreversible processes leading to the destruction of Earth.[461] While their claim was not formulated in the language of existential risk or catastrophe, it is clear that the alleged disaster scenarios are consistent with the notion of existential catastrophe. The court ultimately dismissed their complaint and the subsequent appeal.

As previously stated, most physicists consider it improbable that experiments carried out in particle accelerators will cause an existential disaster.[462] In the context of the LHC, two separate safety assessments (and other studies) have concluded that there is no need to worry.[463] Despite such reassurances, it remains worthwhile to analyse the LHC Controversy, with a focus on the LHC.[464] This point has been made above. In short, valuable lessons can be derived from the LHC Controversy regardless of whether the objective risk is zero in this context. Indeed, many x-risk-related resources refer to high-energy physics experiments when discussing hazards that may threaten to destroy humanity's future. These works often recognise that it is improbable that such experiments will cause a catastrophe. Still, very recent works continue to, at least, mention high-energy physics experiments.[465] With the above in mind, the upcoming Part will consider the risk originator in the LHC context.

3.3 THE EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

CERN is the primary x-risk originator in the LHC context, and one of the defendants in the *Sancho complaint*.[466] Whereas the x-risk landscape may eventually consist of various originators like individuals, companies and organisations, it is interesting to consider some legal aspects of the actual risk originator in the LHC context.

[461] *Sancho case* (n 87) 1259-1261.
[462] Bostrom and Ćirković, 'Introduction' (n 12) 18.
[463] John Ellis and others, 'Review of the Safety of LHC Collisions' (LHC Safety Assessment Group 2008) <http://lsag.web.cern.ch/LSAG-Report.pdf> accessed 27 April 2023 [hereinafter LSAG Report].
[464] Bostrom and Ćirković, 'Introduction' (n 12) 5.
[465] Boyd and Wilson, 'Existential Risks to Humanity' (n 154) 2307; Baum, 'Quantifying the Probability' (n 125); Cotton-Barratt, Daniel and Sandberg, 'Defence' (n 37) 274; Ord, *Precipice* (n 16) 325; Beard, Rowe and Fox, 'An Analysis and Evaluation of Methods Currently Used to Quantify the Likelihood of Existential Hazards' (n 124); Nick Bostrom, 'The Vulnerable World Hypothesis' [2019] Global Policy 455, 462.
[466] See *Sancho complaint* (n 460); See also the *Sancho case* (n 87).

There are two other reasons why looking at some legal aspects of CERN can prove valuable. Firstly, a post-LHC era will likely dawn in the future.[467] New accelerators, some of which may be hosted by CERN, will perhaps rekindle the concerns that haunted particle accelerators in the past. And so, the LHC may not be the last accelerator to face legal obstacles.[468] Secondly, future concerns might arise that the scientific activities of other international organisations (IOs) pose an x-risk. A problem in either of these situations is that IOs often have things in common,[469] which can make it difficult to challenge their activities in domestic courts. National courts frequently decline to exercise jurisdiction over disputes involving international organisations by adopting one or more avoidance techniques.[470] Reinisch explores such techniques in his book.[471] Immunity from legal process is cited as one, but not the only reason in this context.[472]

3.3.1 The Organisation

CERN is an IO.[473] Given its status as such, some implications follow, including the ability to bring claims against it in domestic courts. CERN is one of the major European

[467] Rüdiger Voss, 'CERN: A global project' (2017) 878 Journal of Physics: Conference Series <https://iopscience.iop.org/article/10.1088/1742-6596/878/1/012001> accessed 27 April 2023; CERN, 'A New Schedule for the LHC and its Successor <https://home.cern/news/news/accelerators/new-schedule-lhc-and-its-successor> accessed 27 April 2023; CERN, 'High-Luminosity LHC' <https://home.cern/science/accelerators/high-luminosity-lhc> accessed 27 April 2023; CERN, 'International Collaboration Publishes Concept Design for a Post-LHC Future Circular Collider at CERN' <https://home.cern/news/press-release/accelerators/international-collaboration-publishes-concept-design-post-lhc> accessed 27 April 2023; CERN, 'Future Circular Collider' <https://home.cern/science/accelerators/future-circular-collider> accessed 27 April 2023.

[468] Adams, 'Honey' (n 309) 154; Marshall Chance Peterson, 'The "Sancho" Effect: Why the Large Hadron Collider Won't Destroy the World, and how it Could Improve Science in the Courts' [2014] American Bar Association 303, 315.

[469] Jan Klabbers, *An Introduction to International Organizations Law* (3rd edn, Cambridge University Press 2015). For example, they generally enjoy immunity from legal process.

[470] August Reinisch, *International Organizations Before National Courts* (Cambridge University Press 2000) 35.

[471] ibid ch 2.

[472] ibid 35-36.

[473] These organisations are *commonly* (1) created by states where membership is limited to states or governments, (2) established by treaty, (3) governed by international law, and (4) in possession of organs distinct from the members. See Klabbers, *An Introduction to International Organizations Law* (n 469) 9-10; Chittharanjan Amerasinghe, *Principles of the Institutional Law of International Organizations* (2nd edn, Cambridge University Press 2005) 9-10; Stephen Bouwhuis, 'The International Law Commission's Definition of International Organizations' [2012] International Organizations Law Review 451, 452-453.

scientific organisations,[474] and its primary research is in fundamental particle physics.[475] CERN itself states that curiosity is its "raison d'être".[476] To expand the boundaries of knowledge, it provides accelerator facilitates to researchers around the world.[477] The Organisation's seat is in Geneva,[478] with French territory hosting parts of CERN's installations.[479] The effect is two separate headquarter agreements.

It has been noted that scientific progress can influence states' capacity to compete in a global economy.[480] That being said, many scientific operations require considerable investments – both human and capital.[481] Thus, some states can only participate in the global science race by combining assets through an IO like CERN.[482]

The origins of CERN can be traced to the 1940s.[483] Following the Second World War, some scientists imagined the creation of a European atomic physics laboratory.[484] Such an undertaking, it was believed, would unite scientists and help to share the increasing costs of nuclear physics facilities.[485] In December 1949, French physicist Louis de Broglie presented the first official proposal for creating such a laboratory at the European Cultural Conference in Lausanne.[486] A further push came in June 1950, at the fifth UNESCO General Conference in Florence, when physicist Isidor Rabi

[474] Marc Cogen, 'Membership, Associate Membership and Pre-Accession Arrangements of CERN, ESO, ESA, and EUMETSAT' [2012] International Organizations Law Review 145, 155.
[475] CERN, 'Fundamental Research', <https://home.cern/about/what-we-do/our-research> accessed 28 April 2023.
[476] ibid.
[477] CERN, 'Our Mission' (n 451).
[478] Convention for the Establishment of a European Organization for Nuclear Research (adopted 1 July 1953, entered into force 29 September 1954) 200 UNTS 149 (CERN Convention) art I [hereinafter CERN Convention].
[479] Marc Cogen, *An Introduction to European Intergovernmental Organizations* (Ashgate Publishing 2015) 189.
[480] Cogen, 'Membership, Associate Membership and Pre-Accession Arrangements' (n 474) 146.
[481] ibid.
[482] ibid.
[483] CERN, 'Where Did it all Begin?' <https://home.cern/about/who-we-are/our-history> accessed 28 April 2023; For an overview of the history of CERN, see Cogen, *An Introduction to European Intergovernmental Organizations* (n 479) 189-190.
[484] CERN, 'The History of CERN' <https://timeline.web.cern.ch/taxonomy/term/89> accessed 28 April 2023.
[485] ibid.
[486] ibid; Cogen, 'Membership, Associate Membership and Pre-Accession Arrangements' (n 474) 148-149.

mooted for UNESCO to assist and encourage the formation of regional research laboratories to increase scientific collaboration.[487]

At a UNESCO meeting in December 1951, the first resolution concerning the establishment of a European Council for Nuclear Research was adopted.[488] The provisional Council (Conseil Européen pour la Recherche Nucléaire) was established two months later, from which the Organisation derives (and retains) its acronym.[489] The sixth session of the Council witnessed the signing of the CERN Convention.[490] Its full title is the *Convention for the Establishment of a European Organization for Nuclear Research*. Following the ratification by France and Germany, the Organisation officially came into being on 29 September 1954.[491]

(i) The Convention

Article II of the CERN Convention relates to the Organisation's purposes. It confirms that CERN shall provide for collaboration among European states in nuclear research of a pure scientific and fundamental character.[492] In doing so, the Organisation shall construct and operate international laboratories for research on high-energy particles.[493] Each laboratory shall include one or more particle accelerators.[494] The members of CERN must participate in at least one of the programmes of activities which form part of the *basic programme*.[495] Each member state must signify to the President of the CERN Council those programmes of activities in which it wishes to participate.[496] The basic programme of the 1954 CERN Convention has expanded

[487] CERN, 'The History of CERN' (n 484); Cogen, 'Membership, Associate Membership and Pre-Accession Arrangements' (n 474) 149.
[488] CERN, 'Origins' <https://timeline.web.cern.ch/origins> accessed 28 April 2023; Cogen, 'Membership, Associate Membership and Pre-Accession Arrangements' (n 474) 149.
[489] CERN, 'The History of CERN' (n 484).
[490] ibid.
[491] CERN Convention (n 478); CERN, 'The History of CERN' (n 484).
[492] CERN Convention (n 478) art II(1). This article confirms that CERN shall have no concern with work for military requirements as well as publish or otherwise make generally available the results of its experimental and theoretical work.
[493] CERN Convention (n 478) art II(2)(a)-(b).
[494] ibid art II(2)(a)(i-iii).
[495] ibid art III(3).
[496] ibid art III (3).

over the years.[497] The basic programmes currently include the Proton Synchrotron and the LHC.[498] It was in December 1994 that the Council decided to "…include the Large Hadron Collider (LHC) project in the Basic Programme of the Organisation…"[499]

On a general note, the CERN Convention has served as a successful legal model for other European scientific organisations.[500] This further reenforces the argument that there are good reasons for considering the risk originator in the LHC context.

(ii) Legal Personality and Privileges and Immunities

Organisations like CERN rank amongst the subjects ('legal persons') of international law.[501] Since this Chapter examines the domestic case of *Sancho v. U.S Department of Energy*, it is appropriate to distinguish CERN's *international* and *domestic* legal personality. Notably, any broad legal intervention utilised in or by domestic courts may be of limited relevance when the risk originator is an IO such as CERN.

Only those organisations having a domestic legal personality are exposable to judicial proceedings in national courts.[502] An entity that does not legally exist in a particular system cannot bring or receive claims before the courts of that particular system.[503] The constituent instrument can – and many do – grant legal capacity to the organisation in its member states.[504] For instance, the CERN Convention confirms that

[497] Cogen, 'Membership, Associate Membership and Pre-Accession Arrangements' (n 474) 163.
[498] ibid; Cogen, *An Introduction to European Intergovernmental Organizations* (n 479) 191.
[499] Council, 'Approval of the Large Hadron Collider (LHC) Project' (16 December 1994) CERN/2075/ Final <http://cds.cern.ch/record/33618?ln=en> accessed 28 April 2023.
[500] Cogen, 'Membership, Associate Membership and Pre-Accession Arrangements' (n 474) 176; See Voss, 'CERN: A global project' (n 467).
[501] For CERN's international personality, see Protocol on the Privileges and Immunities of the European Organization for Nuclear Research (adopted 18 March 2004, entered into force 22 February 2007) 2432 UNTS 110 (CERN Protocol) art 2 [hereinafter CERN Protocol].
[502] Reinisch, *International Organizations Before National Courts* (n 470) 38.
[503] ibid.
[504] Amerasinghe, *Principles of the Institutional Law of International Organizations* (n 473) 69. In such cases, the *members* are under an obligation to recognise the organisation's personality in their legal systems. See Reinisch, *International Organizations Before National Courts* (n 470) 41.

CERN enjoys legal personality in its member states.[505] An agreement from 2004 also confirms CERN's domestic personality.[506]

An organisation's constituent instrument does not affect the status of the organisation in non-member states.[507] Nevertheless, courts in states which are not members can find alternative ways to recognise the domestic personality of organisations.[508] The details are largely irrelevant here because bringing a claim against CERN is not an issue of personality.[509] The core issue relates to jurisdiction and CERN's treaty-based privileges and immunities.

It has been said that in order to function efficiently,[510] IOs require minimum standards of freedom and legal security for their assets, headquarters and establishments.[511] As a result, they often enjoy privileges and immunities.[512] Four broad categories commonly attach to the IO itself.[513] These are (1) immunity from legal process, (2) inviolability of the organisation's premises and archives, (3) protections relevant to assets, currency and other fiscal matters, and (4) freedom of communication.[514] The agreements between CERN and its host as well as member states reflect these categories. For a brief illustration, see Table 3.1 and 3.2 below:

[505] CERN Convention (n 478) art IX.
[506] CERN Protocol (n 501) art 2.
[507] A treaty will not create obligations or rights for third-party states without their consent. Vienna Convention on the Law of Treaties (adopted 23 May 1969, entered into force 27 January 1980) 1155 UNTS 331 (VCLT) art 34.
[508] Amerasinghe, *Principles of the Institutional Law of International Organizations* (n 473) 70-71 discusses how national courts can use their conflict of laws principles to recognise the *domestic* personality of an organisation because it possesses *international* personality.
[509] Reinisch writes that declining to exercise jurisdiction based on non-recognition of the organisation's domestic personality is rare. Reinisch, therefore, deems this method to be the 'most radical' one for avoiding disputes involving organisations. See Reinisch, *International Organizations Before National Courts* (n 470) 37-38.
[510] Michael Singer, 'Jurisdictional Immunity of International Organizations: Human Rights and Functional Necessity Concerns' [1995] Virginia Journal of International Law 53, 56; Amerasinghe, *Principles of the Institutional Law of International Organizations* (n 473) 315-316.
[511] Crawford, *Brownlie's Principles of Public International Law* (n 408) 171.
[512] See e.g., Reinisch, *International Organizations Before National Courts* (n 470) 13-14.
[513] Crawford, *Brownlie's Principles of Public International Law* (n 408) 175-176.
[514] ibid.

Table 3.1

The Headquarter Agreement (CERN – Switzerland)[515]	
Article(s)	**Immunity/Privilege**
6	(1) Legal process
3, 4	(2) Inviolability
6, 8, 9	(3) Property, Currency, Fiscal matters
10, 11	(4) Communication

Table 3.2

The Protocol (CERN – All Member States Except for its Host States)[516]	
Article(s)	**Immunity/Privilege**
5	(1) Legal process
3, 4	(2) Inviolability
6, 7	(3) Property, Currency, Fiscal matters
8	(4) Communication

3.3.2 The Machine

The LHC has a derivative name. It is *large*, accelerating either protons or ions which belong to the group of particles called *hadrons* and making these particles *collide*.[517] The construction of the LHC required 1,170 price enquiries and tender invitations, the negotiation, drafting and placing of 115,700 purchase orders and 1,040 contracts, and the commitment of 6,364 different contractors and suppliers.[518]

[515] Agreement Between the Swiss Federal Council and the European Organization for Nuclear Research Concerning the Legal Status of that Organization in Switzerland (11 June 1955, entered into force 3 May 1955) 249 UNTS 405 (Swiss Headquarter Agreement).
[516] CERN Protocol (n 501).
[517] CERN, 'LHC: The Guide' (CERN-Brochure-2017-002-Eng 2017) 15 <https://home.cern/sites/home.web.cern.ch/files/2018-07/CERN-Brochure-2017-002-Eng_0.pdf> accessed 28 April 2023 [Hereinafter LHC Guide].
[518] CERNCOURIER, 'Procurement at the Forefront of Technology' (31 August 2018) <https://cerncourier.com/a/procurement-at-the-forefront-of-technology/> accessed 28 April 2023.

The thesis will not describe the technical workings of the LHC in detail. There is a wealth of information online, notably on CERN's website.[519] CERN has published a guide comprising a collection of facts and figures about the LHC.[520] Those interested in gaining a technical understanding can start by consulting this guide. The LHC achieves energies that no other accelerator has reached before.[521] It is by using such a powerful machine that the physicists can venture further towards understanding the mysteries of the universe.[522] The guide suggests that humanity's understanding is incomplete.[523] Relevant here is the Standard Model – a collection of theories embodying the current understanding of fundamental particles and forces.[524] Whereas the Standard Model is highly successful in many regards, it leaves some questions unsolved; the LHC helps to elucidate these.[525]

The Chapter's background explained that experimental research strives to go beyond what is currently known. Given the inability to forecast the precise outcome in advance, uncertainty will surround the scientific enterprise. This is not negative as noted by Johnson because it aligns with the aspiration to explore and discover.[526] Uncertainty also underpins the LHC where the desire to find something fascinating motivates the experiments and justifies the costs.[527] The LHC is, no doubt, part of a fascinating quest to expand humanity's knowledge.[528] However, the machine is primarily designed for pure science with useful spinoffs being the exception rather than the rule.[529]

---

[519] The webpage contains many key documents. However, a technical insight into the machine is not too relevant for the purpose of understanding the points made in this Chapter.
[520] LHC Guide (n 517).
[521] ibid 54.
[522] ibid.
[523] ibid 22.
[524] ibid 6; CERN, 'The Standard Model' <https://home.cern/science/physics/standard-model> accessed 28 April 2023; Pallab Ghosh, 'What next for the Large Hadron Collider?' *BBC* (5 April 2015) <https://www.bbc.co.uk/news/science-environment-32106990> accessed 28 April 2023.
[525] LHC Guide (n 517) 6, 22.
[526] Johnson, 'Judicial Review' (n 68) 67.
[527] See e.g., Calogero, 'Might a Laboratory Experiment Destroy Planet Earth?' (n 446) 192; Peterson, 'The "Sancho" Effect' (n 468) 305.
[528] Ghosh, 'What next for the Large Hadron Collider?' (n 524).
[529] Johnson, 'The Black Hole Case' (n 68) 827-828; This is not to say that CERN's research is not relevant to the development of new technologies. For instance, the World Wide Web is a CERN technology - invented to allow scientists to share information. See CERN, 'Fundamental Research' (n 475).

Exploring CERN's website reveals that the pursuit of knowledge is the principal reason for operating the machine. This aspect is noteworthy because a risk arguably becomes less tolerable when the direct reward is predominantly limited to pure science. Conversely, a risk might be easier to justify if taking it may come with large benefits.[530] From a theoretical perspective, it is possible to question whether an x-risk is ever worth taking when the immediate benefit is privy to a small community of researchers. The question is asked from a theoretical perspective because it foregoes quantitative aspects. By avoiding quantitative conclusions in the form of probabilistic estimates of the LHC risks materialising, it is tricky to engage in cost-benefit analysis at all.

3.4 THE DISASTER SCENARIOS

It has been mentioned before that the LHC is not the only accelerator that has generated concerns about the possible catastrophic outcomes of physics experiments. Another famous machine is the RHIC at the Brookhaven National Laboratory (BNL).[531] The safety concerns associated with RHIC led the director of the BNL to commission a risk assessment.[532] According to the published version, the "…catastrophe scenarios at RHIC are firmly excluded by compelling arguments based on well-established physical laws."[533]

The LHC has also been the subject of safety assessments, notably the LSAG Report.[534] The LHC Safety Assessment Group (LSAG) prepared this report. It reviews an earlier assessment from the LHC Safety Study Group (LSSG), which had concluded that collisions at the LHC presented no danger. Building upon additional experimental results and theoretical understanding, the LSAG Report confirmed, updated, and extended the LSSG's conclusions.[535] One issue with these reviews is

[530] See the discussion in Baum, 'The Great Downside Dilemma' (n 22).
[531] See Calogero, 'Might a Laboratory Experiment Destroy Planet Earth?' (n 446) 191; Posner, *Catastrophe* (n 90) 30.
[532] ibid 31.
[533] Robert Jaffe and others, 'Review of Speculative "Disaster Scenarios" at RHIC' [2000] Reviews of Modern Physics 1125.
[534] LSAG Report (n 463).
[535] ibid.

their deep complexity. For non-physicists, the theories and evidence relied upon by these assessments are not, as noted by Posner, fully intelligible.[536] Such complexities can shield those who may wish to downplay the risk.

The three main theorised disaster scenarios linked to the LHC are as follows: The production of (a) Strangelets, (b) Micro Black Holes, and (c) Magnetic Monopoles. These are addressed in the LSAG Report and cited in the *Sancho complaint*.[537] There is little point in discussing these speculative scenarios in detail. As already mentioned, the objective risk of the LHC causing an x-catastrophe is largely irrelevant for the purpose of the broad legal intervention in this thesis. The LHC-inspired intervention is also disconnected from the adjudication of competing scientific theories.

The LSAG Report contains a discussion of these scenarios. Those without a technical background may instead want to consult Adams's article. It is written from a legal perspective and seeks to provide "…a layman's description of each of these phenomena and briefly describe some of the major arguments for and against these theories."[538]

3.5 THE SANCHO CASE

The theorised disaster scenarios crystallised into lawsuits aimed at delaying the LHC from commencing its operations. This Part considers one such dispute in detail: *Sancho v. U.S Department of Energy*. As seen below, this case was never heard on the merits. Neither was the subsequent appeal.

3.5.1 Background

In March 2008, Luis Sancho and Walter Wagner filed a complaint in the United States District Court against the U.S Department of Energy, Fermilab, the National Science Foundation and CERN.[539] They sought to halt the defendants from operating the LHC

[536] Posner, *Catastrophe* (n 90) 195.
[537] *Sancho case* (n 87).
[538] Adams, 'Honey' (n 309) 139-146.
[539] *Sancho complaint* (n 460) 1-2.

since doing so could, according to the claimants, trigger irreversible processes leading to the destruction of Earth.[540] The claimants argued that the LHC would create conditions never previously existent on Earth. They maintained that competing theories of physics predicted different outcomes from the LHC collisions, with no agreement amongst physicists as to the exact outcome.[541] The complaint set out three theories dealing with the potential adverse consequences of the LHC experiments. These correspond to the disaster scenarios mentioned above: (a) Strangelets, (b) Micro Black Holes, and (c) Magnetic Monopoles. Under these, operating the LHC could allegedly lead to the end of humanity[542] meaning that the claimants asked the court to address an x-risk scenario (as defined in this thesis).

Sancho and Wagner claimed that scientific publications had articulated these theories. They also noted the lack of absolute refutation of the theorised disaster scenarios,[543] and criticised the perfunctory safety reviews that sought to falsify them. The claimants even argued that the reviews contained fundamental flaws.[544] In support of their assertions, Sancho and Wagner filed affidavits dealing with the perceived safety and ethical flaws. The claimants proclaimed themselves (and their associates) "…experts in physics and other fields of science, technology and ethics who are capable of reviewing…" the safety reviews for flaws and errors.[545] At the time of the complaint, the LSAG Review was not available to the public, and the claimants wanted around four to six months to review the pending review to determine whether it was, once again, fundamentally flawed.[546] Subsequently released and submitted by the defendants as an attachment to the court,[547] the LSAG report confirmed the earlier finding that there was no basis for any conceivable threat from the LHC.[548]

The legal basis for the claim is hard to identify, but the appropriate remedy is less obscure: The claimants wanted to obtain an injunction. But their requests failed in the

[540] *Sancho case* (n 87).
[541] *Sancho complaint* (n 460) 3.
[542] ibid 3-4.
[543] ibid 4-5.
[544] ibid.
[545] ibid 5.
[546] ibid.
[547] Peterson, 'The "Sancho" Effect' (n 468) 306.
[548] LSAG Report (n 463) 2.

US. Likewise, they would probably have been unsuccessful in obtaining a preliminary or permanent injunction as far as the English remedial landscape is concerned. Injunctions are dealt with in detail in Chapter 4.

3.5.2 Violations

An overarching question concerns the legal basis upon which the claimants relied. Sancho and Wagner claimed that the defendants failed to comply with the obligations as required by the National Environmental Policy Act (NEPA).[549] Additionally, the complaint maintained that the defendants failed to adhere to the European Council's Precautionary Principle.[550]

(i) National Environmental Policy Act

NEPA deals with environmental policies in the US.[551] One principal purpose is to declare a national policy which will encourage harmony between humans and the environment.[552] NEPA requires federal agencies to assess the environmental effects of their proposed actions before making decisions.[553] In other words, NEPA integrates environmental policy considerations into federal government decision-making.[554]

As mentioned previously, two defendants in the LHC context were federal agencies—the United States Department of Energy and the National Science Foundation.[555] Whereas NEPA includes various procedural requirements, only some parts merit closer examination for the purpose of this thesis.[556] In particular, NEPA stipulates that all federal agencies shall:

---

[549] *Sancho complaint* (n 460) 6-9.
[550] ibid 9-10.
[551] Council on Environmental Quality, 'A Citizen's Guide to the NEPA: Having Your Voice Heard' (December 2007) 2 [hereinafter NEPA Guide].
[552] National Environmental Policy Act of 1969, 42 U.S.C. § 4321 (2018) [hereinafter NEPA].
[553] NEPA Guide (n 551) 2.
[554] Matthew Villmer, 'Procedural Squabbling Ahead of Global Annihilation: Strengthening the National Environmental Policy Act in a New Technological Era' [2010] Florida Coastal Law Review 321, 321.
[555] *Sancho case* (n 87) 1261.
[556] See also Villmer, 'Procedural Squabbling' (n 554) 324.

include in every recommendation or report on proposals for legislation and other major Federal actions significantly affecting the quality of the human environment, a **detailed statement** by the responsible official on—

(i) the environmental impact of the proposed action,

(ii) any adverse environmental effects which cannot be avoided should the proposal be implemented,

(iii) alternatives to the proposed action,

(iv) the relationship between local short-term uses of man's environment and the maintenance and enhancement of long-term productivity, and

(v) any irreversible and irretrievable commitments of resources which would be involved in the proposed action should it be implemented.[557]

This statement is called an Environmental Impact Statement ('EIS').[558] The point is clear: Agencies must prepare an EIS when there are major Federal actions significantly affecting the quality of the human environment. According to Villmer: "When private citizens or environmental organizations hear of a federal agency's construction or testing plans, they often file a NEPA violation suit if the government failed to prepare an EIS."[559]

If the answer is not readily apparent, the agency may prepare an environmental assessment (EA) to determine whether a full EIS is necessary.[560] If the proposed action does not significantly affect the human environment, the agency can issue a Finding of no Significant Impact ('FONSI').[561] Thus, the EA provides a basis for deciding whether to conclude with a FONSI or prepare a full EIS.[562]

[557] NEPA (n 552) § 4332(C) (2018).
[558] *Sancho case* (n 87) 1265; Villmer, 'Procedural Squabbling' (n 554) 324.
[559] ibid 326.
[560] ibid 324; *Sancho case* (n 87) 1265.
[561] ibid 1265.
[562] NEPA Guide (n 551) 8 and 11-12.

With this overview in mind, it is to be noted that the *Sancho complaint* maintained that the Federal defendants and CERN engaged in a partnership to construct the LHC.[563] Implicit in the complaint is that the Federal defendants' involvement triggered the NEPA requirements. In other words, by omitting to prepare an EA concluding with a FONSI or a full EIS, the defendants allegedly failed to comply with NEPA.[564] Among other things, the claimants requested a preliminary injunction enjoining the defendants from operating the LHC until completion of either a FONSI or EIS.[565]

The court held that NEPA did not apply. In effect, the court lacked subject-matter jurisdiction to adjudicate the claim.[566] It is relevant to note why the Court reached this conclusion. As outlined above, NEPA applies to *major* Federal actions.[567] As explained by judge Gillmor:

> The jurisdiction of the Court…depends on whether Federal Defendants have undertaken a "major Federal action" with respect to the construction of the LHC. To determine if the Court does have jurisdiction under NEPA, the Court must examine two factors: (1) the amount and nature of Federal Defendants' funding, and (2) the extent of Federal Defendants' involvement and control.[568]

Having analysed these factors, the Court concluded that the Federal defendants' involvement with the LHC did not qualify as a major Federal action.[569] The defendant submitted that an International Cooperation Agreement from 1997 with CERN governed their involvement with the LHC.[570] They argued and gave evidence to the effect that their involvement and total funding was limited and that CERN had sole responsibility for the construction, operation and management of the LHC.[571]

According to Villmer, the rush to commence the LHC experiment, together with the multibillion-dollar expenditure, virtually guaranteed a favourable judicial decision,

[563] *Sancho complaint* (n 460) 3.
[564] ibid 6-9.
[565] ibid 11.
[566] *Sancho case* (n 87) 1268.
[567] ibid 1265; See also NEPA (n 552).
[568] *Sancho case* (n 87) 1266.
[569] ibid 1268.
[570] ibid 1262.
[571] ibid 1262-1263.

despite possible NEPA noncompliance.[572] But what if NEPA had applied, and so imbued the court with subject-matter jurisdiction? Relevant to this question is the observation that the defendants maintained that the proposed LHC experiments had been thoroughly reviewed and were completely safe.[573] In effect, the best-case scenario likely entailed a delay in starting the machine until the defendants complied with NEPA. Since the defendants would have continued to insist that the LHC was safe, NEPA would have led to a circular dispute. As proposed by Johnson, NEPA is not a useful constraint because the defendants need only demonstrate that they have complied with the formality requirements of the law.[574]

What is more, there existed no effective control over the organisation's operations.[575] A decision favouring the claimants in the US would, therefore, not have prevented CERN from operating the LHC in Europe.[576] However, Peterson suggests that the LHC operations could potentially have been impacted by the issuance of an injunction since Fermilab is assisting with parts of the operations.[577]

Let us momentarily presume that a decision in favour of the claimants would have convinced CERN to await the federal defendants' compliance with NEPA. It has been recognised that such a delay might have brought about public opposition followed by a permanent stop to the LHC through political means.[578] This shifts the focus from the courts to the political process. On this note, the courts may potentially dismiss similar complaints as constituting a political as opposed to judicial matter. For example, judge Gillmor stated that Congress did not enact NEPA to allow the LHC debate to proceed in federal courts.[579] Whereas she recognised that the action reflected disagreement among scientists about the ramifications of operating the LHC, she explained that NEPA does not give citizens a general opportunity to air policy objections to federal

[572] Villmer, 'Procedural Squabbling' (n 554) 335.
[573] *Sancho case* (n 87) 1261.
[574] Johnson, 'Agencies and Science-Experiment Risk' (n 68) 565*; Hawaii County Green Party v Clinton*, 980 F Supp 1160 (D Haw 1997) 1167-1169.
[575] *Sancho appeal* (n 460) 611-12.
[576] ibid.
[577] Peterson, 'The "Sancho" Effect' (n 468) 307.
[578] Johnson, 'The Black Hole Case' (n 68) 862.
[579] *Sancho case* (n 87) 1269.

actions.[580] The political process was said to provide the appropriate forum in which to air such disagreements.[581]

(ii) Precautionary Principle

The *Sancho complaint* also argued the defendants failed to adhere to the requirements of the European Council's Precautionary Principle and the European Commission's "Science and Society Action Plan".[582] The court was not impressed by these arguments. In September 2008, it granted the Motion to Dismiss the claim that the United States was bound by international law or the agreements highlighted in the complaint. As explained by the court, these had not been incorporated into domestic law by international treaty or otherwise.[583] For the sake of argument, it is worthwhile to briefly entertain the essence of the claim that the defendants failed to adhere to the principle in EU law.

The precautionary principle can be understood as a guide to coping with scientific uncertainties in the management of risk.[584] In the EU, the Treaty on the Functioning of the European Union mentions the principle in Article 191.[585] However, its scope goes beyond environmental protection, and the European Institutions are said to have played a vital role in fleshing out the principle.[586] A starting point is the Commission's

[580] ibid.
[581] ibid; *Metropolitan Edison Co v People Against Nuclear Energy,* 460 U.S. 766, 777, 103 S.Ct. 1556, 75 L.Ed.2d 534 (1983).
[582] *Sancho complaint* (n 460) 9-10.
[583] *Sancho case* (n 87) 1260.
[584] Petrenko and McArthur recognise the difficulties in agreeing on formulating the principle. They also explore different formulations and variations of the principle. See Anton Petrenko and Dan McArthur, 'Between Same-Sex Marriages and the Large Hadron Collider: Making Sense of the Precautionary Principle' [2010] Science and Engineering Ethics 591, 591-592; According to one study, the precautionary principle acknowledges the limits of scientific understanding, and the difficulties of decision-making when conclusive evidence cannot be produced. See 'Study on the Precautionary Principle in EU Environmental Policies: Final Report' (prepared by Milieu Ltd for the European Commission's DG Environment, 2017) 6 [hereinafter Study on the Precautionary Principle]; See also Communication from the Commission on the Precautionary Principle (OM/2000/0001 final) [hereinafter Commission Communication].
[585] Consolidated Version of the Treaty on the Functioning of the European Union [2012] OJ C326/47 (TFEU).
[586] Study on the Precautionary Principle (n 584) 19.

Communication on the precautionary principle (2000) which set out guidelines on the principle's application.[587] A study from 2017 made the following observation:

> The Communication provides that the precautionary principle may be invoked when a phenomenon, product or process may have a dangerous effect, identified by scientific and objective evaluation, if this evaluation does not allow the risk to be determined with sufficient certainty.[588]

The claimants argued there was a lack of absolute refutation of the adverse outcomes of the LHC experiments.[589] They also claimed the relevant safety reviews were fundamentally flawed. The prime risk, they argued, derived from the lack of an ethical governance institution for independent government regulation of the LHC risk.[590]

However, some have noted that numerous studies have discredited the disaster scenarios.[591] As opposed to failing to adhere to the principle, CERN's actions arguably exemplified a proper precautionary approach to the alleged risk scenarios.[592] The Commission's Communication also maintains that it is necessary to clarify the distinction between reliance on the precautionary principle and the search for 'zero risk', something which, as noted, is rarely found in reality.[593] The overall approach by CERN arguably aligned with the Commission's approach in the Communication.

Petrenko and McArthur provide a relevant analysis.[594] Whereas the authors explore the formulation of a refined version of the principle, some parts of their analysis provide support for the conclusion that a court might accept that CERN demonstrated a proper precautionary approach to the alleged risk scenarios. Interestingly, they also use the LHC as a test case for their refined version of the principle.

---

[587] ibid 19-20; Commission Communication (n 584).
[588] ibid 3; Study on the Precautionary Principle (n 584) 6, 20, 30, 76.
[589] *Sancho complaint* (n 460) 4-5.
[590] ibid 10.
[591] Petrenko and McArthur, 'Between Same-Sex Marriages and the Large Hadron Collider' (n 584) 606.
[592] ibid 607.
[593] Commission Communication (n 584) 8.
[594] Petrenko and McArthur, 'Between Same-Sex Marriages and the Large Hadron Collider' (n 584) 591.

According to Petrenko and McArthur, it is necessary to find an account of when a harm scenario is adequately falsified.[595] They write that falsification "…implies that the predications of a proposed harm scenario fail to obtain in multiple studies."[596] If a general trend is evident in numerous studies, then the trend ought to be taken at face value.[597] Petrenko and McArthur go on to explore how numerous studies and safety reviews (i.e., the trend) adequately falsified the harm scenarios advocated by the LHC critics. According to the authors' analysis of the principle, the weight of the evidence produced by the LHC's risk assessment team and other studies ruled out the claimants' fears and eliminated the need for further precautionary measures.[598]

Whereas one can doubt the value of taking the trend at face value, a court might very well use analogous reasoning to dismiss the claim the defendants and CERN failed to adhere to the precautionary principle.

To summarise, the district court dismissed the complaint in September 2008 for lack of subject matter jurisdiction.[599] NEPA was not applicable because the defendants' involvement with the LHC did not constitute a major federal action. Furthermore, the precautionary principle and the cited agreements were not binding on the United States. The complaint was, therefore, never heard on the merits.

3.5.3 The Appeal

In October 2008, Wagner appealed the district court's decision, but the Ninth Circuit Court of Appeals ultimately dismissed his request in August 2010.[600]

This time around, lack of subject matter jurisdiction was not the basis for doing so. Instead, the Court held that Wagner could not demonstrate that he had standing to bring the claim.[601] The Court of Appeals cited *Lujan v Defenders of Wildlife* explaining

[595] ibid 605.
[596] ibid.
[597] ibid.
[598] ibid.
[599] *Sancho case* (n 87) 1268.
[600] *Sancho appeal* (n 460).
[601] ibid 611.

that the constitutional minimum of standing contains three elements.[602] To establish standing Wagner had to demonstrate: (1) an injury in fact, (2) a causal connection between the injury and the conduct complained of, and (3) a likelihood that a favourable decision will redress the injury.[603]

An injury in fact is an invasion of a legally protected interest which is (a) concrete and particularised and (b) actual or imminent, not conjectural or hypothetical.[604] In the LHC appeal, the Court of Appeal explained that injury in fact required some "credible threat of harm".[605] Wagner had only alleged "potential adverse consequences".[606] According to the Court, speculative fear can and did not constitute an injury in fact sufficient to confer standing.[607] The decision demonstrates a significant barrier to challenging similar disputes in the judicial system because:

> [t]he bulk of existential risk over the next century may…reside in rather speculative scenarios to which we cannot assign precise probabilities through any rigorous statistical or scientific method.[608]

Furthermore, even if Wagner had successfully demonstrated injury in fact, the Court maintained that he failed to satisfy the elements of *causality* or *redressability*.[609] As previously explored, the U.S. Government had no effective control over CERN's operations. The alleged harm was not attributable to the defendants' failure to comply with NEPA (i.e., there was no causal connection between the injury and the complained conduct). Furthermore, a decision in favour of the appellant would not have impacted CERN's ability to operate the LHC.[610] Since the injunction would not have afforded Wagner the relief he sought, a favourable decision would not redress the injury.

---

[602] See *Lujan v Defenders of Wildlife* (1992) 504 U.S. 555, 560-61.
[603] *Sancho appeal* (n 460) 611; *Lujan* (n 602) 560-61.
[604] ibid 560.
[605] *Sancho appeal* (n 460) 611.
[606] ibid.
[607] ibid 612.
[608] Bostrom, 'Existential Risk Prevention' (n 15) 16; Wilson, 'Minimizing' (n 151) 336.
[609] *Sancho appeal* (n 460) 612.
[610] ibid.

Some scholars have suggested that the requirements for standing in the US can undermine the interests of future generations in the judiciary. As written by Martinez and Winter:

> As part of the injury-in-fact requirement, plaintiffs must show that they have suffered or imminently will suffer actual injury (i.e., neither conjectural nor hypothetical; not abstract), whereas virtually any harm suffered by future generations will in the present be hypothetical or abstract (that is, neither actual nor imminent).[611]

This is problematic from the perspective of x-risk mitigation because the protection of humanity's future (both in terms of its potential and those who may inhabit the future) is at the forefront of x-risk reduction. While Martinez and Winter highlight that the requirements for standing can vary greatly in different jurisdictions, they note that *locus standi* has not been extended to future generations in any jurisdiction to date.[612]

3.5.4 Other cases

Suits to stop the LHC have reportedly also been initiated in Switzerland, Germany and the ECtHR.[613] There is no need to go over each dispute in detail.[614] Instead, the essential point is that, like with the *Sancho complaint*, it is reported that no court reached the merits of the claim.[615]

Wilson maintains that these disputes show that neither domestic nor international courts can handle disputes involving low-probability, high-consequences advanced technologies.[616] Peterson also suggest that the courts went to great lengths to avoid deciding the LHC Controversy on the merits.[617] Rather than being procedurally

[611] Martinez and Winter, 'Protecting Future Generations: A Global Survey of Legal Academics' (n 394) 13.
[612] ibid.
[613] Wilson, 'Minimizing' (n 151) 335; Adams, 'Honey' (n 309) 152-153; Rostam Neuwirth, 'Law and Magic: A(Nother) Paradox' [2014] Thomas Jefferson Law Review 139, 155-156.
[614] For an overview see Wilson, 'Minimizing' (n 151) 335-337.
[615] Peterson, 'The "Sancho" Effect' (n 468) 308; Johnson, 'The Black Hole Case' (n 68) 870; Neuwirth, 'Law and Magic' (n 613) 156.
[616] Wilson, 'Minimizing' (n 151) 337.
[617] Peterson, 'The "Sancho" Effect' (n 468) 308.

prevented from adjudicating the claims on the merits, he suggests they more likely felt incompetent to assess the relevant scientific theories.[618]

3.5.5 The Reverse Claim

The next Part will examine the 'review barriers' that may complicate different modes of external review in relation to the LHC Controversy. Before that, this Section briefly highlights an aspect that arguably sits in the background of the litigation to stop the LHC. That aspect concerns the *right to science*.

The claimants in the *Sancho case* challenged the defendants' liberty to pursue a research endeavour. As a side-effect, a successful claim would possibly have prevented derivative benefits flowing from the LHC project.

Boggio and Romano write that the 'right to science' includes the right of (i) scientists to research and, (ii) everyone to benefit from that research.[619] As demonstrates below, many legal systems recognise the right to science. In effect, research limitations and restrictions ought to require justification. Accepting certain limitation is less controversial when an experiment involves human subjects. As one example, the International Covenant on Civil and Political Rights states that "…no one shall be subjected without his free consent to medical or scientific experimentation."[620] Yet, not all experiments involve human subjects, and experimental particle physics experiments provide a point in hand.

Santosuosso, Sellaroli and Fabio highlight the constitutional protection of freedom of research in various countries. They demonstrate that some European constitutional systems expressly recognise and protect freedom of research.[621] Other systems,

---

[618] ibid.

[619] Andrea Boggio and Cesare Romano, 'Freedom of Research and the Right to Science, from Theory to Advocacy' in Simona Giordano (ed), *The Freedom of Scientific Research: Bridging the Gap between Science and Society* (Manchester University Press 2019) 162.

[620] Article 7.

[621] See Amedeo Santosuosso, Valentina Sellaroli and Elisabetta Fabio, 'What Constitutional Protection for Freedom of Scientific Research?' [2007] Journal of Medical Ethics 342, 342. Examples given by the authors include Article 5(3) of the German Constitution ('Art and scholarship, research, and teaching shall be free'); Article 33(1) of the Italian Constitution ('The arts and sciences shall be free, and free

including the US, do not expressly protect this freedom. However, in the US, freedom of research falls to be protected "…as a specific aspect of the wider freedom of thought and expression…"[622] This conceptual route of protection has long been recognised.[623] A similar route is likely possible in the UK. The Human Rights Act ('HRA') gives effect to rights and freedoms guaranteed under the ECHR.[624] The ECHR protects freedom of thought (Article 9) and freedom of expression (Article 10). According to the HRA, a court must "take into account" the decisions of the European Court of Human Rights ('ECtHR').[625] This thesis does not engage in a discussion about the relationship between Strasbourg and the UK courts. However, decisions of the ECtHR arguably support that freedom of research is implicated in the right of freedom of thought and right of freedom of expression.[626] The right to science is also recognised under international and regional law.[627] Examples (beyond the ECHR) include Article 27 of the Universal Declaration of Human Rights[628] and Article 15 of the International Covenant on Economic, Social and Cultural Rights.[629] Concerning the European Union, the Charter of Fundamental Rights of the European Union states: "The arts and scientific research shall be free of constraint."[630]

shall be their teaching') and Article 59 of the Slovenian Constitution ('The freedom of scientific and artistic endeavour shall be guaranteed'). Some of these go further and involve the governments in *promoting* and supporting research. Article 9 of the Italian Constitution is provided as an example.

[622] ibid.

[623] Consider the First Amendment to the US constitution. Ferguson explains the basic notion: "…the constitutional claim of scientific inquiry holds that the research enterprise of scientists has a first amendment importance because it is essential to the ability of individuals to engage in scientific expression. The argument thus proceeds on the assumption that scientific expression is itself protected by the free speech clause of the first amendment." James Ferguson, 'Scientific Inquiry and the First Amendment' [1979] Cornell Law Review 639, 644.

[624] Human Rights Act 1998.

[625] ibid s 2(1).

[626] See Voeneky, 'Human Rights and Legitimate Governance' (n 305) 154.

[627] ibid 154-155; Boggio and Romano, 'Freedom of Research and the Right to Science, from Theory to Advocacy' (n 619) 163-164.

[628] This Article stipulates: "(1) Everyone has the right freely to participate in the cultural life of the community, to enjoy the arts and to share in scientific advancement and its benefits. (2) Everyone has the right to the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author."

[629] Article 15(1)(b) stipulates that the Parties to the convention recognises the right of everyone "[t]o enjoy the benefits of scientific progress and its applications…" Article 15(3) reads: The States Parties to the present Covenant undertake to respect the freedom indispensable for scientific research and creative activity.

[630] Article 13

The upshot is this: Restricting the defendants in the context of the *Sancho case* may conflict with the right to science. Boggio and Romano highlight the uncertainty surrounding the content of the right to science.[631] They suggest that 'judicial mobilisation' can provide conceptual clarity to the right to science, and further that the judicial path entails bringing claims against states whenever the right to science is violated.[632] Judicial mobilisation, they write, could target laws and regulations of states if these prohibit or unreasonably restrict the freedom of scientific research.[633] Another potential target for judicial mobilisation is restrictions on the publication of scientific findings. A relevant article here is *The Limits of Government Regulation of Science*.[634] It deals with the controversy surrounding the U.S. government's request – directed at two journals – to redact key research findings before publication: The research was linked to the genetic modification of H5N1 avian influenza viruses.[635] As set out by Kraemer and Gostin, the request did not violate the First Amendment in the absence of legal force, undue inducements or penalties.[636] In contrast, trying to halt the publication or research teams by legal means or adverse consequences like loss of funding, could have violated the First Amendment.[637]

To summarise, the right to science is important in considering how the LHC Controversy should be navigated. In particular, it suggests that an absolute ban is not often appropriate. However, the right to science cannot be taken to give actors an unrestricted right to pursue potentially risky. In particular, the right to research should be restricted when there are reasonable doubts about the reliability of the scientific work assessing x-risk linked to a particular research endeavour.

[631] Boggio and Romano, 'Freedom of Research and the Right to Science, from Theory to Advocacy' (n 619) 165.
[632] ibid 168.
[633] ibid 169-170.
[634] John Kraemer and Lawrence Gostin, 'The Limits of Government Regulation of Science' [2012] 1047, 1047.
[635] ibid.
[636] ibid 1047-1048.
[637] ibid 1047.

3.6 THE REVIEW BARRIERS

This Part discerns a few review barriers that may problematise various modes of external review in relation to the LHC Controversy. To discern these obstacles, the thesis considers the aspects that made—or would have made—judicial intervention in the *Sancho case* difficult. It is helpful to repeat that the LHC-inspired intervention should respond to some of the barriers outlined below.

This Part categorises the review barriers under two sub-headings: *Jurisdiction Over Risk Originator* and *Core Obstacles*. The LHC-inspired intervention is only designed with the so-called core-obstacles in mind. The reason for this selective approach is explained in Section 3.6.2.

3.6.1 Jurisdiction Over Risk Originator

A pressing practical obstacle is that the external reviewer may not be able to exercise jurisdiction and control over the risk originator. As noted in Chapter 1, reference to an external reviewer means any independent third party tasked with considering the reliability of the x-risk assessment. For instance, in dealing with the identity of the external reviewer, Chapter 6 briefly considers international and domestic courts and existing and novel international agencies.

Chapter 2 noted that an x-catastrophe will affect every jurisdiction, and this can cause jurisdictional issues. As the creation of x-risk may ultimately become more localised, this obstacle may not be relevant in every situation. For example, if the originator is a national private company, personal jurisdiction will probably be less problematic than if the originator is an IO with immunity from legal process in the member states.

(i) CERN

CERN is the primary risk originator in the LHC context.[638] Whereas the district court did not address the issue of jurisdiction over CERN,[639] CERN's treaty-based immunities in its member states can be viewed as a hurdle to jurisdiction.

It is possible to consider the jurisdictional obstacle beyond the CERN's treaty-based immunities. For example, Johnson notes the issue of obtaining personal jurisdiction over CERN in the US, which is a non-member state. Nonetheless, he suggests that CERN left itself open to suits because it had established sufficient contacts with the US—owing to certain LHC operations being performed in Illinois—to provide personal jurisdiction.[640] Nevertheless, recall that the *Sancho* appeal (dismissed in 2010 after the publication of Johnson's article) support that a decision in favour of the claimants in the US would not have prevented CERN from operating the LHC in Europe.[641]

What about the position in the member states and host states? As a domestic legal person, CERN is exposable to judicial proceedings before the members' national courts. However, this position is subject to agreement(s) to the contrary. In 2004, CERN concluded a Protocol extending the privileges and immunities granted by Switzerland and France (the host states) by offering similar facilities in all member states.[642] Article 5 of the Protocol confirms that CERN enjoys immunity from legal process unless waived by the Council.[643] Immunity from legal process can be understood as an exemption from the adjudicative as well as enforcement procedures of national courts.[644] If a domestic court deems that the entity is immune from legal process, it is not amenable to suit.[645] A frequently used avoidance technique of national courts involves organisations' immunity from legal process.[646] A suit to stop

[638] The federal defendants alleged they had no control over the operation or management of the LHC.
[639] Peterson, 'The "Sancho" Effect' (n 468) 307.
[640] Johnson, 'The Black Hole Case' (n 68) 867-869.
[641] *Sancho appeal* (n 460) 611-12.
[642] CERN Protocol (n 501).
[643] ibid art 5(1).
[644] Reinisch, *International Organizations Before National Courts* (n 470) 16.
[645] ibid 127. On page 139, he continues by writing that the procedural effect of respecting immunity normally lies in the termination of legal proceedings against the organisation.
[646] ibid 127.

the LHC in Switzerland was reportedly dismissed based on CERN's treaty-based immunity in said state.[647]

(ii) Other Actors

The question of jurisdiction and immunity is not exclusive to IOs. It is beyond the scope of this thesis to deal with this issue in depth. However, it has been suggested with reference to the United States that the doctrine of sovereign immunity may complicate suits against federal agencies when such agencies allegedly expose the public to catastrophic risk through their activities.[648] Noteworthy, sovereign immunity was reportedly invoked as a defence against a suit to stop RHIC (the collider in the US).[649]

To summarise, it might not be possible to subject risk originators to effective review owing to jurisdictional issues and the inability to excerpt control over its undertakings. These kinds of practical issues underpinned the LHC Controversy. That being said, the thesis does not attempt to address these issues. Instead, the LHC-inspired mechanism aims to circumvent the review barriers outlined in the upcoming Section.

3.6.2 Core Obstacles

This Section highlights the review barriers relevant to the LHC-inspired intervention (the 'core obstacles'). It is useful to clarify why the core obstacles matter more than those covered above. As explained in Chapter 1, the thesis is less concerned with the practical implementation of the LHC-inspired intervention and more concerned with its theoretical building blocks, function, and logic. The focus on the core obstacles aligns with that focus.

(i) Unintelligible and Uncertain Science

The first obstacle that may complicate different modes of external review in relation to the LHC Controversy has to do with the complex subject matter underlying the dispute.

[647] Johnson, 'The Black Hole Case' (n 68) 860.
[648] Johnson, 'Agencies and Science-Experiment Risk' (n 68) 528-531 and 556.
[649] ibid 555-556.

As written by Posner, modern science is conceptually tricky, and even when it is not highly mathematised, it often employs mathematical and statistical procedures, intimidating vocabulary and other complex notions beyond the comprehension of many laypeople.[650] Naturally, judges, lawyers, legal scholars and policymakers are affected by such complexities. According to Posner, those within the legal profession might even be more vulnerable since many deliberately turned their back on science when they decided to pursue law.[651] He goes on to suggest that the disciplines are intrinsically different: Whereas the field of law stress verbal skills, science is heavily dependent on mathematics.[652]

The unintelligible science involved in the LHC Controversy can make it challenging for an external reviewer, such as a court, to deal with and independently review the scientific merits underlying the x-risk assessment. For example, in the *Sancho case*, the claimants filed several complex motions, declarations and affidavits to support their allegations, and the defendants responded in kind.[653] Neuwirth notes that it is possible to view the LHC disputes as signifying a fundamental conflict between law and science.[654] He proposes that the claimants attempted to use law as a means to validate competing scientific theories.[655] This request is problematic given the complexity of the subject matter of the dispute. It is even more problematic when the experiment itself is designed to expand humanity's current knowledge. As Johnson notes, the disputed theories can only be validated through knowledge that the experiment itself is designed to supply.[656]

Peterson also suggests that the *Sancho* case reveals that courts are baffled by arcane theoretical science.[657] He argues that the courts probably felt incompetent to deal with the theories that were involved in the dispute.[658] To bridge the gap, Peterson examines the potential to equip the courts with scientific skill and aptitude so they can

[650] Posner, *Catastrophe* (n 90) 96.
[651] ibid.
[652] ibid.
[653] Peterson, 'The "Sancho" Effect' (n 468) 306.
[654] Neuwirth, 'Law and Magic' (n 613) 155.
[655] ibid 156.
[656] See Johnson, 'Judicial Review' (n 68) 74.
[657] Peterson, 'The "Sancho" Effect' (n 468) 309.
[658] ibid 308.

independently examine the merits of conflicting scientific claims. His article advances the idea that "…courts must, to the extent possible, *become* the experts."[659]

A contrasting view is that courts should not function as validators of science. Theoretical and experimental particle physics is highly complex. It is one thing to read and compare the papers and evaluations. It is another thing to appreciate and evaluate them in a meaningful manner. Can an external reviewer, such as a judge, ever be expected to independently understand the essence of the subject matter in a situation such as the one involving the LHC? It is perhaps unrealistic to expect courts (and other external reviewers) to determine what constitutes the scientific truth.[660]

For example, in 1986, two scholars wrote that the disciplines involved in environmental regulation alone are so obscure that a judge cannot be expected to gain an understanding of these disciplines.[661] This general notion applies to physics and many other scientific disciplines that can be linked to x-risk today and probably in the future. Nevertheless, whereas a judge cannot necessarily become the expert, she can learn who the experts are, and encourage that the experts' views are considered by the relevant parties.[662] However, the Sub-section immediately below covers the possible shortcomings of this approach in the context of the LHC dispute.

(ii) Lack of Disinterested Experts

Another core obstacle interlinked with unintelligible science is that there is only a handful of people who can understand and evaluate the underlying theories in the LHC Controversy context.

---

[659] ibid 310.

[660] See Kenneth Abraham and Richard Merrill, 'Scientific Uncertainty in the Courts' [1986] Issues in Science and Technology 93, 98 and 106. The authors make the point that it is undesirable to attempt to make courts expert arbitrators of scientific uncertainty. For example, the idea of establishing a court for the resolution of scientific issues could potentially saddle the scientific community with unwanted features underpinning the legal system, like preoccupation with process, formalisation of dialogue and disputes.

[661] ibid 98.

[662] ibid.

An overarching problem with experts is that they are often to be found on either side of the case. While it is always possible to critically question someone's expert title, the point is that both supporters and critics often accompany any given claim. The experts' motivations can also differ. In the LHC disputes, the experts were, as written by Johnson, either afraid for their livelihoods or lives.[663]

According to Posner, physicists will have professional, pecuniary and emotional stakes in their projects and thus an incentive to downplay the risks.[664] Other authors, including Rees,[665] Bostrom and Ćirković,[666] Calogero,[667] Lehmann,[668] and Peterson[669] paint a similar picture. Given the potential for biased experts in the LHC Controversy, Wilson recommends that self-assessments of safety by those involved with a project should be subject to independent review.[670] Adams also notes that self-regulation is somewhat awkward since the scientists have a personal stake in seeing the experiment(s) through.[671]

In the *Sancho complaint*, the claimants were self-proclaimed experts in physics and other fields of science with an alleged capacity to question the safety reviews produced by other experts such as those working at CERN.[672] However, it has been noted by Peterson that the experts at CERN most likely viewed the claimants' theories as junk science.[673] But the obscurity of the underlying science becomes a double-edged

[663] Johnson, 'The Black Hole Case' (n 68) 873.
[664] Posner, *Catastrophe* (n 90). On page 99, he writes about scientists being reluctant to acknowledge that their research may do more harm than good. On page 133, he explains that the particle physicists have professional and pecuniary (or both) stakes in their projects and thus an incentive to downplay risks.
[665] Rees writes that some theorists seemingly strived to reassure the public as opposed to making an objective analysis. See Rees, *Our Final Century* (n 226) 127.
[666] Bostrom and Ćirković explain that physicists, while having the necessary expertise to assess the risks, are members of a professional community with a stake in the experiments going forward. Bostrom and Ćirković, 'Introduction' (n 12) 18-19.
[667] Calogero notes many experts in the LHC Controversy appeared to be overly concerned with public relations See Calogero, 'Might a Laboratory Experiment Destroy Planet Earth?' (n 446) 198-199.
[668] Lehmann, *No Canary in the Quanta* (n 94).
[669] Peterson writes: "[M]uch of the safety review was performed by CERN employees, creating a significant risk of bias in the reviews." See Peterson, 'The "Sancho" Effect' (n 468) 316.
[670] Wilson, 'Minimizing' (n 151) 337-338 and 360-361.
[671] Adams, 'Honey' (n 309) 155-157.
[672] *Sancho complaint* (n 460) 5.
[673] Peterson, 'The "Sancho" Effect' (n 468) 309.

sword. Whereas the unintelligible science can be used to downplay the risk, it can also be misused to produce unwarranted concerns.

If an external reviewer should refrain from arbitrating the scientific truth, the issue may instead become one of distinguishing good from bad expert testimony. In the context of the *Sancho case*, such a role could have been given to the court. One option for the court is to deal with expert testimony before it reaches the courtroom.[674] In *Daubert v. Merrell Dow Pharmaceuticals*, the U.S. Supreme Court articulated a standard for admitting scientific testimony.[675] The *Daubert* decision make judges gatekeepers for scientific evidence.[676] As explained by Schwartz, the core premise of *Daubert* is that judges decide whether relevant scientific testimony is based on the scientific method without taking a position vis-à-vis the truth of particular scientific conclusions.[677] *Daubert* ensured that a scientific theory would be empirically reliable, known to the scientific community as well as pass the test of review.[678] The Federal Rules of Evidence was amended following the *Daubert* judgement.[679] Section 702 instructs the expert to base her opinion on sufficient facts or data, utilise reliable principles and methods and reliably apply these to the facts of the case.[680]

Nevertheless, utilising a *Daubert* like standard in the LHC context is problematic. The LHC critics' and proponents' theories are only confirmable and falsifiable, respectively in the unhelpful event of Earth's obliteration.[681] The problem with the framework arguably goes beyond the work of the critics and proponents. For instance, Johnson suggests that particle physics has reached a level of 'theoreticalization' where it is divorced from the real-world concepts such as testability.[682] In short, the abovementioned legal framework clashes with the field of particle physics. Whereas it is, therefore, possible to question whether particle-physics testimony should be

[674] Johnson, 'The Black Hole Case' (n 68) 873.
[675] *Daubert v Merrell Dow Pharmaceuticals Inc* (1993) 509 U.S 579.
[676] Adina Schwartz, 'A "Dogma of Empiricism" Revisited: Daubert v. Merrell Dow Pharmaceuticals, Inc. and the Need to Resurrect the Philosophical Insight of Frye V. United States' [1997] Harvard Journal of Law & Technology 149, 156.
[677] ibid 157.
[678] See J Davies, 'Admissibility of Scientific Evidence in Courts' [2005] Medicine and Law 243, 248.
[679] ibid 248-249.
[680] ibid 249; Federal Rules of Evidence, s 702 (a)-(d).
[681] Johnson, 'The Black Hole Case' (n 68) 873-874.
[682] ibid 873-876.

allowed in the courtroom, excluding such testimony is equivalent to excluding judicial intervention[683] because the court cannot be expected to singlehandedly deal with the scientific merits of the LHC Controversy.

(iii) Objectivity and Conflict of Interest

Chapter 1 highlighted the following quote by renowned physicist John Ellis: "Humanity is safe from the LHC…but is the LHC safe from humanity?"[684] From the physicists' perspective, this question is likely more pressing than the scientific worry.[685] The statement does not necessarily imply that proponents of the LHC will maliciously downplay the theorised risks. However, the statement invites one to think seriously about the objectivity of the physicists as risk-assessors.[686] It also highlights the possibility that the experts' judgements might be unconsciously biased.[687]

A scientist is not more admirable than any other type of worker, as noted by Posner.[688] Most aspire to advance scientific knowledge, not protect society from science.[689] While this mentality is understandable, it further highlights the feature that particle physicists have professional and financial stakes in their projects and, therefore, some incentive to downplay the risks.[690] Besides, some scientists are possibly reluctant to accept that their research can do more harm than good.[691] On a related note, the CERN Convention stipulates that the organisation shall have no concern with work for military requirements. Instead, CERN provides for collaboration in nuclear research of a pure scientific character.[692] As argued by Posner, possible psychological factors, as well as financial and career incentives, make it undesirable to entrust scientific policy and regulation to the scientists themselves.[693] How, then, can the potential bias and lack

[683] ibid 877.
[684] Johnathan Ellis, 'The LHC is Safe' (n 9) [00:44:11].
[685] ibid [00:44:24].
[686] See Wilson, 'Minimizing' (n 151) 337-338.
[687] Bostrom and Ćirković, 'Introduction' (n 12) 19.
[688] Posner, *Catastrophe* (n 90) 98.
[689] ibid.
[690] ibid 133; Adams, 'Honey' (n 309) 156.
[691] Posner, *Catastrophe* (n 90) 99.
[692] CERN Convention (n 478) art II.
[693] Posner, *Catastrophe* (n 90) 99.

of objectivity complicate external review in the context of the LHC Controversy? Below is an example placing the judiciary in the position as the external reviewer.

Building on what has been said above, it is possible to start from the premise that courts are gatekeepers of scientific evidence and not arbitrators of its truth. Let's presume – momentarily disregarding Sub-section 3.6.2 (ii) – that judges can adequately sort out expert testimony before trial by looking to the scientific method as opposed to merit in the field of particle physics. Given that the court is not validating competing theories, the experts' testimony reaching court will naturally influence the outcome of the matter quite substantially. Nevertheless, if personal interests motivate *all* experts, the evidence will inevitably be biased. External intervention is hardly meaningful if the outcome is a derivative such evidence and evaluations.

If a group in a particular lab is planning a potentially dangerous experiment, then risk evaluations should, as noted by Calogero, be prepared by scientists with no interests in its performance.[694] The tricky point, however, is that this task is easier said than done because "…the more knowledgeable experts are often to be found precisely among those who are also most keen to see the experimental results in question."[695]

Calogero talks about the public relations issues in relation to particle accelerators in his paper. He recognises that peremptory statements are sometimes necessary to allay fears of the public.[696] He further notes that an alarmed public and unjustified fears can impede scientific progress.[697] The LHC case is probably the archetypal example in the eyes of some particle physicists. Notably, Calogero sensibly suggests that those in the scientific community must not address such problems and manage public opinion by avoiding candour and transparency.[698] He asserts that many scientists were, in his opinion, overtly concerned "…with the public relations impact of what they, or others, say and write…"[699]

[694] Calogero, 'Might a Laboratory Experiment Destroy Planet Earth?' (n 446) 198-199.
[695] ibid 199.
[696] ibid 201.
[697] ibid.
[698] ibid.
[699] ibid 198.

Despite agreeing with Calogero's argument, one should note that the particle physicists were in a uniquely precarious situation. For example, in talking about the dissemination of inaccurate and negative images of scientists as irresponsible and certain experiments as dangerous, Larsen notes that "…no branch of science is [perhaps] more susceptible to this negative press than particle physics, especially experiments involving particle accelerators."[700]

The PR problems may feed into the other aspects discussed in this Section. In other words, the problems may elucidate some of the things that the particle physicists have done to allay the fears of the public. For instance, it may explain the physicists' united approach when it comes to the safety of the LHC. It may also explain why they have felt it necessary to resort to qualitative (rather than quantitative) language when considering the safety of high-energy physics experiments.

(iv) Veiled Probability

Chapter 2 explored the low-probability, high-impact nature of many x-risks. In general, the low and sometimes unknown probability can complicate efforts to respond rationally to the risk scenarios.[701] According to Wilson, x-risks with extremely low probabilities are nearly impossible to challenge in the judicial system.[702] Similarly, Adams writes that claimants will find it problematic to demonstrate that there is a danger when relying on theoretical physics.[703] The *Sancho appeal* supports this conclusion given that the Court of Appeals made it clear that speculative fear does not constitute an 'injury in fact' enough to confer standing.[704]

Is the LHC dangerous? "No", concludes CERN with a single word.[705] Overall, physicists have avoided to publish any quantification of the odds of a black-hole disaster, and according to Johnson, they moved away from expressing risk in

[700] Larsen, *Particle Panic!* (n 265) viii.
[701] Posner, *Catastrophe* (n 90) 6.
[702] Wilson, 'Minimizing' (n 151) 336.
[703] Adams, 'Honey' (n 309) 153.
[704] *Sancho appeal* (n 460) 612.
[705] CERN, 'Is the Large Hadron Collider Dangerous?' <https://home.cern/resources/faqs/large-hadron-collider-dangerous> accessed 28 April 2023; See also the LSAG Report (n 463).

quantitative terms.[706] As an example, consider CERN's webpage *The Safety of the LHC*.[707] It contains reassuring quotes from highly influential scientists and academics, all of which decisively dismiss the risks. In general, the experts seem inclined to allay fears 'beyond reasonable doubt' as was Calogero's impression.[708] Ideally, scientists should prioritise quantitative conclusions (in the form of probabilistic estimates) over ways to manage public relations.[709] Of course, once such estimates are out there, "…the non-trivial question arises of what an 'acceptable' value for the probability of a disaster happening should be – namely a value small enough that the risk is considered worth taking."[710] By avoiding quantitative conclusion, the experts also avoid the thorny question of when the risk of planetary destruction (or the destruction of humanity's future) is unacceptable.[711] A cost-benefit analysis is, thus, circumvented.

What is relevant is that the particle physicists played these difficulties in their favour.[712] In one of his works, Johnson highlights that they frustrated "…the quantification of the benefits of the experiment…"[713] By doing so, the physicists contributed to a state of affairs where cost-benefit analysis was not feasible.[714] Posner also observes this issue in relation to the RHIC.[715] For example, he writes that "…the difficulty of valuing basic research plays into the hands of the particle-accelerator community."[716]

Johnson also records that the particle physicists used emotionally charged statements to justify the experiments rather than doing so by measuring the benefits.[717] He states: "A recurrent theme in the non-quantified argument for particle colliders is to make a special claim of importance for particle physics over other scientific fields."[718] The scientists also played the situation by "…refusing to use numerical values to discuss

[706] Johnson, 'Judicial Review' (n 68) 75-76.
[707] CERN, 'The Safety of the LHC' (n 127).
[708] Calogero, 'Might a Laboratory Experiment Destroy Planet Earth?' (n 446) 192.
[709] ibid 199-201.
[710] ibid.
[711] Johnson, 'Judicial Review' (n 68) 77.
[712] ibid 76-80.
[713] ibid 76.
[714] ibid 76-77.
[715] Posner, *Catastrophe* (n 90) 147-148.
[716] ibid 134.
[717] Johnson, 'Judicial Review' (n 68) 76.
[718] ibid 77.

the chance of disaster."[719] As above, this meant that they could frustrate cost-benefit analysis, albeit this time "…on the other end of the…formula…"[720]

The scientists have continued to use – as seen on CERN's website[721] – qualitative language to defeat arguments that the x-risk(s) were not worthwhile. It is possible to witness this general approach by looking at CERN's online publications, which deals with the safety of the LHC.[722] CERN also directs those interested to the LSAG Report, which concludes: "There is no basis for any concerns about the consequences of new particles or forms of matter that could possibly be produced by the LHC."[723]

Interestingly, and reminiscent of Johnson's observation, Ord highlights that people often prefer qualitative language when discussing x-risks. In other words, they often avoid putting numbers on the risks and resort to expressions such as *improbable* or *highly unlikely*.[724] While Ord makes his remarks in discussing existential risk more generally,[725] his observation certainly finds application to the LHC Controversy.

Furthermore, Johnson highlights how—in a chain of events involving, first the paper by Giddings and Mangano,[726] then the LSAG Report[727] and finally the SPC Report[728]—the probability of harm, as a matter of rhetoric, went from "…*insignificant* to *inconceivable* to *impossible*."[729] The removal of the quantitative aspect can hinder traditional risk analysis and, therefore, judicial intervention. It has been suggested that the traditional risk analysis is expert-centred, relies on statistical and mathematical

[719] Ibid.
[720] ibid.
[721] CERN, 'The Safety of the LHC' (n 127).
[722] ibid.
[723] LSAG Report (n 463).
[724] Ord, *Precipice* (n 16) 165.
[725] ibid 165-170.
[726] Steven Giddings and Michelangelo Mangano, 'Astrophysical Implications of Hypothetical Stable TeV-scale Black Holes' (2008) 78)(3) Physical Review D <http://dx.doi.org/10.1103/PhysRevD.78.035009> accessed 28 April 2023.
[727] LSAG Report (n 463).
[728] Peter Braun-Munzinger and others, 'SPC Report on LSAG Documents' (SPC Panel 2008) <http://cdsweb.cern.ch/record/1113558/files/cer-002766289.pdf> accessed 28 April 2023.
[729] Johnson, 'The Black Hole Case' (n 68) 852. For the discussion see pages 845-852.

calculations, and involves cost-benefit analysis.[730] The courts can use this traditional view to deal with problems of uncertainty in some cases.[731]

Finally, the dynamics of the debate outside the courtroom are likely reflective of the debate that would take place inside the courtroom.[732] The general approach of *not-talking-about-probability*[733] might contribute to a judicial standpoint that the alleged risks are merely speculative.

(v) Infinite Loss?

Johnson submits that it is difficult to quantify the benefits of high-energy physics experiments and assign a numerical value for the probability of catastrophe.[734] Posner's book, to mention another example, also shed light on these issues.[735]

Leiss writes that "[w]e calculate risk as the product of probability times consequences (R = P x C), but, as Johnson says…, 'The insertion of infinity into the risk equation causes it to blow up.'"[736] Since Leiss cites Johnson in talking about infinity, it is possible to get some insights by looking at Johnson's article.[737] Some background information is necessary before dealing with the relevant pages of Johnson's article. Johnson considers the use of cost-benefit analysis to tackle the LHC case and an accompanying injunction request.[738] Hypothetically, if the *benefits* (of the LHC) outweigh the *costs* (of the LHC), an injunction should not be issued and vice versa.

However, Johnson's overarching argument is that such analysis begins to "break down" when exposed to the LHC case. More as a thought experiment, then, Johnson crunches some numbers to analyse the situation and make his point. Johnson begins

---

[730] Molly Walker Wilson, 'Cultural Understandings of Risk and the Tyranny of the Experts' [2011] Oregon Law Review 113, 123.
[731] Johnson, 'Judicial Review' (n 68) 68-69.
[732] ibid 80.
[733] Johnathan Ellis, 'The LHC is Safe' (n 9) [01:05:22].
[734] Johnson, 'The Black Hole Case' (n 68) 876-880.
[735] Posner, *Catastrophe* (n 90).
[736] Leiss, *The Doom Loop in the Financial Sector, and Other Black Holes of Risk* (n 16) 17.
[737] Johnson, 'The Black Hole Case' (n 68) 876-878.
[738] ibid 876-883.

by calculating the *price of risk* to add it to the costs. He calculates *R* – the price of risk – like this:

> $R = PL$ (where, P is the probability and L is the loss).

Johnson initially assumes that the value for *P* is zero and that the value of *L* is infinite. Thus, $R = 0 \times \infty$. This means that $R = 0$. Thus, under the presumptions above, Johnson finds that the price of risk is zero. If this is true, "…the LHC is a no-lose proposition, and we would not even need to analyze the LHC's benefits to reach the conclusion that no injunction should be issued."[739] However, Johnson goes on to change the assumption for the value for *P*. He assigns an "…incredibly tiny number…" to *P* – namely, one in one trillion.[740] In effect:

> $R = (1 \times 10^{-12}) \times \infty$
>
> $R = \infty$

Under this assumption, the price of risk has "…shot up from nil to infinity."[741] This clarifies why Leiss suggests that the presumption of zero risk is necessary from the perspective of the proponents of the LHC.[742] In other words, he suggests that it is necessary to displace the very concept of risk when $L = \infty$. Impossibility means that there is no risk which can be deemed unacceptable considering the high stakes.

Nevertheless, it is possible to suggest that it is not a zero risk presumption that is necessary. Instead, if the aim is to calculate acceptability of risk, it is necessary to change the presumption that the loss is one of infinite magnitude. In fact, Johnson goes on to reconsider the assumption that the loss is infinite. This allows him to get some functional results for analysing whether the *benefits* outweigh the *costs*.[743]

---

[739] ibid 878.
[740] ibid.
[741] ibid.
[742] Leiss, *The Doom Loop in the Financial Sector, and Other Black Holes of Risk* (n 16) 17.
[743] Johnson, 'The Black Hole Case' (n 68) 878-883.

It is arguable that even an x-risk can be deemed too improbable to merit concern. This should also mean that a zero risk presumption is unnecessary even when the value of *L* is postulated as being 'infinite'. Thus, it is arguable that certain undertakings can be justified even if they pose an x-risk (without involving an existential trade-off). For example, in examining the RHIC controversy, Calogero writes:

> I would probably advise in favour of authorising a worthwhile undertaking, be it a scientific experiment or some other human enterprise, if I were *reliably* guaranteed that it entailed a risk of ultimate catastrophe per year less than one over one trillion (probability per year less than 10^{-12}).[744]

The concept of "ultimate catastrophe" in Calogero's paper appears to denote the destruction of Earth without taking into account the possibility that such catastrophe will destroy humanity's future. It is unclear whether he would have reached a different conclusion and figure if he had accounted for humanity's future (i.e., had he valued future lives and humanity's long term potential).

Additionally, it is debatable what amounts to a *worthwhile* undertaking and when a guarantee is reliable enough. Some might also disagree with Calogero's recommended figure (i.e., 10^{-12}). He recognises this in his paper.[745] However, he also emphasises why it might be unrealistic for someone to insist that a particular undertaking should pose zero probability of catastrophe before it can be allowed to go ahead.[746] Despite some grounds of disagreement, accounting for the value of humanity's future do not precludes the overarching argument above. Namely, it is possible to contend that some undertakings (such as the experiments at the LHC) should be allowed to go ahead even if there is a "non-zero probability" of x-catastrophe and even if the expected benefits are not in the league of an existential trade-off. Posner also suggests there are instances when "…[u]tterly trivial probabilities of even large harms must be ignored…"[747]

[744] Calogero, 'Might a Laboratory Experiment Destroy Planet Earth?' (n 446) 199.
[745] ibid.
[746] ibid.
[747] Posner, *Catastrophe* (n 90) 141.

The essential point is that the LHC Controversy, such as the unknown probabilities and the difficulty in assigning value to the future of humanity can hinder effective cost-benefit analysis or any intervention which emphasises the point at which a risk is acceptable and unacceptable.

Part 6.3 has introduced some of the review barriers that can affect different modes of external review in relation to the LHC Controversy. The LHC-inspired intervention aims to circumvent the so-called core obstacles highlighted above.

3.7 NO HARM AND NO RISK

This Part considers the idea that persons might have a 'right against risking' and whether this logic might work in the context of the *Sancho case*. This Chapter has already noted that the x-risks associated with experimental particle physics are uncertain.[748] In other words, it is far from clear whether the experiments pose any risk at all. The effect of this is that the 'right against risking' logic might not work in the LHC Controversy context. In effect, this Part provides a further example of the review barriers can disrupt a specific mode of external review.

3.7.1 No Risk of Future Harm

It has been argued that persons might enjoy a 'right against risking'.[749] In the context of human rights, Voeneky writes that "…the prohibition against actions that kill another person can be extended to actions that involve a risk that a person is killed, which means that a right not to be risk-exposed exists…"[750] Oberdiek explores, more broadly, whether people have a valid claim that others must not treat them in certain risky ways.[751] Oberdiek's answer is in the affirmative, and he builds this theory around the idea that people have an interest in leading an autonomous life. Autonomy requires a range of *acceptable options* on how to live one's life. Being subject to risk can narrow

[748] See Section 3.6.2 (iv).
[749] John Oberdiek, 'Towards a Right Against Risking' [2009] Law and Philosophy 367.
[750] Voeneky, 'Human Rights and Legitimate Governance' (n 305) 154.
[751] Oberdiek, 'Towards a Right Against Risking' (n 749) 367.

the acceptable options by narrowing one's safe options.[752] Yet, not all risks diminish autonomy, and Oberdiek identifies that some risks are too improbable to do so.[753]

Could the 'right against risking' logic have been of assistance in the *Sancho case* and in the LHC Controversy more generally? In 2010, the German Constitutional Court considered the argument that, unless the experiments at CERN were shut down, Germany would violate the duty to protect the claimant's life based on the constitutional right to life.[754] As translated and summarised by Voeneky:

> The court did not rule out this interpretation but held that reasonable doubt that the experiments are safe requires the plaintiff to at least try to rebut the majority view – that there is no existential risk – through arguments that are based on natural science and are part of the scientific discourse. Furthermore, the court held that it was necessary (and in the case at hand sufficient) that the German Government made a prior assessment of the risks that are posed by the experiments.[755]

Rebutting the majority view is not an easy task, and failing to do so will foreclosed the 'right against risking' argument. As mentioned, most experts agreed that experimental particle physics experiments pose no x-risk. The perceived absence of a credible risk would have foreclosed the 'right against risking' logic. Ultimately, the *Sancho appeal* largely ended the matter (in the US) by finding that speculative fear could not constitute an injury in fact sufficient to confer standing. It is noteworthy to recall that the experts have avoided probabilistic estimates. One might think back Calogero's point here that once such estimates are available, "…the non-trivial question arises of what an 'acceptable' value for the probability of a disaster happening should be..."[756] Circumventing probabilistic estimates also meant that the claimants in the *Sancho case* could not argue the risk of harm was unacceptable.[757]

---

[752] ibid 371-374.

[753] ibid 371-376. It is, therefore, uncertain whether the right against risking can extent to the *Sancho case*.

[754] Voeneky, 'Human Rights and Legitimate Governance' (n 305) 157-158.

[755] ibid 157-158. The thesis cites Voeneky's summary of the decision because the judgement is in German, and Voeneky is a professor at the University of Freiburg.

[756] Calogero, 'Might a Laboratory Experiment Destroy Planet Earth?' (n 446) 199-201.

[757] It is possible to transform the issue into risk-benefit assessment. However, Voeneky notes that the questions which must be answered to undertake such an assessment are tricky. Even if the risk is quantifiable, we might not be able to quantify the benefit. See Voeneky, 'Human Rights and Legitimate Governance' (n 305) 152.

Although criminal law is of little relevance to the *Sancho case*, and the LHC Controversy one might highlight the somewhat analogous notion of *endangerment*.[758] As explained by Duff, endangerment involves creating, by an act or omission, a significant risk that someone will suffer harm.[759] However, and again, the uncertain nature of the risk seemingly defeats this line of thinking.

(i) Analogy: Climate Litigation

This Sub-section briefly compares the phenomenon of climate litigation and the *Sancho case*. To keep the discussion short and to the point of comparing the position of the claimants in the *Sancho case* and claimants in relation to climate litigation, this Sub-section only refers to the *Urgenda* case.[760]

In the *Urgenda* case, the Urgenda Foundation sued the Dutch Government seeking to compel it to do more to prevent climate change. Of central importance were Articles 2 ('right to life') and 8 ('right to respect for private and family life') of the ECHR. Setzer and Byrnes report that a continuing trend in climate litigation is the reliance on human rights arguments.[761] According to the Court of Appeal in the *Urgenda* case, the Government had a positive obligation (using the phrase 'duty of care') to prevent a *future violation* of the interests protected under Articles 2 and 8. It reportedly said: "If the government knows that there is a real and imminent threat, the State must take

[758] See R.A Duff, 'Criminalizing Endangerment' [2005] Louisiana Law Review 941; Findlay Stark, *Culpable Carelessness: Recklessness and Negligence in the Criminal Law* (Cambridge University Press 2016) ch 1. Stark writes that Anglo-American criminal law concentrates on *unjustified* risk taking.
[759] Duff, 'Criminalizing Endangerment' (n 758) 944-945 and 952-953. When talking about endangerment, the defendant does not display hostility towards his victim. Instead, by taking the risk and failing to take adequate precautions, the defendant reveals that he does not care for the victim's interests Whereas endangerment need only involve an *actus reus*, Duff suggests that recklessness is the appropriate fault paradigm *if* there should be a fault element for endangerment.
[760] *State of the Netherlands v Urgenda Foundation*, ECLI:NL:HR:2019:2007, Judgment (Sup. Ct. Neth. Dec. 20, 2019) (Neth.).
[761] Joana Setzer and Rebecca Byrnes, 'Global Trends in Climate Change Litigation: 2020 Snapshot' (2020) <https://www.lse.ac.uk/granthaminstitute/wp-content/uploads/2020/07/Global-trends-in-climate-change-litigation_2020-snapshot.pdf> accessed 28 April 2023, 14.

precautionary measures to prevent infringement…"[762] The Court concluded that the State contravened its duty under the Articles by doing too little to prevent dangerous climate change.[763]

The dispute ultimately reached the highest court within the common judiciary in the Netherlands. The Supreme Court upheld the decision that the Government was under an obligation to reduce the emission of greenhouse gases pursuant to Articles 2 and 8.[764] The State did not dispute that there was "…a real threat of dangerous climate change, resulting in the serious risk that the current generation of citizens will be confronted with loss of life and/or a disruption of family life."[765] However, the State did dispute that it was obligated under Articles 2 and 8 to mitigate that threat.[766] Referring to jurisprudence from the European Court of Human Rights, the Supreme Court made several noteworthy remarks in disagreeing with the State's position.

Firstly, the positive obligation to take appropriate steps to safeguard the lives of individuals applies to situations entailing hazardous industrial activities and natural disasters. The obligation applies when there is a *real and immediate ris*k to persons and when the State is aware of that risk. The Court understood the phrase 'real and immediate' as referring to a risk that is genuine and imminent. The term 'imminent' meant that the risk must directly threaten the persons involved.[767] According to the Court, the positive obligation in Articles 2 and 8 overlaps vis-à-vis activities that are hazardous to the environment.[768] This finding allowed the Court to refer to these obligations collectively.

---

[762] Jonathan Verschuuren, '*The State of the Netherlands v Urgenda Foundation*: The Hague Court of Appeal Upholds Judgment Requiring the Netherlands to Further Reduce its Greenhouse gas Emissions' [2019] Review of European, Comparative & International Environmental Law 94.
[763] ibid.
[764] Translation of the Supreme Court's Decision: <http://blogs2.law.columbia.edu/climate-change-litigation/wp-content/uploads/sites/16/non-us-case-documents/2020/20200113_2015-HAZA-C0900456689_judgment.pdf> accessed 28 April 2023.
[764] *Urgenda Foundation* (n 760) 4.8 and 5.1.
[765] ibid 4.7 and 4.8 referring to the Court of Appeal's conclusion.
[766] ibid 4.8 and 5.1.
[767] ibid 5.2.2.
[768] ibid 5.2.4.

Secondly, the protection afforded by the Articles is not limited to specific persons, but to the population as a whole.[769] The Court regarded it as inconsequential that the risk would crystallise in the future and impact large parts of the population as opposed to specific persons.[770]

Thirdly, the obligation to take appropriate steps includes a duty to take *preventative measures*. According to the Court, the obligation to take appropriate steps may entail mitigation measures – that is, measures designed to prevent the threat from ever materialising.[771] It said: "The mere existence of a sufficiently genuine possibility that [the] risk will materialise means that suitable measures must be taken."[772] Considering Article 13 of the ECHR, the Court further explained that a remedy is *effective* if it prevents or ends the violation or offers redress for a crystallised violation.[773] When dealing with more serious violations, the available remedy must provide for both the prevention or end and redress.[774]

The *Urgenda* case demonstrates that a State's failure to curtail a risk can raise human rights questions. Following this general logic, can one argue that the State is breaching its 'duty' under Articles 2 and 8 by doing too little to mitigate x-risk associated with a scientific undertaking such as the LHC? The answer apparently depends on how far future courts are willing to stretch the positive obligations, and at what stage a risk becomes *real and immediate*.

An obvious issue is that climate litigation concerns the adverse consequences of dangerous climate change. The risk, at least in the *Urgenda* case, is not in dispute. It is sufficiently real. Contrast this with the *Sancho case* where the risk itself is clouded with uncertainty. Most experts reject the risk of future harm. The brief overview of the *Urgenda* case once again illustrates that the uncertainty of the underlying risk in the *Sancho case* makes it difficult to focus on the notion of the acceptability of the alleged

[769] ibid 5.3.1.
[770] ibid 5.6.2.
[771] ibid 5.3.2.
[772] ibid 5.6.2.
[773] Article 13 states: "Everyone whose rights and freedoms as set forth in this Convention are violated shall have an effective remedy before a national authority…"
[774] *Urgenda Foundation* (n 760) 5.5.3.

risk itself. Indeed, had the risk associated with climate change been highly disputed, then the *Urgenda* case might have been decided differently.

3.8 CONCLUSION

This Chapter has explored some legal aspects of the LHC Controversy. While some claim that the LHC poses an x-risk, others dismiss experimental particle physics as a serious hazard within the x-risk landscape. In particular, this Chapter has considered the *Sancho case* and highlighted some "core obstacles" that, in combination, may affect and hinder different modes of external review in relation to the LHC Controversy. The joint effect of these core obstacles is to make external and independent review problematic. It should be recalled that one of the self-imposed conditions for our broad legal intervention is that it should circumvent the so-called 'core obstacles'. This theme is picked up in Chapter 6, which will deal with the theoretical configuration of the LHC-inspired intervention in much greater detail.

4 CHAPTER FOUR: THE ULTIMATE INJUNCTION

> A person operating internationally cannot so easily defeat the judicial process. There is not a black hole into which a defendant can escape out of sight and become unreachable.[775]

4.1 BACKGROUND

Using the National Environmental Policy Act ('NEPA') and Precautionary Principle, the claimants in the *Sancho case* sought to obtain an injunction to halt the LHC. They argued that the LHC could create a black hole, essentially triggering an x-catastrophe. The District Court dismissed the complaint for want of subject matter jurisdiction.[776] The Court of Appeal, taking a different approach, held that they lacked standing.[777]

As written by Johnson, "…in American courts, the question remains: What cause of action could be used to force CERN to defend a suit on the merits?"[778] Similarly, according to Adams, it is difficult to imagine a body of law that is applicable where scientists aim to undertake an experiment that may enhance knowledge, but that may also give rise to a minimal and theoretical, yet non-zero chance of destroying the world.[779] In such a situation, much like in the *Sancho case* itself, the defendants will (i) not likely have caused any harm,[780] (ii) not have done anything illegal,[781] and (iii) positively believe, relying on scientific assessments, that the endeavour is safe. The

---

[775] *Mercedes-Benz AG v Leiduck* [1996] AC 284, 305 (Lord Nicholls). This was said in the context of jurisdictional issues. It does, in no way, relate to x-risks, particle accelerators or black holes.
[776] *Sancho case* (n 87) 1268.
[777] *Sancho appeal* (n 460) 611.
[778] Johnson, 'The Black Hole Case' (n 68) 861.
[779] Adams, 'Honey' (n 309) 148.
[780] The ex ante nature of the *Sancho* claim may exclude significant areas of law. The tort law system is a noteworthy example. The absence of any actualised (e.g., physical) damage to the claimant is fatal for certain damage-based torts in English law. See Jenny Steele, *Tort Law – Text, Cases and Materials* (3rd edn, Oxford University Press 2014) 112; Furthermore, Wansley explains that tort law is generally ineffective at reducing certain type of environmental and health risks (including risks to future generations) due to problems of causation, standing and evidence. See Matthew Wansley, 'Regulation of Emerging Risks' [2016] Vanderbilt Law Review 401, 423.
[781] Overall, it is difficult to analyse the *Sancho case* in terms of *actus reus* and *mens rea*. There was an absence of actualised harm, criminal conduct and malice intent or fault in the *Sancho case*. In contrast, those who intentionally mean to cause harm can more easily be condemned than those who intend to solve mysteries about the universe. If someone aspires to cause harm, society can often stop him before he succeeds. In England, for instance, the Criminal Attempts Act 1981 applies where a person intends to commit an offence and does an act which is more than merely preparatory to its commission.

uncertain nature of the risk itself is also problematic and will exclude anything akin to a ‘right against risking’ argument.

Despite the difficulties in finding an applicable body of law, the appropriate remedy is more readily identifiable. In *Sancho*, the claimants sought injunctive relief. In other words, they sought to halt—or at least delay—the defendants’ scientific endeavour from commencing. Ultimately, if an x-risk materialises, no one is left to deal with liability issues. As such, some ex ante measure, such as obtaining an injunction, was the only sensible option from the claimants’ perspective in the *Sancho case*.

4.1.1 Aims of Chapter 4

This Chapter considers Johnson’s *meta-analysis*. Johnson suggests that his analytical framework could have enabled a generalist judge in the US to deal with the preliminary injunction request in the *Sancho case*. Notably, Johnson does not situate his discussion within the wider remits of x-risk mitigation. Likewise, his approach has largely gone unnoticed within the x-risk circle.

Chapter 1 explained that the LHC-inspired intervention can operate to evaluate the reliability of an x-risk assessment by inviting an external reviewer to analyse a set of interconnected deficiency factors. These deficiency factors require consideration of (i) the actors who assessed the relevant x-risk and (ii) the possibility that they based their conclusion on flawed information or an imperfect state of knowledge or understanding. Chapter 6 will use Johnson’s meta-analysis to compartmentalise these factors. As such, Chapter 4 is vital since it allows the thesis to build upon Johnson’s meta-analysis. Beyond outlining the meta-analysis, Chapter 4 will contextualise injunctions given that this is the context within which Johnson situates the meta-analysis. It should also be recalled that the claimants in the *Sancho case* sought injunctive relief. Therefore, an overview of injunctive relief is interesting because it further contextualises the *Sancho case*.

4.2 INJUNCTIVE RELIEF: ENGLISH LAW

The *Sancho case* claimants requested a preliminary and perpetual injunction to halt the LHC until the defendants could meet certain conditions.[782] As a thought experiment and to introduce injunctions in general terms, this Part will consider whether the *Sancho case* claimants could have obtained a preliminary injunction had the dispute arisen in England. Part 4.3 will then consider injunctions from the American law perspective before Part 4.4 outlines the 'meta-analysis' that Johnson promotes within the context of a preliminary injunction request.

4.2.1 Brief Overview

The injunction is an equitable remedy underpinned by equitable principles.[783] In England, the Senior Courts Act 1981 stipulates that the High Court may grant an injunction "…in all cases in which it appears…just and convenient to do so."[784] The jurisdiction is seemingly wide, and according to Lord Nicholls, the criterion for the grant of relief is the prevention of *injustice*.[785] Unsurprisingly, the courts have restricted their discretion and there are various guidelines which dictate whether they are likely to grant injunctive relief.[786]

An injunction is a court order which requires someone do a specific act (*mandatory* injunction) or refrain from doing so (*prohibitory* injunction).[787] The remedy is available in a host of different circumstances.[788]

---

[782] *Sancho complaint* (n 460) 11.

[783] For instance, injunctive relief is not available where common law remedies (such as damages) can achieve justice between the parties. Alastair Hudson, *Equity and Trusts* (9th edn, Routledge 2016) 1034-1035.

[784] Section 37(1); The County Courts Act 1984, s 38(1) gives the County Court jurisdiction to grant an injunction: "…the court may make any order which could be made by the High Court if the proceedings were in the High Court."

[785] *Mercedes-Benz AG v Leiduck* (n 775) 308 (Lord Nicholls).

[786] Robert Pearce and Warren Barr*, Pearce & Stevens' Trusts and Equitable Obligations* (7th edn OUP 2018) ch 35.

[787] David Bean, Andrew Burns and Isabel Parry, *Injunctions* (13th edn, Sweet & Maxwell 2018) para 1-01.

[788] The injunction can, for instance, be used to compel the removal and demolition of structures, a partner to leave the matrimonial home. It can also be granted to restrain trespass to land, interference with contractual rights and trustees from committing a breach of trust. ibid 1-14 –1-28.

One can classify an injunction according to the period for which it remains in force: A *perpetual* injunction is a final remedy granted after a trial on the merits.[789] In contrast, an *interim* injunction is granted before the case goes to trial.[790] Different criteria will, therefore, govern these two types.[791] The injunction, once granted, remains in force until the court discharges it,[792] and breach thereof may constitute a contempt of court.[793] Indeed, without sanctions, "…an injunction would be a paper tiger."[794]

For the claimants in the *Sancho case*, the desirability to secure an injunction is self-evidential. Fearing that the defendants' endeavour would cause harm for which no compensation is possible, they sought to prevent rather than address the harm.

Unlike a tort or breach of contract, an injunction is not a cause of action. It is a *remedy*.[795] As such, there is an overriding requirement that the applicant have an underlying cause of action entitling her to relief.[796] For instance, in *Day v Brownrigg*, James LJ said: "This Court can only interfere where there is an invasion of a legal or equitable right."[797] Thesiger LJ also explained that no legal right had been invaded in the case.[798] In *Paton v British Pregnancy Advisory Service Trustees*, the court stated: "[T]here must be a legal right enforceable in law or in equity before the applicant can obtain an injunction from the court to restrain an infringement of that right."[799] Decisions from the House of Lords also enshrine that an injunction is not a cause of action.[800] It would seemingly have been difficult for the claimants in the *Sancho case*

[789] ibid 1-01.
[790] ibid 1-01.
[791] ibid.
[792] ibid 1-02.
[793] Pearce and Barr*, Pearce & Stevens' Trusts and Equitable Obligations* (n 786) 832.
[794] *Attorney General v Punch Ltd [2003] 1 AC 1046* [32] (Lord Nicholls).
[795] Bean, Burns and Parry, *Injunctions* (n 787) 1-03.
[796] ibid.
[797] *Day v Brownrigg* (1878) 10 Ch D 294, 305 (Lord Justice James) The defendant had changed the name of his house to *Ashford Lodge*, which was also the name of the claimant's home. The claimant asked for an injunction to restrain the defendant from continuing to use the name *Ashford Lodge.* The Court refused.
[798] ibid 305.
[799] *Paton v British Pregnancy Advisory Service Trustees* [1979] QB 276, 278. A husband unsuccessfully sought an injunction to prevent his wife's abortion.
[800] "A right to obtain an interlocutory injunction is not a cause of action. It cannot stand on its own. It is dependent upon there being a pre-existing cause of action against the defendant arising out of an

to obtain an injunction for the purpose of English law. This is because it is problematic to see how the defendants had invaded or threatened to invade the claimants' legal or equitable rights.

4.2.2 Perpetual Injunctions

There is an issue with *perpetual* injunctions in cases such as the *Sancho case*. A perpetual injunction is granted at the end of a final hearing.[801] The remedy may be moot if the defendants go ahead with the experiment before the final hearing. This means that the claimant may, as a matter of urgency, need to seek an interim remedy.

4.2.3 Interim Injunctions

Bean, Burns and Parry highlight the significance of interim injunctions. An interim hearing is not a trial on the merits, and neither side will prove their case at this stage.[802] Referring to *United States of America v Abacha*,[803] they explain that the purpose is to "hold the ring" pending final determination of the dispute.[804] While recognising that it is impossible to stop the world pending trial, Lord Hoffmann explains that "[i]t is often said that the purpose of an [interim] injunction is to preserve the status quo…"[805]

The question below is whether a court would potentially have granted an interim injunction in the *Sancho case* for the purpose of the English remedial landscape. While the courts may grant an interim injunction when it appears just and convenient to do so,[806] they have defined this discretion. When dealing with interim (*prohibitory*) injunctions, it is common to refer to *American Cyanamid v Ethicon*.[807] This case, and the guidelines by Lord Diplock, is taken to have clarified the court's approach and

---

invasion, actual or threatened by him, of a legal or equitable right of the plaintiff for the enforcement of which the defendant is amenable to the jurisdiction of the court." *The Siskina* [1979] AC 210, 256 (Lord Diplock); See also *Channel Tunnel Group Ltd v Balfour Beatty Construction Ltd* [1993] AC 334, 362 (Lord Mustill).

[801] Bean, Burns and Parry, *Injunctions* (n 787) 1-01 and 2-01.

[802] ibid 3-01.

[803] [2015] 1 WLR 1917.

[804] Bean, Burns and Parry, *Injunctions* (n 787) 3-01.

[805] *National Commercial Bank Jamaica Ltd v Olint Corp Ltd* [2009] 1 WLR 1405 [16].

[806] The Senior Courts Act 1981, s 37(1).

[807] *American Cyanamid Co v Ethicon Ltd* [1975] AC 396.

jurisdiction to grant interim injunctions.[808] While the case does not lay down exact and universal principles,[809] one can certainly speak of the *Cyanamid* guidelines as a leading source of law.

Scholars often compartmentalise the guidelines. For instance, Bean, Burns and Parry discuss the guidelines under the following headings: (a) a serious question to be tried, (b) inadequacy of damages to either side, (c) balance of convenience, and (d) special cases.[810] Similarly, in *R-Squared Holdco Limited*, the Court listed the principles deriving from *American Cyanamid* as follows: "…an applicant must show that: (1) there is a serious issue to be tried; (2) damages would be an inadequate remedy; and (3) the balance of convenience favours the grant of interim injunctive relief."[811]

(i) Serious Question

The starting point, sometimes described as the gateway to an interim injunction, is whether there is a serious question to be tried.[812] The claimant is not required to convince the court that she is more likely than not to obtain a final injunction at trial.[813] As explained by Lord Diplock, the court's function at the interim stage is not to resolve conflicts of evidence or difficult questions of law.[814] *Re R-Squared Holdco* echoes this point: When dealing with an application for an interim injunction, *"…*the court would need to be satisfied *only* that there was a serious issue to be tried…"[815] In contrast*, during the final determination, the court* "…would need to be persuaded to the full standard applicable in the final determination of civil matters, namely, on the balance of probabilities." [816]

---

[808] Guidelines for the exercise of the court's jurisdiction to grant interim injunctions were laid down by your Lordships' House in *American Cyanamid…*in the speech of Lord Diplock in that case, with which the remainder of their Lordships concurred." See *R v Secretary of State for Transport Ex p Factortame Ltd* (No.2) [1991] 1 AC 603, 671 (Lord Goff).
[809] ibid; Bean, Burns and Parry, *Injunctions* (n 787) ch 35.
[810] ibid 3-12.
[811] *R-Squared Holdco Limited* [2020] EWHC 23 (Ch) [37] (Mr James Pickering).
[812] *Allfiled UK Ltd v Eltis* [2015] EWHC 1300 (Ch) [67].
[813] Bean, Burns and Parry, *Injunctions* (n 787) 3-13.
[814] *American Cyanamid* (n 807) 407 (Lord Diplock).
[815] *R-Squared Holdco* (n 811) [22] (Mr James Pickering).
[816] ibid.

There are good reasons for the thresholds. In *American Cyanamid*, Lord Diplock said that "…the evidence available to the court at the hearing of the application for an [interim] injunction is incomplete. It is given on affidavit and has not been tested by oral cross-examination."[817] However, while the threshold is lower during the interim hearing, it is still necessary to satisfy the court "…that the claim is not frivolous or vexatious, in other words, that there is a serious question to be tried."[818] The threshold is not too high. Lord Diplock's judgement also reads:

> [U]nless the material available to the court at the hearing of the application for an [interim] injunction fails to disclose that the plaintiff has any real prospect of succeeding in his claim for a permanent injunction at the trial, the court should go on to consider whether the balance of convenience lies in favour of granting or refusing the [interim] relief that is sought.[819]

Accordingly, there is a serious question to be tried unless the material fails to reveal that the applicant has any real prospect of succeeding at trial.[820]

This gateway requirement would possibly have barred the claimants in the *Sancho case* from having obtained an interim injunction. As mentioned above, the injunction is not a cause of action. It is a remedy that is incidental to the enforcement of a substantive right.[821] How could the claimants have convinced the court that there was a serious issue to be tried? More specifically, *what is* the serious issue to be tried? Does it concern the materialisation of the alleged risk or the failure to comply with some procedural requirement such as with NEPA in the *Sancho case*?

(ii) Inadequacy of Damages / Balance of Convenience

If there is a serious issue to be tried, the court will examine if damages would be an adequate remedy for the party injured by the *grant* or *refusal* of the injunction. Withholding the injunction may adversely affect the claimant. Contrariwise, granting it can adversely affect the defendant. In *National Commercial Bank Jamaica,* Lord

[817] *American Cyanamid* (n 807) 406 (Lord Diplock).
[818] ibid 407.
[819] ibid 408.
[820] *Allfiled UK* (n 812) [67].
[821] *The Siskina* (n 800) 256 (Lord Diplock); *Channel Tunnel Group* (n 800) 362 (Lord Mustill).

Hoffmann explained that "…restrictions on the defendant's freedom of action will have consequences, for him and for others, which a court has to take into account."[822] At this stage, it helps to understand that a successful applicant must generally give a cross-undertaking in damages. That is, the claimant undertakes to compensate the defendant for the loss he sustains if the injunction cannot be justified at trial.[823] With the above in mind, consider the following statement by Lord Hoffmann:

> As the House of Lords pointed out in *American Cyanamid Co v Ethicon Ltd*…if damages will be an adequate remedy for the plaintiff, there are no grounds for interference with the defendant's freedom of action by the grant of an injunction. Likewise, if there is a serious issue to be tried and the plaintiff could be prejudiced by the acts or omissions of the defendant pending trial and the cross-undertaking in damages would provide the defendant with an adequate remedy if it turns out that his freedom of action should not have been restrained, then an injunction should ordinarily be granted.[824]

The alleged consequence of withholding the injunction in the *Sancho case* is the potential destruction of humanity's entire future. Even if one disregard the value of the future, the entire present would be destroyed. Clearly, damages would not adequately compensate the claimants. In contrast, there is more scope to argue that the cross-undertaking in damages could provide the defendant with an adequate remedy.

It is not always clear if damages will be an adequate remedy. In such an instance, the court may need to consider the balance of convenience more widely. In *American Cyanamid*, Lord Diplock explained "[i]t is where there is doubt as to the adequacy of the respective remedies in damages available to either party or to both, that the question of balance of convenience arises."[825]

In *National Commercial Bank Jamaica Ltd*, Lord Hoffmann recognised that it is often difficult to assess the adequacy of damages in practice. In effect, the court will try predicting whether the grant or refusal of the interim measure is "…more or less likely to cause irremediable prejudice (and to what extent) if it turns out that the injunction

---

[822] *National Commercial Bank Jamaica* (n 805) [16] (Lord Hoffmann).
[823] Bean, Burns and Parry, *Injunctions* (n 787) 3-03.
[824] *National Commercial Bank Jamaica* (n 805) [16] (Lord Hoffmann).
[825] *American Cyanamid* (n 807) 408 (per Lord Diplock).

should not have been granted or withheld…"[826] As a basic principle, he suggested, the court should take the approach which is likely to cause the least irremediable prejudice.[827] The balance of convenience in the context of the *Sancho case* arguably tips clearly in favour of the claimants.

Overall, the main problem for the claimants in the *Sancho case* is the *gateway* to any interim injunction. It is questionable whether the claimants could have shown that there was a serious question to be tried had the *Sancho case* arisen in England.

4.3 INJUNCTIVE RELIEF: THE AMERICAN PERSPECTIVE

This Part examines *preliminary injunctions* under American law in the context of the *Sancho case*. As in the English legal system, the court can grant relief before final judgement in the form of an interim injunction. The court should resort to established factors as opposed to unfettered discretion in deciding whether to grant preliminary injunctive relief. In *Chaplaincy of Full Gospel Churches v England*, the Court of Appeals noted:

> To warrant preliminary injunctive relief, the moving party must show (1) a substantial likelihood of success on the merits, (2) that it would suffer irreparable injury if the injunction were not granted, (3) that an injunction would not substantially injure other interested parties, and (4) that the public interest would be furthered by the injunction. [828]

In *Winter v Natural Res. Def. Council, Inc.*,[829] the Supreme Court of the United States articulated this standard in the following way:

> A plaintiff seeking a preliminary injunction must establish that he is likely to succeed on the merits, that he is likely to suffer irreparable harm in the absence of preliminary relief, that the balance of equities tips in his favor, and that an injunction is in the public interest.

[826] *National Commercial Bank Jamaica* (n 805) [17] (Lord Hoffmann).
[827] ibid.
[828] *Chaplaincy of Full Gospel Churches v England*, 454 F.3d 290, 297 (D.C. Cir. 2006)
[829] *Winter v Natural Res. Def. Council, Inc* 555 U.S. 7, 20 (2008).

Moore notes that a unified standard to preliminary injunctions emerged only after the decision in *Winter.* He also writes that the Supreme Court has not provided further guidance on how to apply the four-factor standard, meaning that lower courts must rely on *Winter* to evaluate requests for preliminary injunctive relief. Nevertheless, his article goes on to show that lower courts continue to disagree vis-à-vis the correct application of the standard.[830]

The four factors outlined above are co-dependent and affect each other. In *Brady Campaign to Prevent Gun Violence v Salazar*, the Court underlined that "[i]n applying this four-factored standard, district courts may employ a sliding scale under which a particularly strong showing in one area can compensate for weakness in another."[831] How does the factual matrix in the *Sancho case* align with the framework above?

According to Peterson, factors (2) and (4) support the request for injunctive relief: The destruction of the Earth is clearly within the notion of irreparable injury, and the injunction furthers the public interest by ensuring the survival of humanity.[832] In isolation, these factors tip the scale in favour of the claimant. However, factors (1) and (3) count against the request: It is exceptionally unlikely (improbable) that the alleged risks will materialise. In other words, Peterson infers that the claimant would fail to show a substantial likelihood of success on the merits because the likelihood of the disaster scenarios—i.e., the production of (a) Strangelets, (b) Micro Black Holes, and (c) Magnetic Monopoles—was far from substantial.[833] Furthermore, the injunction would substantially injure the other party, both financially and in taking away its opportunity to pursue and advance knowledge of the universe.[834]

"[W]ith two factors at either extreme…" writes Peterson, the scale is evenly balanced, and the resulting puzzle deprives the court of guidance on whether it could grant a preliminary injunction.[835] He, therefore, writes that in cases involving low-probability

[830] Devon Moore, 'The Preliminary Injunction Standard: Understanding the Public Interest Factor' [2019] Michigan Law Review 939. 943-944.
[831] *Brady Campaign to Prevent Gun Violence v Salazar* 612 F. Supp. 2d 1 (D.D.C. 2009).
[832] Peterson, 'The "Sancho" Effect' (n 468) 310-311.
[833] ibid 310.
[834] ibid.
[835] ibid 311.

and catastrophic risk, the traditional preliminary injunction analysis starts to unravel.[836] Johnson[837] and Peterson[838] make the same essential point that the preliminary injunction-analysis breaks down when confronted with the extreme facts of the LHC dispute. However, both authors highlight an alternative to the four-factored standard examined above.[839] This alternative may facilitate preliminary injunctions in *speculative cases*. Both authors refer to *Earth Island Institute v U.S. Forest Serv* to illustrate the alternative approach.[840] In this case, the Court recognised the "traditional" four-factored criteria.[841] However, it also recognised the alternative where a court can grant a preliminary injunction should the claimant demonstrate that (1) *serious questions are raised* and (2) the *balance of hardships tips sharply in his favour*.[842] This approach is somewhat reminiscent of the *American Cyanamid* guidelines used in the English remedial landscape.

Compare the *balance of hardships*: The injunction will halt a tremendously expensive scientific undertaking and adversely impact physicists around the world. On the other hand, the injunction is sought to prevent an x-catastrophe which will destroy the present and humanity's future. According to Johnson, the alleged black-hole disaster tips the balance of hardships *sharply* in the claimant's favour. [843] On the other hand, and as recognised by Johnson, one potential objection is that the 'gross amount of hardship' on behalf of the claimant should be discounted by the exceptionally low probability of the disaster.[844] However, and without resolving this balancing issue, the main problem under the alternative approach arguably resides elsewhere; the claimant must prove that 'serious questions' are raised. As clarified by Johnson, the court must decide if there are questions which ought to be *taken seriously*.[845] The underlying science is highly complex, the experts vouch for the safety of the project and the probabilities are uncertain; in effect, it appears implausible that a generalist

[836] ibid.
[837] Johnson, 'The Black Hole Case' (n 68) 870-873.
[838] Peterson, 'The "Sancho" Effect' (n 468) 310-312.
[839] ibid 311; Johnson, 'The Black Hole Case' (n 68) 871.
[840] Peterson, 'The "Sancho" Effect' (n 468) 311; Johnson, 'The Black Hole Case' (n 68) 871.
[841] *Earth Island Inst. v. U.S. Forest Serv.*, 351 F.3d 1291 (9th Cir. 2003) 1297-98.
[842] ibid.
[843] Johnson, 'The Black Hole Case' (n 68) 871.
[844] ibid.
[845] ibid.

judge can evaluate the scientific merits of the claimant's arguments. It appears unlikely that she could say – based on the strength of the scientific arguments themselves – that serious questions arise since the LHC poses an x-risk.

Both Johnson and Peterson suggest that the complex theories and science make it difficult to determine whether the claimant has demonstrated that serious questions are raised.[846] A similar point was made above in relation to English law and interim injunctions; how can the claimant demonstrate that there is a serious issue to be tried?

Nevertheless, Johnson goes on to consider an alternative approach as to how the claimants might prove that there are serious questions to be tried. This is where the meta-analysis comes into play. This is examined in the Part below. As noted in the introduction to this Chapter, Johnson's meta-analysis will eventually be utilised to compartmentalise the deficiency factors that play an essential role in relation to the LHC-inspired intervention.

4.4 THE META-ANALYSIS AND PRELIMINARY INJUNCTIONS

This Part introduces Johnson's meta-analysis. In particular, it considers his article *'The Black Hole Case: The Injunction Against the End of the World'*.[847] The thesis refers to this article as the Black Hole Case Article below. Beyond this piece, Jonson has produced other relevant works, including '*Agencies and Science-Experiment Risk*', '*Uncertain Risk, Science Experiments, and the Courts*', and '*Judicial Review of Uncertain Risks in Scientific Research*'.[848]

It is useful to clarify some key terminology before proceeding. The following terminology should be noted:

[846] Johnson, 'The Black Hole Case' (n 68); Peterson, 'The "Sancho" Effect' (n 468).
[847] Johnson, 'The Black Hole Case' (n 68).
[848] Johnson, 'Agencies and Science-Experiment Risk' (n 68); Johnson, 'Uncertain Risk, Science Experiments, and the Courts' (n 68); Johnson, 'Judicial Review' (n 68).

Table 4.1

Black Hole Case	A legal dispute (not involving particle accelerators) where review barriers similar to those identified in Chapter 3 re-emerge and make external review impractical or difficult.
LHC Case	Any court case (such as *Sancho v US Department of Energy)* that spawned from the LHC Controversy.
LHC Controversy	The disagreements and uncertainty as to whether high-energy physics experiments facilitated by particle accelerators (i) pose any existential risk to humanity, (ii) pose an unacceptable existential risk to humanity, or (iii) should be delayed or suspended.

4.4.1 Overview

The Black Hole Case Article deals with the LHC Controversy—mainly the LHC dispute—from a legal perspective. In this 90-page paper, Johnson explores various themes, including the scientific, social and political aspects of the controversy and the challenges that may adversely affect claimants and judges in case like the LHC case.

Furthermore, his article offers "…a set of analytical and theoretical tools that are usable in the courts for dealing with [the LHC case] and cases like it."[849] Johnson intends for the tools to find application beyond the LHC case. He proceeds to say: "If litigation over the LHC does not put a judge in the position of saving the world, another case soon might."[850] Highlighting climate change, genetic engineering, nanotechnology and artificial intelligence, he suggests that there is a non-trivial probability that courts will come face to face with a real doomsday scenario in the future.[851] As such, he proposes that courts need analytical methods to make fair and principled decisions despite the challenges that such cases present.[852]

Recognising that it may be beyond courts to settle disputes on the scientific merits in black-hole cases, Johnson proposes that courts are "…well equipped to look at the human aspects in a prospective catastrophe and render a decision on an injunction

[849] Johnson, 'The Black Hole Case' (n 68) 822.
[850] ibid.
[851] ibid.
[852] ibid.

request on the basis of those factors."[853] His article deals with four categories of meta-analysis that can apply to the LHC case and other black-hole cases.[854]

Conducing their analysis on a higher level, "[u]sing a kind of meta-analysis" explains Johnson, courts should examine the potential for (i) *Defective Theoretical Groundings*, (ii) *Faulty Scientific Work*, (iii) *Credulity and Neglect*, and (iv) *Bias and Influence*.[855] The following lines summarises the approach that Johnson advocates:

> Using a kind of meta-analysis, courts should gauge the risk that scientific judgments are wrong. Relevant subjects of inquiry include organizational culture, group politics, and psychological context. The particular aspects of scientific arguments should also be scrutinized on a meta level. Relevant issues here include the newness of underlying theory, the complexity of the chain of argument, the likely reliability of underlying data, and so on. Also relevant is what history has to say about the durability of pronouncements made in the field.[856]

One may wonder why Johnson promotes "a kind of meta-analysis"[857] for dealing with the LHC case. This is where the review barriers exposed in Chapter 3 come into play. For instance, the scientific substance underpinning the dispute is exceptionally complex. Few lawyers and judges would be able to deal with the scientific merits.[858] According to Johnson, the LHC casefile is "…replete with the infinite and the unknowable," and "[t]he facts are, quite literally, more complex than anything on Earth."[859] The point is that neither judges nor lawyers can realistically be expected to make an independent evaluation of the underlying science. Few would perhaps be ready to disagree with this point. And this brings to light another issue – the experts.

Another reason for the meta-analysis is that there are good reasons for being cautious when relying upon those who understand the underlying science. For instance, Johnson makes the point that "[e]very expert has a very personal stake in the

[853] ibid 885; See also Johnson, 'Uncertain Risk, Science Experiments, and the Courts' (n 68) 367-375.
[854] Johnson, 'The Black Hole Case' (n 68) 886.
[855] ibid 883-907; Johnson, 'Uncertain Risk, Science Experiments, and the Courts' (n 68) 369-375.
[856] Johnson, 'The Black Hole Case' (n 68) 885.
[857] ibid.
[858] Posner, *Catastrophe* (n 90).
[859] Johnson, 'The Black Hole Case' (n 68) 821 and 870-872.

matter."[860] The potential lack of objectivity is not a novel feature in the context of the LHC Controversy as seen in Chapter 3.

In turn, a related problem revolves around testing and confirming the theories offered by the disputants. As professed by Johnson, it is not possible to "…eschew the experts in favor of some hard, physical proof."[861] By way of example, the theories of the critics can only be confirmed "…through the obliteration of the court, the parties, and the planet."[862] This observation is central within the x-risk landscape: "Theories that involve the end of the world are not amenable to experimental verification—or at least, not more than once."[863] Citing a piece in *The New Yorker*, Ord attributes the lines above to Sagan.[864] In one his articles, Sagan writes in slightly different words that:

> Part of the resistance to serious consideration of such apocalyptic pronouncements is their necessarily theoretical basis. Understanding the long-term consequences of nuclear war is not a problem amenable to experimental verification—at least not more than once.[865]

The bottom line: Johnson promotes the meta-analysis because such an approach may allow a court to provide meaningful adjudication despite the formidable challenges—including those mentioned above—that characterised the LCH Controversy.

It is important to recall that Johnson explores the analytical tools within the context of preliminary injunction requests. As mentioned earlier in Chapter 4, and as noted in Johnson's article, a judge may—as a matter of American law—issue a preliminary injunction if "serious questions" are raised and if the hardship caused by the injunction would tip "sharply" in favour of the claimant.[866] It is also worth recalling that this formulation is similar to the *American Cyanamid* guidelines used in English law when a judge considers whether or not to issue a preliminary injunction.[867]

---

[860] ibid 873.
[861] Johnson, 'The Black Hole Case' (n 68) 821; Wilson, 'Minimizing' (n 151) 357.
[862] Johnson, 'The Black Hole Case' (n 68) 873.
[863] Ord, *Precipice* (n 16) 197.
[864] ibid 197.
[865] Carl Sagan, 'Nuclear War and Climatic Catastrophe: Some Policy Implications' [1983] Foreign Affairs 257, 258.
[866] Johnson, 'The Black Hole Case' (n 68) 871.
[867] See Chapter 4.

Given that withholding the injunction in the LHC case may destroy humanity's past, present and future, the hardship caused by withholding the injunction most definitely tips "sharply" in favour of the claimant. In effect, and as noted previously, the analytical aspect that is more problematic is that of considering whether "serious questions" are raised. How can the court determine whether there are such questions given the review barriers? What methodology can be employed by the court? Again, this is where the meta-analysis is relevant.

The analytical tools are what the judiciary can use "…in evaluating whether there are the sort of "serious questions" that justify a preliminary injunction…"[868] Therefore, one is to appreciate that the court's ultimate aim, according to Johnson, is to analyse whether there are "serious questions"[869] by considering the potential for (i) Defective Theoretical Groundings, (ii) Faulty Scientific Work, (iii) Credulity and Neglect, and (iv) Bias and Influence. Chapter 6 argues that these grounds align well with the field of x-risk and can be used as a device in compartmentalising the deficiency factors that might be analysed as part of evaluating the *reliability* of the x-risk assessment.

4.4.2 The categories of meta-analysis

Having outlined how the analytical tools intersect with preliminary injunction requests, this section will briefly summarise the four categories of meta-analysis. This examination serves as a precursor to building upon Johnson's analysis in proposing a broad legal intervention for x-risk mitigation. According to Johnson, the categories are not "strictly distinctive", and there is "considerable overlap" between them.[870] The categories of meta-analysis that, according to Johnson, can be applied to the LHC case as part of a preliminary injunction request are as follows:[871]

[868] Johnson, 'The Black Hole Case' (n 68) 885.
[869] Naturally, if the court finds that "serious questions" arise, a judge may issue a preliminary injunction.
[870] Johnson, 'The Black Hole Case' (n 68) 886.
[871] ibid.

Table 4.2

CATEGORY	SUMMARY
i. Defective Theoretical Groundings	➢ Concerns **macro-scale** scientific error. ○ Analyse and give weight to the potential that the relevant risk assessment is based on flawed or incomplete scientific theory.
ii. Faulty Scientific Work	➢ Concerns **micro-scale** scientific error. ○ Analyse and give weight to the potential mistakes – such as miscalculations – of the individuals having produced the relevant risk assessment(s).
iii. Credulity and Neglect	➢ Concerns **innocent and unintentional** mistakes. ○ Analyse and give weight to the potential mistakes in risk assessment work and decision-making attributable to cognitive biases and sociological factors.
iv. Bias and Influence	➢ Concerns **non-innocent** errors. ○ Analyse and give weight to the potential self-interest of the individuals that produced the relevant risk assessment and sanctioned the endeavour.

Categories (i) and (ii) above focus on the scientific work. In contrast, categories (iii) and (iv) concern the scientists.[872] These categories are neither meant to operate as four absolute requirements nor cumulative conditions. In other words, according to Johnson, the court may "…decide on the basis of any one of these categories, or some combination of more than one, that an injunction is warranted."[873]

In a subsequent article from 2021, Johnson maintains his standpoint that courts can and should undertake a "…kind of meta-analysis that gets above the level affected by uncertainty."[874] For this purpose, he revisits the four categories set out in the Black Hole Case Article.[875] However, he also adds a new category which invites the judge to consider "…the potential for fraud, lies, and faked results."[876]

[872] ibid.
[873] ibid.
[874] Johnson, 'Uncertain Risk, Science Experiments, and the Courts' (n 68) 367 and 369-371.
[875] ibid 369-375.
[876] ibid 369.

4.4.3 A predecessor to the Black Hole Case Article

Before formalising his argument in the Black Hole Case Article, Johnson circulated his general idea in some blog posts on *PrawfsBlawg*.[877] Portions of the Black Hole Case Article are "closely based" upon these blog posts, a relationship that Johnson notes in his article.[878]

This Section will briefly consider these posts because they are accompanied by some interesting comments that question Johnson's meta-analysis. In addition, other scholars have considered and cited Johnson's blog post, but few have critically analysed the meta-analysis in the Black Hole Case Article itself. There are five posts on *PrawfsBlawg.* The posts offer an introduction to the LHC Controversy and Johnson's suggested approach for how the judiciary might cope with the challenges flowing from it. In the third post, Johnson begins to sketch the foundation for the meta-analysis that was later expanded upon in the Black Hole Case Article.[879]

There are a few notable differences between the blog posts and the article. A big difference is this: The Black Hole Case Article considers four categories of meta-analysis. Two focus on the *scientific work*, whereas the other two focus on the *scientists*. The blog posts focus only on the scientists and the notion that courts should "…scrutinize the culture of CERN and the particle-physics community, as well the political, social, and psychological context in which their decisions are made."[880] In other words, the posts do not consider the categories of meta-analysis concerning the scientific work.

Another interesting aspect is that the blog posts come with comments that readers have left in response to Johnson's suggestions. Some of these are remarkable

[877] Eric Johnson, 'Could Bad Judging Cause the Earth to Be Sucked Into a Black Hole? Maybe.' (*PrawfsBlawg*, 21 October 2008) <https://prawfsblawg.blogs.com/prawfsblawg/2008/10/could-bad-judgi.html> accessed 28 April 2023.
[878] Johnson, 'The Black Hole Case' (n 68) 819.
[879] Eric Johnson, 'Culture and Inscrutable Science: An Analytical Method for Preliminary Injunctions in Extreme Cases' (*PrawfsBlawg*, 24 October 2008) <https://prawfsblawg.blogs.com/prawfsblawg/2008/10/culture-and-ins.html> accessed 28 April 2023.
[880] ibid.

because they highlight potential weaknesses in Johnson's suggested approach. Moreover, if taking the comments at face value, both Walter Wagner and Luis Sancho – the claimants in the *Sancho case* – left separate comments in response to the posts.

Furthermore, some comments seemingly influenced the substance of the Black Hole Case Article. For example, one poster under the name 'James Blodgett' writes (and sets out to show) that "[t]he idea that scientists are adequately careful and knowledgeable about risks…is refuted by the history of 'safety factor' arguments for colliders."[881] This idea became a prominent theme in the Black Hole Case Article. For instance, in the article Johnson writes that "…what history has to say about the durability of pronouncements made in the field…" is a relevant component in considering if "serious questions" exist.[882] This theme is particularly relevant for the category of meta-analysis that revolves around analysing the potential for *defective theoretical groundings*.[883]

This Section will not deal with all comments. Instead, it mentions a selected few that may allow a critic to transpose similar arguments against the Black Hole Case Article. Some posters question the involvement and suitability of generalist courts when it comes to Johnson's approach. A poster under the name 'James Grimmelmann' writes that it is not acceptable – neither socially nor rhetorically – for judges to render decisions that admit that they do not understand the underlying science.[884] His comment also reads: "Even if they actually make their decisions based on social factors, as…[Johnson advocates], any opinion that admits the judge has simply given up on the science involved is going to be profoundly unpersuasive."[885] Ellis makes a somewhat similar observation in response to the Black Hole Case Article. In *Physics*

[881] See James Blodgett's comment in the comment section to Eric Johnson, 'Scientocracy and the Need for Judicial Process' (*PrawfsBlawg*, 10 November 2008) <https://prawfsblawg.blogs.com/prawfsblawg/2008/11/scientocracy-an.html> accessed 28 April 2023.
[882] Johnson, 'The Black Hole Case' (n 68) 885.
[883] ibid 887-890.
[884] See James Grimmelmann's comment in the comments section to Johnson, 'Culture and Inscrutable Science' (n 879).
[885] ibid.

World, it is reported that Ellis "…does not see how a court can avoid the question of whether the science is right, adding that 'the job of the court is to get at the truth'."[886]

One may perhaps respond by suggesting that the court is not asked to disregard the truth due to opaque science. Instead, the court is guided to take an alternative route towards the truth given the many obstructions hindering it from reaching the same destination by taking another, perhaps more traditional, route. This general position is also applicable to the comment left by 'James Grimmelmann'. There is no reason to frame the approach as one where the judge is giving up on the science involved. Instead, mindful of the analytical difficulties of the LHC case, the judge is focusing on alternative aspects that are still relevant to the dispute at hand.

An interesting paradox is also noted by a poster under the name 'JP' who writes that courts are probably "…not up to the tasks…" that Johnson proposes.[887] In particular, and in relation to this paradox, 'JP' suggests that any judge who halts the LHC will likely face a lot of criticism and attract an unwanted reputation. The same will not happen if the judge simply leaves the project alone. The paradox, then, is that judges may find themselves in the same position as the hypothetical particle physicist who might consider whether or not to speak up against the scientific project. In fact, the judge may even be in a worse position owing to the lack of expertise which may exaggerate the issue.[888] While this thesis will not deal with this issue in any detail, Winter, for example, notes in his article that "[i]t has long been recognized that judges are not able to exclusively decide cases based on legal principles."[889] As an example, and with reference to two legal studies, Winter highlights that potential gains for the judge's personal career can play an important role when making a judgment.[890] It seems reasonable to believe that a judge would think twice about, e.g., reputational ramifications of halting a very real and fascinating multi-billion dollar project based on speculative 'end of the world' theories extended by non-physicists.

---

[886] Edwin Cartlidge, 'Law and the End of the World' *Physics World* (02 February 2010) <https://physicsworld.com/a/law-and-the-end-of-the-world/> accessed 28 April 2023.
[887] See JP's comment in the comments section to Johnson, 'Culture and Inscrutable Science' (n 879).
[888] ibid.
[889] Christoph Winter, 'The Value of Behavioral Economics for EU Judicial Decision-Making' [2020] German Law Journal 240, 240.
[890] ibid.

In another comment, the (presumably same) poster under the name 'JP' accepts Johnson's premise that it might be problematic to leave regulation to self-interested scientists. However, the poster (once again) questions Johnson's confidence in judges and the judicial process writing: "There are multiple democratically-accountable political government entities that could regulate projects like the LHC. Why would judge-made law be better?"[891]

Rather than questioning the involvement of courts, some posters take a stab at the general attitudes underlying Johnson's approach. For example, a poster under the name 'Ben' writes that Johnson is placing shockingly "…little faith in scientists."[892] Another poster under the name 'C.E. Petit' shares this sentiment and points to, among other things, the openness and transparency by which the LHC advocates' and CERN have dealt with the safety questions.[893] This section will be concluded with an excerpt from the same comment, where 'C.E. Petit' writes:

> From the scientists' perspective, the arguments being raised resemble an argument that Columbus should not be allowed to travel west seeking Indian spices because he would unbalance the planet when he sails off the edge in the middle of the Atlantic.[894]

(i) A procedural approach?

Euan MacDonald has dealt with Johnson's analysis on *PrawfsBlawg*.[895] MacDonald's post appears on a blog dedicated to the Global Administrative Law ('GAL') Project. The GAL Project is part of the research of the Institute for International Law and Justice

[891] See JP's comment in the comments section to Johnson, 'Scientocracy and the need for Judicial Process' (n 881).
[892] See Ben's comment in the comments section to Johnson, 'Could Bad Judging' (n 877).
[893] See C.E. Petit's comment in the comments section to Johnson, 'Scientocracy and the need for Judicial Process' (n 881)
[894] ibid.
[895] Euan MacDonald, 'GAL - and The End Of The World As We Know It' (20 January 2009)<https://globaladminlaw.blogspot.com/2009/01/global-administrative-law-and-end-of.html> accessed 28 April 2023.

(IILJ) at the NYU School of Law.[896] Navigating to the web page that introduces the GAL Project, the Project is said to focus on

> [T]he increasing use of administrative law-type mechanisms — in particular those related to transparency, participation, accountability, and review — within the regulatory institutions of global governance.[897]

With this in mind, MacDonald's post can be examined in some further detail. The first point to note is that he is generally positive towards Johnson's approach and arguments. However, like some comment-posters on *PrawfsBlawg*, he ultimately questions whether courts are ideal for the tasks proposed by Johnson.[898]

There is an underlying theme in MacDonald's post that is interesting. That theme is the interrelationship between GAL and Johnson's analysis. According to MacDonald, GAL seems to be conceptually suitable for framing the issues and solutions that Johnson highlights. For example, MacDonald writes that Johnson is, in essence, proposing a procedural approach for overcoming the challenges in the LHC case, and such an approach, claims MacDonald, fits neatly together with the GAL Project.

It might be useful to briefly repat Johnson's suggestion on *PrawfsBlawg*.[899] The suggestion is for courts to focus on, e.g., social dynamics, psychological factors and organisational cultures to determine whether "serious questions" exist. As mentioned, MacDonald believes that this is an investigation into the procedures rather than the scientific merits of the theories underlying the dispute. Following this observation, he makes a proposition that involves transposing Johnson's analysis beyond its original remits. More specifically, MacDonald notes the potential for a more mature GAL

---

[896] Institute for International Law and Justice, 'Projects' <https://www.iilj.org/projects/> accessed 28 April 2023.

[897] Institute for International Law and Justice, 'Global Administrative Law' <https://www.iilj.org/gal/> accessed 28 April 2023.

[898] McDonald's post concludes with a discussion of immunity and jurisdiction, which leads him to question whether national courts would be "…the best fora in which to review the CERN's evaluation of the safety of its own operations." Euan MacDonald, 'GAL - and The End Of The World As We Know It' (n 895).

[899] Again, MacDonald is not responding to Johnson's article(s). Instead, he is looking at the blog posts.

insisting upon broad sets of procedural guarantees. As one may have guessed, these guarantees can take inspiration from Johnson's analysis.

MacDonald considers that a set of broad procedural guarantees—realised through GAL—may avert the need for courts to resort to *ad hoc* categories and tests. Another good observation from MacDonald is that such guarantees may, in fact, benefit organisations like CERN (and possibly other x-risk originators). This is because the procedural guarantees may enhance the legitimacy of an organisation's safety assessments. In turn, this may help the organisation to fend off some of the bad press that might accrue. And bad press is, as Larsen demonstrates, certainly not a novel problem for CERN and particle accelerators more generally.[900] The Black Hole Case Article ended up referring to MacDonald's post at one point.[901] However, MacDonald did not persuade Johnson to reconsider his court-centric approach.

The posts by Johnson on *PrawfsBlawg* are also cited in Adams's article dealing with the regulation of particle physics experiments.[902] More specifically, Adams refer to these posts in discussing some of the obstacles that complicated the lawsuits aiming to halt the LHC.[903] A few pages down (and separately) from having cited Johnson's posts, Adams displays doubt as to whether existing courts or hypothetical science courts can singlehandedly regulate experiments within particle physics.[904]

Irrespective of whether one believes that a science court is suitable as a governing body, what is needed the most, writes Adams, is a set of guidelines or rules that applies to this area of science.[905] To this effect, his article goes on to consider the outline of a regulatory framework for experiments within particle physics. The important takeaway from Adams's article for present purpose is that a court-centric approach is, once again, questioned.

[900] Larsen, *Particle Panic!* (n 265).
[901] Johnson, 'The Black Hole Case' (n 68) 861.
[902] Adams, 'Honey' (n 309) 148-149.
[903] ibid.
[904] ibid 153-154 and 157-158.
[905] ibid 158.

4.5 JUDICIAL REVIEW VERSUS INTERAGENCY REGULATION

The pivotal issue in this Part concerns Johnson's court-centric approach. As seen above, some have questioned his suggestion on the ground that the judiciary might not be the ideal institution for the tasks proposed by Johnson.

In short, should courts (as opposed to e.g., another agency) undertake the kind of analysis that Johnson proposes in controlling certain research endeavours? The Black Hole Case Article is about demonstrating that the judiciary *could* use the so-called meta-analysis. However, in *Agencies and Science-Experiment Risk* from 2016, Johnson makes the argument that courts *should* deal with issues such as those in the LHC Controversy (by applying the meta-analysis).[906]

In his 2016 article, Johnson considers *catastrophic and irreversible ultrahazards*, and government agencies undertaking scientific research. A preliminary observation is that Johnson's typology of risk is problematic if one accepts the ASP since it includes both existential and non-existential risks. In any event, Johnson suggests that self-regulation is inadequate as a mode of governance when an agency's research activities produce a risk falling within his classification scheme of risk.[907]

He writes: "If an agency's own scientific program presents a plausible risk of catastrophe, then the agency itself should not be trusted to make the decision about whether that risk is acceptable."[908] The two case studies in his article give weight to this argument.[909] A number of general factors, some of which played out in these case studies, are cited in support of this proposition.[910] These factors include the notion that (i) agencies are acting in the interest of self-preservation so that safety issues – that could result in delays or resistance – may be undermined by the organisations and their employees, [911] (ii) scientists may – as seen vis-à-vis the history of nuclear physics

[906] Johnson, 'Agencies and Science-Experiment Risk' (n 68).
[907] ibid 535-555 and 567-583.
[908] ibid 567.
[909] Involving NASA and the Department of Energy. ibid 535-553.
[910] ibid 567-582.
[911] ibid 571-573.

– be prone to conflicts of interest and misjudge risk,[912] and (iii) insights from behavioural economics and cognitive psychology may indicate that scientists and agency managers – like people more generally – are prone to making irrational judgements when evaluating the safety of their activities.[913]

To reinforce the final point, Johnson explores probability neglect, optimism bias and the availability heuristic.[914] He then suggests that the political process will likely not make up for issues "…such as agency self-interest and the cognitive biases of agency decision makers."[915] The ensuing argument is that the judiciary can solve the problem and governance gap by reviewing agency actions when these allegedly create a catastrophic and irreversible ultrahazard.[916]

In the final part of the paper, Johnson writes that "[c]ourts are an excellent venue for addressing small-probability/large-harm risks."[917] This is because:

> The judiciary provides a unique opportunity to get beyond mistake-inducing mental heuristics. Courts lack the personal and institutional biases to be found inside agencies. And unlike the public, whose economy of attention is necessarily limited, courts have the luxury of being able to explore a problem in depth. Moreover, civil discovery and the adversarial process can uncover important facts that agency insiders may be prone to ignore. And the questions and arguments of opposing counsel can sharpen the factual picture. Simply put, courts are good for providing review of agency experiment-risk questions for the same reasons courts are good for contract disputes, criminal charges, and everyday negligence claims: Courts have the time, resources, and independence needed to get to a fair result.[918]

Coming in at about 60 pages, Johnson only devotes a little over two pages to expressly justify why courts should be seen as providing an apt governance solution. The article mainly considers why *self-regulation* is not ideal and why flawed decision-making may flow from such a governance scheme.

---

[912] ibid 573-578.
[913] ibid 578-582.
[914] ibid 578-580.
[915] ibid 582.
[916] ibid 583-586.
[917] ibid 583.
[918] ibid 583-584.

While it is reasonable to believe that self-regulation is problematic, it does not follow that external review and oversight should come from the judiciary. Why not from another agency? This question is revisited below. Before that, it is useful to note that the judiciary has already proven unsuccessful in dealing with the two catastrophic and irreversible ultrahazards examined in Johnson's article from 2016. In fact, his own Article highlights:

> [A] lack of judicial scrutiny is not because no one has tried to get the courts involved. Both the Cassini-plutonium and RHIC-strangelet controversies precipitated lawsuits…But none of it precipitated meaningful review on the issue of acceptable risk. In fact, none of the plaintiffs even achieved what could be charitably described as a moral victory.[919]

In the former lawsuit, NASA could easily demonstrate compliance with the relevant law (i.e., NEPA), and the latter suit was dismissed on procedural grounds.[920] Likewise, the judiciary failed to deal with the alleged risk examined in the Black Hole Case Article.[921] These failures may be taken as indicative of the future role of the judiciary. For instance, Wilson contends that the LHC cases illustrate that "…neither domestic nor international courts are equipped to handle disputes involving low-probability, high consequence advanced technologies. There seems to be no court that offers reliable judicial relief for such situations…."[922]

On the other hand, these past failures do not necessarily bar the argument that the judiciary should review agency actions that purportedly generate a catastrophic and irreversible ultrahazard or an x-risk. However, they indicate that other options also demand attention and exploration.

One of the more pressing questions that arise in relation to Johnson's proposal is this: Why should courts be viewed as offering the optimal solution to the governance gap instead of another institution? This is a comparative question. In fact, in addition to judicial review and court-issued injunctions, Johnson acknowledges that ex ante

[919] ibid 555.
[920] ibid 565.
[921] See *Sancho case* (n 87).
[922] Wilson, 'Minimizing' (n 151) 337.

regulation can spring from another agency.[923] In other words, and as written in his article, it is possible to ensure that "…one agency's research activity is regulated by another agency."[924] However, Johnson neither explores the *interagency regulation approach* nor compares it to the *judicial review approach* in serious detail.[925] This is unfortunate because the nature of irreversible risks (or x-risks) and the past failures involving courts may, prima facie, suggest that regulation by another agency is more appropriate than judicial oversight.

One question to consider is whether Johnson offers any justifications for choosing the judicial review path rather than the interagency regulation path. One possible explanation is that the latter approach will lead nowhere, owing to a lack of political incentive to make it happen in practice. In a similar vein, Johnson notes that Congress can, in theory, ensure the availability of judicial review of relevant agency actions by, in particular, amending the Administrative Procedure Act or the Federal Tort Claims Act.[926] Despite this theoretical possibility, Johnson views congressional action as unlikely "…absent a catastrophe of sufficient magnitude to set the political process in motion."[927] In the x-risk circle, Ord[928] and Bostrom have expressed similar notions. For instance, Bostrom writes: "[I]t is possible that [existential-risk mitigation] will at some point receive a boost from the occurrence of a major (non-existential) catastrophe that underscores the precariousness of the present human condition."[929] Given Johnson's view that congressional action is improbable, he proposes that "[t]he institution best-suited to assert the need for judicial review is the judiciary itself."[930]

In sum, the perceived lack of political incentive to address the governance gap by, for example, facilitating interagency regulation may provide some justification for Johnson's court-centric approach. But the lack of political incentive cannot solely

[923] Johnson, 'Agencies and Science-Experiment Risk' (n 68) 553-555.
[924] ibid 553; Matthew Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies' [2016] Harvard Journal of Law & Technology 353, 381-388.
[925] However, the article considers 'interagency regulation'. Johnson, 'Agencies and Science-Experiment Risk' (n 68) 553-555.
[926] ibid 584. For a discussion of the APA and the FTCA see ibid 557-564.
[927] ibid 584-585.
[928] Ord, *Precipice* (n 16) 200.
[929] Bostrom, 'Existential Risk Prevention' (n 15) 28.
[930] Johnson, 'Agencies and Science-Experiment Risk' (n 68) 585.

justify the argument that the judiciary should provide the solution. Furthermore, it does not explain why generalist judges are better positioned to deal with the acceptability of a catastrophic risk from an agency's actions than other external persons like those in another regulatory agency.

Another justification for choosing the judicial review approach is found not in his 2016 article but in a separate book chapter authored by Johnson.[931] Having considered how courts could deal with catastrophic risks and extreme uncertainty from science experiments, Jonson writes:

> While the upshot of my analysis is that the courts can, indeed, do a good job in handling uncertainty in the science-experiment-risk context, a word should be said in closing about whether courts should perform this function.[932]

He continues:

> The courts are well-positioned to provide both trust and a meaningful sense of control. The courts supply an avenue to trust through their role of gathering and impartially weighing evidence. And the openness of the courts to hearing complaints of affected parties can provide people everywhere with a sense of control over their own destinies. In the overall analysis, it does not appear that the existence of uncertainty militates against courts engaging in questions of uncertain risk. Instead, thoughtful reflection seems to show that judicial resolution is particularly appropriate.[933]

As seen from these lines, there are perhaps some justifications for Johnson's court-centric approach. However, whether the judiciary *should* deal with the risks manifested in the LHC Controversy—whether the risk is framed as an x-risk or not—is a question that should ideally be considered against how well other institutions can handle the same situations. It seems that one of the stronger arguments for a court-centric approach is that courts are better equipped than other institutions to facilitate external oversight of similar actors and risks. However, such an argument is not made out.

---

[931] Johnson, 'Judicial Review' (n 68).
[932] ibid 84.
[933] ibid.

To question the court-centric approach, the upcoming Section will consider whether external agencies can, prima facie, provide a more suitable alternative than the judiciary in regulating alleged x-risk originators such as CERN. This question requires one to consider and compare the institutional competence of courts and agencies more generally. The aim is not to deal with this point in detail. Instead, the Section below will highlight some general aspects that may lead someone to, prima facie, favour the *interagency regulation approach* over the *judicial review approach*. Scherer's article is suitable for this very purpose.[934]

4.5.1 Institutional Competence

Scherer's article considers the regulation of AI,[935] which, as seen in Chapter 2, is amongst the hazards within the x-risk landscape. As part of this ambit and given the characteristics of AI, Scherer compares the competencies of legislatures, agencies and courts. Before such comparative analysis, his article highlights some of the challenges when it comes to regulating AI.[936] Scherer also notes some of the literature suggesting that AI may pose GCRs and x-risks.

After outlining the regulatory challenges, his article considers "…the comparative competencies of [legislatures, agencies, and courts] for managing the public risks associated with AI."[937] Scherer recognises that "…no institution has a monopoly on any particular competence."[938] For example, while agencies generally have more subject-matter expertise than courts, the latter can enlist the support of experts to close the gap.[939] As another example, agencies may have more flexibility to act ex ante than courts. However, as Scherer notes, it is questionable how often agencies exercise this freedom.[940] He acknowledges that all characteristics examined in his paper are subject to similar qualifications. Despite this recognition, Scherer's view is

[934] Scherer, 'Regulating Artificial Intelligence Systems' (n 924).
[935] ibid.
[936] These include laying down a satisfactory definition of AI for regulatory purposes, attaching liability when an *autonomous* AI system cause harm, and controlling the nature of AI R&D. ibid 359-373.
[937] ibid 376.
[938] ibid 377.
[939] ibid.
[940] ibid 378.

that the general characteristics are "…instructive in their implication for whether and how AI might be effectively regulated."[941]

There is a wealth of material considering the relative competencies of courts and agencies. Nevertheless, the aim here is simply to draw on some general observations that can be instructive as to whether agencies or courts have the prima facie upper hand in dealing with x-risks from complex technologies and scientific endeavours. Whereas Scherer considers the intuitional strengths and weaknesses of legislatures,[942] we will only consider some of the competencies of agencies and courts.

Firstly, agencies enjoy more institutional flexibility than courts.[943] They can be staffed by experts, make decisions based on wide social factors and be "tailor-made…for the resolution of a particular social problem."[944] From the wide ambit of protecting human health and the environment to the much narrower one of licensing acupuncturists, the given mission scope of any given agency can greatly.[945] Secondly, agencies and their staff can become specialists, whereas judges are often generalists.[946] However, one can question how well an agency's expertise will hold up and stand the test of time in the context of, for instance, AI regulation. This is because AI research draws from many fields and is open to quick and sudden change.[947] Thirdly, Scherer notes that agencies can undertake *independent* investigations, whereas courts must generally rely on the record developed and presented by the opposing parties.[948] Fourthly, courts can be seen as "…inherently reactive institutions…" whereas agencies can often intervene before harm occurs.[949] Nevertheless, Scherer highlights that an agency's capacity to act ex ante may be weakened in practice, and especially when tasked with regulating emerging technologies. For instance, there may be hasty changes in the underlying research, the expertise of the agency's staff may wane owing to these sudden changes, and the process of promulgating or amending rules

[941] ibid.
[942] ibid 378-381.
[943] ibid 382.
[944] ibid 381.
[945] ibid 383.
[946] ibid 383-384.
[947] ibid 384-385.
[948] ibid 386.
[949] ibid 387.

might be too slow to enable the agency to act in a timely manner.[950] Despite such observations, an agency's greater ability to act ex ante can offer some benefits, such as being able to disseminate information about risk as well as encourage voluntary industry standards.[951]

As mentioned above, courts generally have a lesser capability to act ex ante when compared to agencies. But not only will many cases be brought only after the relevant harm has materialised, "…the ability to introduce information regarding broader social and economic considerations is limited…" once the claim has been brought.[952] Scherer highlights that the ensuing effect is that courts may focus more on the risks and less on the benefits of the relevant technology.[953] This is generally speaking not desirable when dealing with technologies and endeavours that may give rise to x-risk. In contrast, agencies have more flexibility to consider wider social and economic factors. Such considerations, writes Scherer, "…are often the whole point…" in the administrative policymaking process.[954] Nevertheless, the features above may also place courts in a better position than agencies to deal with, for instance, liability questions. For example, the process of litigation can expose and uncover information pertaining to the dispute at hand such as information about the safety features (or lack thereof) of a product that has caused harm.[955] Still, whereas the "…intensive discovery and fact-finding processes of civil litigation…"[956] can give courts the upper hand in allocating responsibility once harm has occurred, the same process can undermine several facets concerning the disputed technology:

> [B]ecause both discovery and the presentation of evidence at trial will focus on the features of the product that led to the harm (and the absence of features that could have prevented the harm), the judge and jury may not have any occasion to consider the broader risk profile of the disputed technology. Each case, taken individually, thus provides an incomplete — and even misleading — factual picture of the technology at issue.[957]

[950] ibid.
[951] ibid 387-388.
[952] This is because "procedural and evidentiary rules act to focus attention on the specific facts that led to harm in that case…" ibid 388.
[953] ibid.
[954] ibid.
[955] ibid 389.
[956] ibid.
[957] ibid 389.

Furthermore, the adversarial system can potentially lead to a situation where scientific rigour is undermined. For instance, Scherer notes that "[s]trategic considerations, rather than scientific rigor, drive the parties' decisions regarding what witnesses to call and what evidence to present."[958] Ultimately, it is often possible to find a 'qualified' expert who can testify in support of each side.[959] This may affect the courts' ability to deal with some of the questions that may arise in the context of emerging technologies and scientific endeavours. This is especially so if the questions focus on the scientific rigour and merit of the competing claims.[960] In contrast, and as mentioned above, an agency can be staffed by experts, and "…an agency consisting of people with prior knowledge of the relevant field is less likely to be hoodwinked than a lay jury or a generalist judge."[961]

Finally, Scherer notes that the common law develops rather slowly, and that "[t]he path from the filing of the suit to final adjudication is long and winding…"[962] Whereas interim remedies, like interim injunctions, may be available, the duration of the process of litigation is something to keep in mind. In fact, the LHC case has been cited an example of the lengthy process with Marchant writing:

> A (hopefully) trivial yet illustrative example of the slow pace of courts is a legal action in the European Court of Human Rights to stop the start up of the Large Hadron Collider in Europe because of an alleged risk it could start a runaway reaction that could destroy the earth. After the court denied an interim order to delay the experiment, a news report quoted a court spokesperson as saying it could "take several years" to decide the merits of the case…[963]

Where does all of this leave us? In short, one may argue that agencies can, prima facie, provide a more promising alternative than the judiciary in facilitating external oversight of actors such as CERN when their activities may pose an existential risk.

---

[958] ibid 392.
[959] ibid.
[960] One will recall that Johnson's approach is one where a generalist judge can deal with issues of risk without understanding "the science as scientists do…". Johnson, 'Agencies and Science-Experiment Risk' (n 68) 584.
[961] Scherer, 'Regulating Artificial Intelligence Systems' (n 924) 392.
[962] ibid 389.
[963] Marchant, Allenby and Herkert, *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight* (n 424) 24.

X-risks often involve big questions. These risks often intersect with complex science and technology. They demand policy deliberations such as thinking about the broader benefits of the technology or endeavour in question. They demand considerations about the value of humanity's future and whether and when a risk is deemed acceptable given the astronomical stakes. They require an ex ante approach. They necessitate a unified and international response. Anthropogenic existential risks are also a relatively new risk species, and so they compel special attention that accounts for some of the obscuring features that can affect and hinder their mitigation.[964] The nature of these risks and a brief comparison between courts and agencies suggest that the *interagency regulation* alternative – albeit not self-regulation – can offer a better starting point in dealing with and reviewing many existential risks, including those allegedly posed by CERN.

4.6 CONCLUSION

Chapter 4 has predominantly considered Johnson's meta-analysis, which, in his view, could have allowed a generalist judge in the US to deal sensibly with the preliminary injunction request in the *Sancho case*. The meta-analysis is helpful because it can be used notwithstanding the 'review barriers' outlined in Chapter 3. Johnson essentially suggests that the court should focus on human failures rather than the scientific merits of the claim to gauge whether "serious questions" exist. Whereas his court-centric approach is questionable, looking at the human aspects can facilitate external and independent review in the LHC Controversy context.

Chapter 6 will use the four categories of meta-analysis to compartmentalise the deficiency factors that form part of the LHC-inspired intervention. These categories are useful because the meta-analysis aligns with wisdom from existential risk studies. Despite this, Johnson's work is not situated within the remits of x-risk mitigation, and his approach has largely gone unnoticed within the x-risk circle.

[964] These were considered in Chapter 2.

Before sketching the theoretical foundation of the LHC-inspired intervention in Chapter 6, the upcoming Chapter deals with the astronomical stakes proposition and some alternative legal interventions for x-risk mitigation. The ASP is viewed as a fundamental component underpinning efforts to mitigate x-risks. It is also central to the LHC-inspired intervention. As such, the next Chapter provides the last steppingstone before the thesis can advance a broad legal mechanism in response to the LHC Controversy.

5 CHAPTER FIVE: THE ASTRONOMICAL STAKES PROPOSITION AND LEGAL INTERVENTIONS

> What makes existential catastrophes especially bad is not that they would…[cause] a precipitous drop in world population or average quality of life. Instead, their significance lies primarily in the fact that they would destroy the future.[965]

> Even if we are mistaken now, there is vast, unimaginably vast, value in keeping our options open, *in keeping history going*. We owe it to the past lives that suffered so we could exist, and to future lives, for whom we suffer so that they may exist better than we do.[966]

> Utopia is the hope that the scattered fragments of good that we come across from time to time in our lives can be put together, one day, to reveal the shape of a new kind of life. The kind of life that yours should have been.[967]

5.1 BACKGROUND

The x-risk landscape is associated with astronomically high stakes. A risk within this landscape threatens to destroy humanity's past, present as well as future. Regarding the past, there has, as noted by Ord, been cooperation across the generations, and this has been vital to humanity's overall success.[968] An x-catastrophe will break the chain of intergenerational cooperation and dismantle every grand project in want of humanity's collective input.[969] According to Ord, succumbing to such a catastrophe may be regarded as failing our ancestors because the effect is a neglect of any sort of duty that the present generation might have to pay forward the work they did for us.[970] Similarly, Bostrom writes that the present generation may "…have custodial duties to preserve the inheritance of humanity passed on to us by our ancestors and convey it safely to our descendants."[971] Kaczmarek and Beard also indicate how the past can

[965] Bostrom, 'Existential Risk Prevention' (n 15) 17.
[966] Moynihan, *X-Risk* (n 332) 423.
[967] Nick Bostrom, 'Letter from Utopia' [2008] Studies in Ethics, Law, and Technology 1, 6.
[968] Ord, *Precipice* (n 16) 12-13, 49-50.
[969] "We do not want to be the failing link in the chain of generations, and we ought not to delete or abandon the great epic of human civilisation that humankind has been working on for thousands of years, when it is clear that the narrative is far from having reached a natural terminus." Bostrom, 'Existential Risk Prevention' (n 15) 23.
[970] Ord, *Precipice* (n 16) 51.
[971] Bostrom, 'Existential Risk Prevention' (n 15) 23.

offer reason to avoid an x-catastrophe.[972] They propose that human extinction – even if everyone agree to bring it about – would wrong our forebears by undermining their sacrifices: "[O]ur forebears have taken costly steps towards improving humanity's lot and those currently alive have the power to further or to undermine those past efforts, and thereby make our forebears' sacrifices more or less worthwhile."[973]

Another more direct concern has to do with the present. An x-catastrophe will claim the lives of everyone presently alive. Measured in terms of lost lives a catastrophe of this class would be the worst incident in our history.[974] As written by Bostrom, if we believe that we have some moral reason to prevent a catastrophe that would claim the lives of a small number of people, and a stronger moral reason to prevent one that would claim the lives of a larger number, then we may speculate that we have an even stronger moral reason to prevent a catastrophe that would kill the entire human population.[975] Psychological factors aside and all else being equal, it is possible to accept that the badness of a catastrophe is linked to the number of lost lives. Many will likely agree that extinction is undesirable as it would kill many concrete humans.[976]

Whereas looking at the past and present offers a strong case for reducing x-risks through legal mechanisms, it is the destruction of humanity's future that makes an x-catastrophe uniquely bad and the stakes astronomically high.

5.1.1 Aims of Chapter 5

The stakes form an essential facet of the x-risk landscape. This Chapter will consider the astronomical stakes proposition (the 'ASP') and argue that this concept should underpin legal interventions for x-risk mitigation. It is useful to recall that the ASP encapsulates two interrelated principles. The first one is that x-risks have

[972] Patrick Kaczmarek and Simon Beard, 'Human Extinction and Our Obligations to the Past' [2020] Utilitas 199.
[973] ibid 201.
[974] Ord, *Precipice* (n 16) 43.
[975] Bostrom, 'Existential Risk Prevention' (n 15) 23-24.
[976] Stefan Schubert, Lucius Caviola and Nadira Faber, 'The Psychology of Existential Risk: Moral Judgments about Human Extinction' (2019) 9 *Scientific Reports* <https://doi.org/10.1038/s41598-019-50145-9> accessed 03 May 2023.

astronomically high stakes. The second is that existential catastrophes are uniquely bad relative to all non-existential catastrophes.

One of the self-imposed conditions for the LHC-inspired intervention is that it utilises the ASP. As such, it is necessary to justify the ASP and explore how this concept may underpin other legal interventions with the potential to mitigate x-risk. To provide a balanced discussion, Chapter 5 also considers why someone might want to reject the idea that the ASP should underpin legal mechanisms. This Chapter will address the following sub-question: What is the justification for the astronomical stakes proposition, and what is the relevance of this concept vis-à-vis legal interventions for existential risk mitigation?

5.2 RELATIVE BADNESS

This Part introduces the relative difference between an *existential* catastrophe and a *non-existential* catastrophe, and the unique badness of the former. Later in this Chapter, the thesis will justify that notion by exploring the potential duration and quality of humanity's future. As noted by Moynihan, research on existential risk is often motivated by optimism about humanity's collective future.[977]

A catastrophe destroying the future is arguably much worse—and much more so than what people may initiatively think—than a catastrophe which 'only' destroys most of humanity's present. This is so even if both catastrophes claim a similar number of lives. The former catastrophe, i.e., the *existential* catastrophe, can be viewed as uniquely bad.

As explained by Schubert, Caviola and Faber, "[w]hether extinction is uniquely bad or not depends on which of these considerations is the stronger: the immediate harm, or the long-term consequences."[978] As mentioned, the latter consideration is dominant for many researchers dealing with x-risks. There is a well-known passage from Derek Parfit that can serve to introduce the relative difference between an *existential*

[977] Moynihan, *X-Risk* (n 332) 21-22.
[978] Schubert, Caviola and Faber, 'The Psychology of Existential Risk' (n 976) 1.

catastrophe and a *non-existential* catastrophe, and the unique badness of the former. Many well-known scholars within the x-risk circle have referred to his passage,[979] which is repeated in the Box below:

> I believe that if we destroy mankind, as we now can, this outcome will be much worse than most people think. Compare three outcomes:
>
> (1) Peace.
> (2) A nuclear war that kills 99% of the world's existing population.
> (3) A nuclear war that kills 100%.
>
> (2) would be worse than (1), and (3) would be worse than (2). Which is the greater of these two differences? Most people believe that the greater difference is between (1) and (2). I believe that the difference between (2) and (3) is very much greater…The Earth will remain habitable for at least another billion years. Civilization began only a few thousand years ago. If we do not destroy mankind, these few thousand years may be only a tiny fraction of the whole of civilized human history. The difference between (2) and (3) may thus be the difference between this tiny fraction and all of the rest of this history. If we compare this possible history to a day, what has occurred so far is only a fraction of a second.[980]

The word *peace* can be replaced with 'no catastrophe' and the words *nuclear war* with 'catastrophe'. Doing so will place Parfit's passage within the x-risk landscape:

Scenario (1) – No Catastrophe.
Scenario (2) – A catastrophe that kills 99%.
Scenario (3) – A catastrophe that kills 100%.

It is presumed here that Scenario (3) amounts to an x-catastrophe. What makes this Scenario uniquely bad is that it destroys humanity's future. This makes the difference between Scenario (2) and (3) much greater than the difference between (1) and (2).

---

[979] Works on existential risk that, in one way or another, talk about Parfit's passage include: Ord, *Precipice* (n 16) 43; Bostrom, 'Existential Risk Prevention' (n 15) 17-18; Schubert, Caviola and Faber, 'The Psychology of Existential Risk' (n 976) 1-2; Farquhar and others, 'Existential Risk' (n 61) 11; Beckstead, 'On the Overwhelming Importance of Shaping the Far Future' (n 342) 56; Torres, *The End* (n 35) 29-30; Häggström, *Here Be Dragons* (n 22) 237-238; Martin Rees, *On the Future Prospects for Humanity* (Princeton University Press 2018) 116-117.
[980] Derek Parfit, *Reasons and Persons* (Oxford University Press 1984) 453-454.

Of course, for this conclusion to work, it must be assumed that the 1% who survived in Scenario (2) can, in time, recover and rebuild civilization. Regardless of the accuracy of this assumption,[981] the general point stands. The difference between an *existential* and a *near-existential* catastrophe is, on some views, extreme; it is much greater than what many people might believe at first.

In his passage, Parfit writes that most people believe that the greater difference is between Scenario (1) and (2). Is this true, and if so, why?

5.2.1 Relative Difference

According to some research, and in line with Parfit's expectation, many people think the greater difference is between Scenario (1) and (2). This is supported by Schubert, Caviola and Faber who examine "…the general public's views of the badness of human extinction."[982] This Section will summarise 'Study 2a' from their paper.

Having observed that people find human extinction bad, Study 2a explores whether people find extinction *uniquely* bad relative to a non-extinction catastrophe. They asked 1,251 participants – who were further divided into one control condition and four experimental conditions – to rank three outcomes from best to worst. The participants who chose the expected ranking were then asked to consider the relative difference (in terms of badness) between the outcomes. In the control condition, 257 participants were asked to rank these outcomes:

(A) There is no catastrophe.
(B) There is a catastrophe that immediately kills 80% of the world's population.
(C) There is a catastrophe that immediately kills 100% of the world's population.[983]

[981] A catastrophe killing 99% of everyone may leave the survivors unable to recover. If so, the catastrophe is existential.
[982] Schubert, Caviola and Faber, 'The Psychology of Existential Risk' (n 976).
[983] The authors made a few deliberate changes from Parfit's experiment. They "…said that 80%, rather than 99%, die in the non-extinction catastrophe, to make it more plausible that humanity could recover." ibid 6.

A majority in the control condition (~82%, 213/257 participants) ranked the outcomes as Parfit expected. That is, (A) as the best outcome and (C) as the worst. This was the 'expected ranking'. As one might have guessed, this ranking was only a preliminary step towards the main aim of the study.

The authors were mainly interested in which difference those who gave the 'expected ranking' found to be greater. Was it greater between (A) and (B), meaning that extinction is *not* uniquely bad? Or was it greater between (B) and (C), meaning that extinction *is* uniquely bad? A minority (~23%, 50/213 participants) judged the difference between (B) and (C) to be greater. Thus, most who gave the expected ranking did not find extinction uniquely bad. The authors used four experimental conditions to explain the results in the control condition. The following can be noted:

1. Schubert, Caviola and Faber theorised that people do not find extinction uniquely bad because they felt strongly for the victims. In other words, people focus more on the immediate consequences of the catastrophes. To test this theory, they used two experimental conditions (including the *animals condition*; 246 participants) to trigger a lesser focus on the immediate harm. The animals condition was largely identical to the control condition. The exception was that the catastrophes in (B) and (C) were described as killing 80% and 100% of the world's *zebra* population. The authors believed that people would focus less on the immediate harm if the catastrophes would affect zebras. A majority chose the expected ranking (~89%, 221/246 participants). In turn, a larger proportion of participants in the animals condition (~44%, 98/221 participants) judged extinction uniquely bad when compared with the control condition.

2. Another theory was that the participants neglected the relevant long-term consequences. The authors included the *salience condition* (248 participants) to test this theory. This condition was largely identical to the control condition. The exception was that the participants were directed to think about the long-term consequences. They were informed that humanity could have a future provided that some humans survived. They were then informed that there would be no future if everyone died. A majority opted for the expected ranking (~77%, 193/248 participants). In turn, a larger proportion in the salience

condition (~50%, 97/193 participants) found extinction uniquely bad when compared with the control condition.

3. The authors also included the *utopia condition* (248 participants) to test whether the quality of the future was relevant for the participants. The authors painted a very bright picture of the future and made clear what would be lost if humanity were to go extinct. Most participants here selected the expected ranking (~86%, 215/248 participants). In turn, a much larger proportion in the utopia condition (**~76%, 165/215** participants) found extinction uniquely bad when compared with the control condition.

Schubert, Caviola and Faber's study is valuable for several reasons. Firstly, an overarching finding is that laypeople do not view extinction as uniquely bad when asked without further qualifications.[984] A reason for this is that they focus on the immediate consequences of the catastrophe.[985] Secondly, inviting people to consider the long-term impacts makes a difference in how people judge extinction.[986] Thirdly, beliefs about the quality of the future are relevant because hints that the future may be very good can contribute to the view that extinction is uniquely bad.[987] The second and third point suggests that it is valuable to think more deeply about what the future may hold in store for humanity. In particular, it is from such reflections that one can extract the astronomical stakes proposition.

5.3 FUTURES

This Part will build upon the considerations above by exploring the potential duration and quality of humanity's future. As mentioned in the preceding Section, it is from deeper reflections on humanity's future that one can extract the ASP.

A preliminary observation merit attention. The focus below is on the many happy future *humans* (understood as inclusive of lifeforms within the definition of *humanity* in

[984] See the control condition. ibid 5.
[985] See the animals condition and the salience condition. ibid.
[986] See the salience condition. ibid.
[987] See the utopia condition. ibid.

Chapter 2) whose existence would be foreclosed by an existential catastrophe. It might be remembered that the term humanity is taken to mean 'Earth-originating intelligent life', and that the definition covers our intelligent descendants, whether they are born on Earth or beyond. Thus, if space colonisation becomes a reality, those born amongst the stars will fall within the definition of humanity.

Nevertheless, the focus on the foreclosing of happy future humans is not the only way to highlight the astronomically high stakes of x-risks. An existential catastrophe will also destroy humanity's *potential*.[988] In the words of Ord, "[t]here is so much that we could be and do, such a variety of flourishing and achievement ahead, that most conceptions of value will find something to mourn should we fail, should we squander this potential."[989] While it may sound naïve to some, humanity has, according to some scholars, the potential to create a future that is grandiose. Sandberg is currently working on a book on "Grand Futures" which reportedly considers, inter alia, the limits of what humanity *could* achieve.[990] He writes that "[p]art of this is an attempt to outline what we know is and isn't physically possible to achieve…"[991] From pushing the boundaries of human well-being to constructing Dyson spheres and moving entire galaxies, the alleged possibilities are wild. His talk – *Grand Futures: How Much is there to Hope For, How Much is at Stake?*[992] explores some of these possibilities, and it can reinforce the conception that humanity has barely scratched the surface when it comes to fulfilling its potential.

5.3.1 Duration & Future Lives

The expected value of humanity's future will depend on, among other things, its duration.[993] The length of time which can remain open to humanity is significant for different reasons. It feeds into what humanity may come to achieve in the far future. It

[988] Ord, *Precipice* (n 16); Moynihan, *X-Risk* (n 332).
[989] Ord, *Precipice* (n 16) 305.
[990] See e.g., Anders Sandberg, 'What kinds of grand futures are there?' (*Andart*, 5 July 2018) <http://aleph.se/andart2/space/what-kinds-of-grand-futures-are-there/> accessed 03 May 2023.
[991] ibid.
[992] Anders Sandberg 'Grand Futures: How Much is there to Hope For, How Much is at Stake?' (The Cambridge Conference on Catastrophic Risk 2020, The Centre for the Study of Existential Risk, 9 December 2020) <https://www.youtube.com/watch?v=9a5uYOoFJvE&t=6s> accessed 03 May 2023.
[993] Ord, *Precipice* (n 16) 411.

may also mean that most 'people' are future people. Thus, if it is accepted that that *future* generations matter and that future lives hold some value, then one will probably be more likely to accept that a catastrophe that destroys the future is uniquely bad.

(i) Finite Existence

While the duration is important, it is to be appreciated that humanity's future is limited by cosmic default. "[I]t is unlikely that civilization can persist in the cosmos indefinitely."[994] For instance, Adams considers the fate of the Earth, galaxy, and universe with reference to various astrophysical processes that are waiting to unfold in the far-future.[995] Whereas he starts by discussing how the Sun will affect our planet over the coming billions of years, his Chapter progresses much farther into the future. At one point, he writes that even the largest black holes will have evaporated when the cosmic age exceeds 10^{100} years.[996] Still, humanity's future may prove much shorter. Carbon-based life cannot survive the catastrophe induced by proton decay (perhaps ~10^{40} years from now).[997] In other words, "…the universe continues to exist, and astrophysical processes continue beyond [the] end of known biology."[998] Ćirković has also offered an overview of various resources dealing with inter alia the future of the Earth, solar system and universe.[999] These topics pertain to physical eschatology which is a branch of astrophysics that deals with the fate of astrophysical objects and the universe itself.[1000]

Similarly, Torres examines the 'really big picture' and writes that there are many "…natural events associated with the aging of our…sun, and universe that will pose a

[994] Baum and others, 'Long-term Trajectories' (n 118) 72.
[995] Fred Adams, 'Long-term Astrophysical Processes' in Nick Bostrom and Milan Ćirković (eds), *Global Catastrophic Risks* (OUP 2008) 33-47.
[996] ibid 41.
[997] ibid.
[998] ibid.
[999] Milan Ćirković, 'Resource Letter: PEs-1: Physical Eschatology' [2002] American Journal of Physics 122.
[1000] ibid; Milan Ćirković, 'Forecast for the Next Eon: Applied Cosmology and the Long-Term Fate of Intelligent Beings' Foundations of Physics [2004] 239, 240.

succession of existential risks."[1001] Whereas Torres (and others)[1002] highlights that space colonisation is a key strategy for humanity's future survival,[1003] he recognises that this strategy will not insulate us from the risk posed by the universe itself.[1004] From the above it is clear that the ultimate end is not that of humanity, but rather that of the universe. In her book, *The End of Everything (Astrophysically Speaking)*, Mack writes:

> The death of a planet…might in principle be survivable. In billions of years, humanity could still conceivably exist, in some perhaps unrecognizable form, venturing out to distant reaches of space, finding new homes and building new civilizations. The death of the universe, though, is final.[1005]

She writes that, "[a]t some point, in a cosmic sense, it will not have mattered that we ever lived."[1006] Her book considers five possible – albeit not equally credible – cosmic ends: The Big Crunch, Heat Death, Vacuum Decay, the Big Rip and the Bounce.[1007] She explains that each scenario presents a different style of apocalypse and is governed by a different physical process.[1008] Of course, and as mentioned, humanity's end may come long before the death of the universe.

The takeaway is that various long-term physical processes mean that humanity's future, on this planet and beyond, is probably (and save for some very speculative strategies)[1009] one of finite duration. In effect, "…extinction is not a question of

---

[1001] Torres, *The End* (n 35) 152.

[1002] Hawking, *Brief Answers to the Big Questions* (n 238). On page 150 he writes: "…[I]n the long run the human race shouldn't have all its eggs in one basket, or on one planet. I just hope we can avoid dropping the basket before we learn how to escape from Earth." Continuing on page 151, he writes that he is convinced that we must leave Earth; Anders Sandberg, Jason Matheny and Milan Ćirković, 'How can we Reduce the Risk of Human Extinction?' *Bulletin of the Atomic Scientists* (9 September 2008) <https://thebulletin.org/2008/09/how-can-we-reduce-the-risk-of-human-extinction/> accessed 03 May 2023; Baum, 'The Great Downside Dilemma' (n 22); Ord, *Precipice* (n 16) 392-393.

[1003] Torres, *The End* (n 35) 230-231; In contrast, see Phil Torres, 'Space Colonization and Suffering Risks: Reassessing the "Maxipok Rule"' [2018] *Futures* 74.

[1004] Torres, *The End* (n 35) 157, 247.

[1005] Katie Mack, *The End of Everything: (Astrophysically Speaking)* (Allen Lane 2020) 2.

[1006] ibid 206.

[1007] Mack deals with the evidence for and against the scenarios. For example, on page 70 she writes that the current evidence suggests that "…we are almost certainly safe from a fiery death in a Big Crunch…"

[1008] ibid 12.

[1009] Torres, *The End* (n 35) 157.

whether, but when." [1010] With the above in mind, the upcoming Sub-section will speculate how long humanity could survive.

(ii) On Earth and Beyond

Beckstead contends that there is a non-negligible probability that humanity will (i) survive on Earth until it becomes uninhabitable as well as (ii) colonise space. It is the latter prospect that will arguably unlock the "…lion's share of the expected duration of our existence."[1011] In effect, humanity may survive for billions or even trillions of years, which makes the expected value of the future astronomically great.[1012]

Beckstead is talking about the *subjective* probability in this context. To be conservative, Beckstead assumes the probability for claim (i) above should be 1%. Recognising the underlying uncertainties, he explains: "Having a very high or low probability in this claim…would require much greater certainty about the future than it is reasonable to have."[1013] Regarding claim (ii), Beckstead proposes to assign a probability greater than 1/100, *conditional* on surviving for a billion years. Thus, the *unconditional* probability of this event is 1/10,000:

$$\frac{1}{100} \times \frac{1}{100} = \frac{1}{10{,}000}$$

Relying on estimates for how long some stars will burn (~100 trillion years), Beckstead writes that humanity may survive for the full 100 trillion years if humanity colonises space. In light of the above, he calculates and suggests that the expected years of civilisation ahead of us is 10 billion years. This is because:

$$\frac{1}{10{,}000} \times 100 \text{ trillion years} = 10 \text{ billion years.}$$[1014]

[1010] Bostrom and Ćirković, 'Introduction' (n 12) 8.
[1011] Beckstead, 'On the Overwhelming Importance of Shaping the Far Future' (n 342) 57.
[1012] ibid 54.
[1013] ibid.
[1014] ibid 57-58.

Beckstead's claims that (i) humanity can survive until the Earth becomes uninhabitable and (ii) colonise space are not unreasonable. As a baseline, presume – as others have done before – that the Earth can remain habitable for complex life for another ~1 billion years. This presumption is not the same as guaranteeing that humanity will survive for as long as Earth remains habitable. For example, it is possible to find statements pointing out that over 99% of all species to have ever existed are now extinct.[1015] Is this indicative of humanity's fate? Not necessarily. To begin with, it should be remembered that the end of our species is not the same as the end of humanity. As written by Beckstead: "When we include our intelligent descendants, it is not absurd to consider the possibility that civilization continues for a billion years, until the Earth becomes uninhabitable."[1016] Ord also notes that many species are succeeded by their siblings or children on the evolutionary tree.[1017]

Furthermore, some features can increase humanity's survivability compared to other species. Examples include humanity's adaptability, intelligence, numbers, global presence and ability to think about x-risks and how to prevent them from materalising. On the other hand, the odds can also be worsened because humanity keeps on adding to and neglecting the catalogue of x-risks. It is difficult to predict with confidence how humanity, civilization, technology and so on will transform over the next aeon. However, the topic of studying the path that human civilisation may take into the long-term future is not beyond legitimate analysis.[1018] The thesis will not further consider the odds of future survival and the factors that can sway these odds. Instead, it will settle with the general observation advanced by Beckstead (and others) that the likelihood of survival depends on the decisions that society makes.[1019]

What about the claim that humanity may colonise space? It is true that the undertaking to expand humanity's home beyond Earth will face many technical and other

[1015] Sandberg, Matheny and Ćirković, 'How can we Reduce the Risk of Human Extinction?' (n 1002); Snyder-Beattie, Ord and Bonsall, 'An Upper Bound for the Background Rate of Human Extinction' (n 167).
[1016] Beckstead, 'On the Overwhelming Importance of Shaping the Far Future' (n 342) 56.
[1017] Ord, *Precipice* (n 16) 220.
[1018] Baum and others, 'Long-term Trajectories' (n 118).
[1019] Beckstead, 'On the Overwhelming Importance of Shaping the Far Future' (n 342) 56; Baum and others, 'Long-term Trajectories' (n 118) 56. Häggström, *Here Be Dragons* (n 22) 245; Ćirković, 'Forecast for the Next Eon' (n 1000) 256.

hurdles.[1020] Beckstead accepts that there are many obstacles to space colonisation,[1021] but also notes that these are surmountable. He writes: "People who have done the most in-depth work on the feasibility of space colonisation generally believe it is possible." He further explains why he does not find claims about its infeasibility to be very convincing.[1022] His paper draws on interviews with people, including Anders Sandberg (a believer in colonisation) and Charles Stross (the most credible sceptic Beckstead could find). Notably, Sandberg is the co-author of an article in *Acta Astronautica*, which argues that colonisation is feasible.[1023] It even suggests that "…intergalactic colonisation is not far beyond our current capabilities today."[1024] One can also refer to Ord's book for a relevant (and less technical) discussion.[1025] He explains that our society do not currently have the technology to reach another star, stop and then build a civilization. Still, he states that there are seemingly no "fundamental barriers" that would make such an endeavour impossible in the future. In his view, "…the biggest challenge will be surviving on Earth for the century or two until it becomes technologically feasible."[1026]

In summary, space colonisation can greatly increase the expected duration of humanity's future. To make an astronomical difference, humanity must venture beyond the solar system. One cannot deny that there are many obstacles affecting this pursuit. Some of these challenges are technological, and others may be political or economic. In addition, humanity must also survive on the Earth until surviving beyond the Earth becomes a realistic option. Nevertheless, with enough time, the

[1020] See Hawking, *Brief Answers to the Big Questions* (n 238) ch 8; Nick Beckstead, 'Will we Eventually be Able to Colonize Other Stars? Notes From a Preliminary Review' (Future of Humanity Institute, 22 June 2014) <https://www.fhi.ox.ac.uk/will-we-eventually-be-able-to-colonize-other-stars-notes-from-a-preliminary-review/> accessed 03 May 2023.

[1021] A few of the obstacles include large energy requirements, health and reproductive challenges, short human lifespans in comparison with great distances for interstellar travel, finding a hospitable location and building another civilization.

[1022] Beckstead notes that these claims – many contained in journalistic articles and blog posts – often fail to properly engage with the applicable counterarguments. He also explains that he found no books or scientific papers which makes the case that space colonisation is in-principle infeasible. Beckstead, 'Will we Eventually be Able to Colonize Other Stars?' (n 1020).

[1023] Stuart Armstrong and Anders Sandberg, 'Eternity in Six Hours: Intergalactic Spreading of Intelligent Life and Sharpening the Fermi Paradox' [2013] 1.

[1024] ibid 2.

[1025] Ord, *Precipice* (n 16) 228-230.

[1026] ibid 229.

endeavour is said to be feasible. Thus, and to summarise it is not unreasonable to make the case that there is indeed a *non-negligible probability* that it will happen.

(iii) Future Lives

A corollary from above is that humanity may survive for billions or trillions of years to come. The potential duration of the future is significant. It means that humanity's overall 'story' may have just begun in the grand scheme of things. In effect, the vast majority of humans who will ever exist are future humans. In other words: "Most generations are future generations."[1027] This recognition matters if it is accepted that a future lives matter. [1028]

This observation intertwines with the following question: How many lives may exist in the future if one accepts certain assumptions aligning with the previous discussion on the potential length of the future? Bostrom estimates that there is potential for 10^{16} human lives if assuming (i) that the Earth will remain habitable for another billion years and (ii) that we enter a mode of sustainable existence with a population of one billion people, which has a lifespan of 100 years.[1029] He also suggests that these lives can be much better than the average life of today.

Bostrom's estimate (i.e., the potential for 10^{16} human lives) is conservative. This is because it "…ignores the possibility of space colonisation and software minds…"[1030] The estimate becomes much less conservative if these are taken into account. Bostrom estimate that "…the number of biological human life-years in the future accessible universe…is 10^{34} years."[1031] This estimate ignores the possibility of software minds. The thesis will not explore the possibility of transferring human minds to computer hardware – i.e., uploading.[1032] All to be appreciated here is that if

[1027] ibid 44.
[1028] Simon Beard and Patrick Kaczmarek, 'On the Wrongness of Human Extinction' [2019] Argumenta 85, 85-86
[1029] Bostrom, 'Existential Risk Prevention' (n 15) 18.
[1030] ibid.
[1031] ibid.
[1032] Häggström, *Here Be Dragons* (n 22). On page 66, he calls uploading the "ultimate life extension technology" On pages 66-81, he explores philosophical and technological issues when it comes to uploading.

uploading is considered, the estimate grows greater still. As written by Bostrom in his paper on the astronomical waste argument, "[w]hat matters…is not the exact numbers but the fact that they are huge."[1033] This is the key point in the present context too. For the present purpose, the huge numbers are simply used to support the notion that x-risks have astronomical stakes. More generally, they can also support the interrelated argument that x-risk *reduction* is extremely valuable and should be made a global priority.[1034]

5.3.2 Duration & the Quality of the Future

The *quality* of the future is an important component of the ASP. As noted, Schubert, Caviola and Faber's paper suggests that more people will view extinction as uniquely bad when they are informed that the future may be extremely good. One interpretation is that people will reflect more profoundly on the long-term consequences when considering the potential quality of the future. Their finding can also be used to support the broader point that the quality of the future matters for its expected value. People may feel a stronger urge to prevent a catastrophe which would foreclosie the existence of quadrillions of *happy* people. This recognition comes with a significant caveat. It can inversely be suggested that the present generation has less incentive to prevent extinction if it is anticipated that the future will contain more suffering than happiness. Non-existence might even be the better option if the future will contain astronomical amounts of suffering. If so, then the mitigation of "s-risks" – risks of astronomical suffering – may be even more important that the mitigation of x-risks.[1035]

However, there is still a non-negligible probability that the future may turn out good. Bostrom, amongst other x-risk scholars, claim that human lives can be "…considerably better than the average contemporary human life…"[1036] The next Sub-section entertains this claim.

[1033] Bostrom, 'Astronomical Waste' (n 17) 309.
[1034] Bostrom, 'Existential Risk Prevention' (n 15) 18-19.
[1035] Althaus and Gloor, 'Reducing Risks of Astronomical Suffering' (n 146); Daniel, 'S-risks' (n 146).
[1036] Bostrom, 'Existential Risk Prevention' (n 15) 18.

(i) Better Lives?

This Section considers Bostrom and other scholars' claims that future lives can be considerably better than the average contemporary human life. The non-negligible probability that such claims are true is significant if one accepts that there exists that policymakers have more reason to protect a 'good' future than a 'bad' one. The idea of what constitutes a good future is contentious. Yet, one reasonable benchmark is the quality of the average human life. Bostrom himself proposes that future lives can be better since technological and moral progress can help humanity to partially overcome disease, poverty, injustice and various biological limitations.[1037]

At this point, it is possible to wonder if the premise of Bostrom's claim is too optimistic. After all, most people reportedly think that the world is, all things considered, getting *worse* as opposed to *better*.[1038] If this is accurate, without any exceptions, Bostrom's claim becomes less persuasive. It is more difficult to build a case that the future can be better than today if things only worsen over time. But this is not necessarily the case, and people from all walks of life are often wrong about significant global trends, with a common theme being that many incorrectly think that things are not improving or are only getting worse.[1039]

If many believe that the world is a hopeless mess, they may, in turn, believe that worrying about the future is a misplaced ambition. In fact, such beliefs can probably

---

[1037] ibid.

[1038] Will Dahlgreen, 'Inside the Mind of an Optimist', *YouGov* (13 May 2015) <https://yougov.co.uk/topics/politics/articles-reports/2015/05/13/optimists-about-world> accessed 03 May 2023; Will Dahlgreen, 'Chinese People are Most Likely to Feel the World is Getting Better' *YouGov* (05 Jan 2016) <https://yougov.co.uk/topics/lifestyle/articles-reports/2016/01/05/chinese-people-are-most-optimistic-world> accessed 03 May 2023; Max Roser, 'Most of us are Wrong About how the World has Changed (Especially Those who are Pessimistic About the Future)' *Our World in Data* (27 July 2018) <https://ourworldindata.org/wrong-about-the-world> accessed 03 May 2023; Pete Etchells, 'Declinism: Is the World Actually Getting Worse?' *The Guardian* (16 January 2015) <https://www.theguardian.com/science/head-quarters/2015/jan/16/declinism-is-the-world-actually-getting-worse> accessed 03 May 2023.

[1039] Hans Rosling, Ola Rosling, and Anna Rosling Rönnlund, *Factfulness: Ten Reasons We're Wrong About the World – And Why Things Are Better Than You Think* (Sceptre 2018)*;* See also the Gapminder Foundation <https://www.gapminder.org/> accessed 03 May 2023; Hans Rosling and Ola Rosling, 'How not to be Ignorant About the World' (*TED Talk*, 11 September 2014) <https://www.youtube.com/watch?v=Sm5xF-UYgdg&t=216s> accessed 03 May 2023; Roser, 'Most of us are Wrong About how the World has Changed' (n 1038).

extinguish a person optimism about the future. As written by Roser, "[o]ur perception of how the world is changing matters for what we believe is possible in the future." [1040] In light of a survey asking participants whether people in the future will be better or worse off, Roser notes that there is a connection between peoples' perception of the past and their hope for the future. He explains that the degree of optimism about the future correlates with how much people know about global development.

Those with the least knowledge of how the world has changed were the most pessimistic about the future.[1041] In contrast, "[t]hose who are optimistic about the future can base their view on the knowledge that it is possible to change the world for the better, because they know that we did."[1042]

Various publications on *Our World in Data* provide reliable data for numerous global problems.[1043] At the time of writing this sentence, the website contains 3,407 charts across 297 topics.[1044] Various developments indicate that the world is changing for the better in *many* ways. For example, in one publication, Roser looks at (inter alia) extreme poverty, literacy, health and freedom, and suggests that global living conditions have improved over the last two centuries.[1045]

Whether looking at *Our World in Data* or elsewhere online, it is possible to find data supporting the argument that the world is, in many significant ways, getting better and that humanity has made remarkable progress over the last centuries.[1046] The purpose of this Section is not to summarise the statistics and data demonstrating this point. Instead, the takeaway is simply that humanity's progress to date can offer credence to the argument that there is a non-negligible probability that future human lives can be "…considerably better than the average contemporary human life…"[1047]

[1040] ibid.
[1041] ibid.
[1042] Max Roser 'The Short History of Global Living Conditions and Why it Matters That we Know it' *Our World in Data* (2020) <https://ourworldindata.org/a-history-of-global-living-conditions-in-5-charts> accessed 03 May 2023.
[1043] Our World in Data <https://ourworldindata.org/> accessed 03 May 2023.
[1044] ibid.
[1045] Roser 'The Short History of Global Living Conditions and Why it Matters That we Know it' (n 1042).
[1046] Steven Pinker, 'Is the World Getting Better or Worse? A Look at the Numbers' (*TED Talk*, 21 May 2018) <https://www.youtube.com/watch?v=yCm9Ng0bbEQ> accessed 03 May 2023.
[1047] Bostrom, 'Existential Risk Prevention' (n 15)18; Moynihan, *X-Risk* (n 332) 22.

It is appropriate to end this Sub-section with two qualifications. Firstly, the world is not only improving. This is also emphasised by Ord. While suggesting that the trends toward progress are clear, he writes that there is "…a danger of choosing selectively from history to create a simple narrative of improvement from a barbarous past to a glorious present."[1048] On a related note, some scholars argue that those demonstrating the positive trends have avoided analysis of negative trends. For instance, Berggren writes that one famous book in this regard—*Factfulness*— "…includes many graphs of 'bad things in decline' and 'good things on the rise' but not a single graph of problematic phenomena that are on the rise."[1049] Secondly, it is not possible to conclude that trends of progress will continue indefinitely.

5.4 ASTRONOMICAL STAKES PROPOSITION

The sections above have explored why, one some views, the destruction of the future makes an existential catastrophe uniquely bad. They have also sought to defend the position that x-risks have astronomically high stakes.[1050]

Building on these recognitions, this thesis proposes a concept which can be referred to as the 'astronomical stakes proposition'. The essence of the ASP is repeated here for convenience: The concept encapsulates the ideas that (i) x-risks have astronomically high stakes, and that (ii) x-catastrophes are uniquely bad relative to *all* non-existential catastrophes.[1051] It is possible to appreciate that point (i) and (ii) are interconnected. X-risks have astronomically high stakes because they threaten to

[1048] Ord, *Precipice* (n 16) 18.
[1049] Christian Berggren, 'Good Things on the Rise: The One-Sided Worldview of Hans Rosling' *Kvartal* (20 September 2018) <https://kvartal.se/artiklar/bra-saker-pa-uppgang-roslings-varldsbild-ar-ensidigt-positiv/> accessed 03 May 2023. For a translation, see <https://www.researchgate.net/publication/328759928_Good_Things_on_the_Rise_The_One-Sided_Worldview_of_Hans_Rosling_Translation_of_an_essay_published_in_Kvartal_Sept_20_2018_httpskvartalseartiklarbra-saker-pa-uppgang-roslings-varldsbild-ar-ensidigt-pos> accessed 03 May 2023.
[1050] "To calculate the loss associated with an existential catastrophe, we must consider how much value would come to exist in its absence. It turns out that the ultimate potential for Earth-originating intelligent life is literally astronomical." See Bostrom, 'Existential Risk Prevention' (n 15) 18.
[1051] For this thesis, it is not necessary to determine whether x-catastrophes are uniquely bad relative to catastrophes that, while not bringing about extinction, bring about suffering on an astronomical scale.

destroy the future, and it is the destruction of the future that would make such a catastrophe uniquely bad relative to catastrophes that do not destroy the future.

It is suggested that the ASP can and should, in some cases, empower and underpin legal mechanisms seeking to mitigate x-risk. The underlying reason for why the ASP should influence legal interventions for x-risk mitigation revolves around the simple view that appropriate precautions and measures should correspond to the risk in question.[1052] This entails a consideration of the stakes (and likelihood) of the risk involved. For instance, Voeneky writes that "…a decisive element of a governance regime should be that the relevant actors…assess and reduce existential and global catastrophic risks in a way that mirrors the actual probability and severity of a certain risk."[1053] She continues by suggesting that the "…higher the severity of the possible damage, the lower the demands on how probable a risk must be."[1054] This is one example of how the ASP can influence a proposed governance regime for the mitigation of existential risk. Although, at this stage it can be noted that the ASP need not underpin all interventions in the same way or to the same extent.

To summarise, the ASP should underpin interventions that may safeguard humanity's future. These interventions can be broad or targeted in nature. In turn, a broad or targeted intervention can be legal or non-legal.

Part 5.5 below deals with broad and targeted intervention (legal and non-legal) in further detail. It is to be remembered that the intervention advanced in response to the LHC Controversy is a *broad legal* mechanism which aims to reduce the probability that an x-risk materialises in the first place. Chapter 6 will consider how the ASP can support the LHC-inspired intervention.

[1052] However, and as noted by Voeneky "Results in behavioral science show that we do not have a rational perception of low probability/high risk scenarios: Our risk perception does not mirror the probability and severity of a certain risk, as very rare events are either ignored by us or we tend to overweight small risks." See, Voeneky, 'Human Rights and Legitimate Governance' (n 305) 151; See also Karin Kuhlemann, 'Complexity, Creeping Normalcy and Conceit: Sexy and Unsexy Risks' [2019] Foresight 35.

[1053] Voeneky, 'Human Rights and Legitimate Governance' (n 305) 151.

[1054] ibid 156.

5.5 BROAD & TARGETED INTERVENTIONS

Beckstead distinguishes between *broad* and *targeted* ways of attempting to shape the far future.[1055] Improving education is given as an example of a broad intervention. In contrast, trying to prevent an asteroid from hitting Earth is a targeted intervention. He recognises that there is a spectrum between the two. Teaching our children to be good stewards of the future is, according to Beckstead, an option falling somewhere in-between a broad and targeted way of trying to shape the far future.[1056]

The distinction between broad and targeted ways and interventions can be transposed to the field of x-risk mitigation. Of course, the distinction is not privy to the mitigation of this specific category of risk.[1057] In any event, the distinction is sometimes embraced by others.[1058] For example, Todd separates '*targeted efforts to reduce specific [existential] risks'* (e.g., better disease surveillance to reduce the risk of pandemics) and '*broad efforts to reduce [existential] risks*' (e.g., better disaster shelters which could reduce the chance of extinction from pandemics, nuclear winter and asteroids).[1059] Farquhar and others explore a similar theme and distinguish *cross-cutting* versus *risk-specific* interventions and *direct* versus *capacity-building* interventions.[1060]

5.5.1 Non-Legal Interventions and Strategies

Interventions that mitigate x-risks can be non-legal in nature. For example, Torres outlines various strategies for minimising the likelihood that humanity perishes to an x-catastrophe.[1061] Thinking about non-legal interventions, Torres considers options

[1055] Beckstead, 'On the Overwhelming Importance of Shaping the Far Future' (n 342).
[1056] ibid 10.
[1057] The distinction is also applicable in considering how to mitigate e.g., s-risks. Tobias Baumann, 'S-risks: An Introduction', (Center for Reducing Suffering, 2017) <https://centerforreducingsuffering.org/intro/> accessed 03 May 2023.
[1058] Phil Torres, 'Facing Disaster' (n 331) 5.
[1059] Benjamin Todd, 'The Case for Reducing Existential Risks', (80,000 Hours, October 2017) <https://80000hours.org/articles/existential-risks/> accessed 03 May 2023.
[1060] Farquhar and others explore the distinction between cross-cutting and risk-specific interventions. The distinction comes down to whether an intervention aims to address a specific risk or whether it helps to address multiple risks. See Farquhar and others, 'Existential Risk' (n 61) 17.
[1061] Torres, *The End* (n 35).

including (i) redesigning the human form, (ii) colonising the universe, (iii) burrowing into the Earth, and (iv) supporting organisations like the FHI and CSER.[1062] Those interested in mitigation strategies more generally can consider Chapter 6 of *Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks* and Chapter 7 of *The Precipice*.

Some of the interventions above, such as supporting relevant organisation like FHI, may involve little to no interaction with legal norms. Other strategies, such as redesigning the human form, will surely involve legal issues and questions.[1063] Space colonisation is another example of a 'non-legal strategy' that will probably involve questions of law. For one, international space law, a body of law said to suffer from legal uncertainty, is relevant to the colonisation of celestial bodies and exploitation of resources.[1064] There is no need to examine the details to appreciate the basic point that some non-legal interventions may intersect with questions of law.

5.5.2 Legal Interventions and Strategies

The distinction between legal and non-legal interventions is not perfect or absolute. Still, it is possible to appreciate that there is a general distinction. As with any strategy, a legal intervention can be (more or less) broad or targeted in nature.

Calling for the present generations to owe up to the astronomical stakes associated with x-risks may be taken as an example of a broad (legal) intervention. The gist of such an intervention can take many shapes and forms. For instance, it is possible to draw an analogy to the UNESCO *Declaration on the Responsibility of the Present Generations Towards Future Generations*.[1065] Without commenting on its effectiveness, the Declaration proposes that the present generations have

[1062] ibid 228-240
[1063] Woodrow Barfield and Alexander Williams, 'Law, Cyborgs, and Technologically Enhanced Brains' (2017) 2(1) Philosophies <https://doi.org/10.3390/philosophies2010006> accessed 03 May 2023.
[1064] Brandon Gruner, 'A New Hope for International Space Law: Incorporating Nineteenth Century First Possession Principles into the 1967 Space Treaty for the Colonization of Outer Space in the Twenty-First Century' [2004] Seton Hall Law Review 299.
[1065] Ord, *Precipice* (n 16) 204.

responsibilities towards the future generations. For instance, Article 4 reads: "The present generations have the responsibility to bequeath to future generations an Earth which will not one day be irreversibly damaged by human activity." The Preamble also reads, among other things, that "…the very existence of humankind and its environment are threatened…" and "…the fate of future generations depends to a great extent on decisions and actions taken today…"

It is easy to appreciate the idea of targeted (legal) interventions. For example, a proposed treaty aiming to regulate a specific technology associated with existential risk may count as a targeted (legal) intervention for x-risk mitigation. The word *may* is appropriate because many relevant technologies are also associated with less serious risks and other problems. Take artificial intelligence (AI) as an example. AI is linked to an array of legal issues that are very much unrelated to existential risk. Examples include intellectual property issues, legal personhood issues, and liability issues.[1066] Addressing any one of these specific issues may be seen as a targeted legal response. However, it is not a targeted intervention for the purpose of *x-risk mitigation*. Chapter 2 highlighted the distinction between *narrow* and *general* AI. Existential risk is mainly associated with future AGI. Thus, a regulatory framework for AGI is more likely to count as a targeted legal intervention for our purpose.

As another example, consider geoengineering. Recall that geoengineering covers different techniques and splits into two broad categories: Carbon Dioxide Removal (CDR) techniques and Solar Radiation Management (SRM) techniques. It has been recognised that any regulatory framework for geoengineering cannot be uniform.[1067] Still, some tailored frameworks, such as the regulation of *SRM research*, can be seen as a targeted legal intervention for existential risk mitigation.[1068]

[1066] Rowena Rodrigues, 'Legal and Human Rights Issues of AI: Gaps, Challenges and Vulnerabilities' (2020) 4 Journal of Responsible Technology <https://doi.org/10.1016/j.jrt.2020.100005> accessed 03 May 2023.
[1067] House of Commons Science and Technology Committee, 'The Regulation of Geoengineering' (Fifth Report of Session 2009–10) para 30.
[1068] Farquhar and others, 'Existential Risk' (n 61) 16-19.

(i) Examples of Legal Interventions

This Sub-section will build on the previous one and highlight some potential legal interventions that may mitigate x-risks. The interventions are crudely categorised into one of two tables: 'Table 5.1: Broad Legal Interventions' or 'Table 2: Targeted Legal Interventions'.

There are a few points to note before moving on. Firstly, this Section further contextualises the distinction between legal and non-legal interventions. It is also useful to highlight a few examples in the two categories. Secondly, in categorising a proposed intervention as broad or targeted, the approach is the same as outlined in Chapter 1. In other words, broad legal interventions are designed to reduce the likelihood of x-risk from two or more hazards. In contrast, targeted interventions are hazard-specific, aiming to reduce risk from one hazard. Thirdly, interventions can be more or less broad or targeted in relation to one another. This conclusion aligns with Beckstead's observation from above that there is a spectrum between broad and targeted ways of attempting to shape the far future. Thus, although intervention (A) and intervention (B) may classify as broad, the former may be broader than the latter or vice versa. It is not necessary to focus on this matter in detail. Nonetheless, a few comments concerning this inter se distinction are made after Table 5.1. Fourthly, some proposals below do not explicitly focus on existential risk but can help to mitigate such risks in effect. Fifthly, the proposals below should not be taken as an indication that this thesis endorse them. Finally, one can note the distinction between interventions that revolves around *prevention*, *response* and *resilience*.[1069] All interventions below largely belong to the 'prevention' camp. That is, they can be seen as trying to reduce the likelihood that a catastrophe strikes in the first place.

[1069] Cotton-Barratt, Daniel and Sandberg, 'Defence' (n 37).

Table 5.1

Broad Legal Interventions	
Author	**Proposal**
1. Ord	Highlights the option of "…a kind of constitution for humanity, and writing into it the paramount need to safeguard our future, along with the funding and enforcement mechanisms required."[1070]
2. Voeneky	Argues that "…a legitimate governance regime of existential and global catastrophic risks [from scientific and technological progress] should be based on human rights…" This is because, among other things, human rights include relevant values (e.g., the freedom of science and the right to life), which need to be considered when assessing risk scenarios. Additionally, human rights can be interpreted to oblige states to evaluate and reduce these risks in a way that reflects a risk's probability and severity. In effect, human rights can provide a *starting point* and offer *general guidance* on how to govern GCRs and ERs.[1071]
3. Wilson	Suggests that the international community should create a binding yet flexible treaty that regulates GCRs and ERs from emerging technologies. Wilson focuses on nanotechnology, bioengineering, and AI. However, he suggests that "…an international treaty could potentially cover all emerging technologies that pose a GCR/ER, beginning with the three in [his] paper…" He proposes the framework of a model treaty with regulatory mechanisms, including the use of the precautionary principle, a body of experts, and binding reforms for scientists.[1072]
4. McKinnon	Proposes a new international criminal offence ('postericide') which is committed by *intentional* or *reckless* conduct that is fit to bring about a state of near human extinction. She highlights the badness of a state of near extinction on account of the damage that would be done to victims in such a state. The damage to which she refers is the damage to human security. In light of this, McKinnon suggests that it is legitimate to criminalise conduct that creates or exacerbates an extinction mechanism making total human extinction more likely. Among other things, McKinnon outlines the actus reus and mens rea for postericide.[1073]
5. Johnson	Suggests that a generalist judge could employ a "meta-analysis" to deal with preliminary injunction requests in cases such as the *Sancho case*. He argues that the meta-analysis can be used to circumvent many of the challenges that arose in the *Sancho case*. In essence, the court should analyse the potential that the relevant safety assessment is based on flawed or incomplete scientific theory, the potential for underlying mistakes in the assessment, the potential that various psychological factors, as well as a conflict of interest amongst the risk assessors, could undermine the quality of safety pronouncements. Notably, the application of the meta-analysis is not contingent upon the court having scientific or subject-matter expertise of, for example, theoretical particle physics. If, having utilised the meta-analysis, the court believes that there are "serious questions" (e.g., that the risk

[1070] Ord, *Precipice* (n 16) 202.
[1071] Voeneky, 'Human Rights and Legitimate Governance' (n 305).
[1072] Wilson, 'Minimizing' (n 151).
[1073] McKinnon, 'Endangering Humanity' (n 180).

	assessment might be flawed), the court should issue a preliminary injunction.[1074]
6. Peterson	Focuses on cases where scientific theories are the subject of litigation – such as the ones involving the LHC – and proposes that creating a new science court can improve the adjudicative process in cases involving complex scientific claims. While exploring different mechanisms that could improve the ability of courts to address complex scientific claims, Peterson ultimately argues that "…the most significant and comprehensive improvements would be achieved through the creation of a specialized science court."[1075]
7. Villmer	Suggesting that the National Environmental Policy Act (NEPA) has lost its impact, Villmer argues that NEPA requires substantive modification to be brought back to its former status as an environmental safeguard. He argues that modification is particularly acute given that we are in an era where scientific and governmental experimentation may bring about complete global destruction. First, Villmer argues that the US Congress *must* create a specialised NEPA court to deal with NEPA violation cases in a timely fashion. Second, he argues that the US Congress must amend NEPA so that preliminary injunctions become mandatory in circumstances when it is reasonably believed that human life is at risk.[1076]

Some of the interventions in Table 5.1 are broader than others. For instance, it is reasonable to think that Proposals (1) and (2) are wider than Proposal (7). This is because the latter concerns NEPA litigation in the US whereas the former proposals are more flexible and have an international dimension. Likewise, Proposal (3) is possibly wider than (4) since the latter – in contrast to the former – revolves around *intentional* or *reckless* conducts. The same principle applies to the LHC-inspired intervention. It is not intended to deal with all hazards within the x-risk landscape. Furthermore, it is not intended to take a spot as the 'broadest' legal intervention for x-risk mitigation.

The list above is not exhaustive, and there are other proposed legal interventions that could be allocated to Table 5.1. Examples include Posner's proposal on a federal catastrophic-risk assessment board,[1077] Verdirame's suggested international treaty on risks to humanity's future,[1078] Boyd and Wilson's line of thinking that the UN may play

[1074] Johnson, 'The Black Hole Case' (n 68); See also Johnson, 'Uncertain Risk, Science Experiments, and the Courts' (n 68).
[1075] Peterson, 'The "Sancho" Effect' (n 468).
[1076] Villmer, 'Procedural Squabbling' (n 554).
[1077] Posner, *Catastrophe* (n 90) 221.
[1078] Guglielmo Verdirame, 'For China, a Legal Reckoning is Coming: We Need a new International Treaty to Deal with Risks to the Future of Humanity' *UnHerd* (21 April 2020) <https://unherd.com/2020/04/for-china-a-legal-reckoning-is-coming/> accessed 03 May 2023.

a more prominent role in x-risk mitigation,[1079] and Kohler's existential risk sanction clause.[1080] The next Sub-section will consider a few targeted legal interventions.

Table 5.2

Targeted Legal Interventions	
Author	**Proposal**
8. Nindler	With a focus on AGI, Nindler explores (i) the UN's capability to manage existential risk and (ii) some changes to international law that may strengthen this capability. For instance, Nindler explores the option of an international treaty regulating AI R&D. He admits that it is difficult to outline the precise contents of a hypothetical treaty. Still, he accepts that there are general components that can probably be regarded as useful. For instance, the treaty will need to contain an obligation of the state parties to conduct AI R&D safely and refrain from the proliferation of knowledge that may contribute to the development of unsafe AI. Recognising that verification and monitoring is a crucial (yet challenging) aspect of such a treaty, he explains that it may be desirable to establish an international enforcement agency.[1081]
9. Ord	Suggests that it may be possible to strengthen existing institutions related to existential risk. As an example, "…the Biological Weapons Convention could be brought into line with the Chemical Weapons Convention: taking its budget from $1.4 million up to $80 million…"[1082]
10. Wilson	Highlights the option of expanding existing international instruments to govern GCRs and ERs from biotechnology: "[R]ather than concluding a new international agreement, states could agree to amend existing international treaties to include increased safeguards over a wider range of activities…" Nevertheless, he ultimately does not recommend this option since existing instruments would have to be radically transformed to fill the gap.[1083]
11. Farquhar and others	Explore and recommend different interventions, such as increasing efforts to develop appropriate governance of SRM research.[1084]
12. Posner	Writes about, among other things, several specific policies that may help to mitigate existential risks. As one example, Posner discusses an international bioterrorism agency that can reduce the risk of bioterrorism by, inter alia, establishing and verifying relevant standards.[1085]
13. Adams	With reference to the LHC Controversy, Adams proposes an international safety agreement that can regulate high-energy physics experiments. Among the proposed regulatory principles is the requirement to analyse and record all data and to (only) gradually increase the collision energy. Adams also explores a compliance

[1079] Boyd and Wilson, 'Existential Risks to Humanity' (n 154).
[1080] Manfred Kohler, 'The Need for Global Protections Against Existential Risks' *The Regulatory Review* (11 June 2020) <https://www.theregreview.org/2020/06/11/kohler-need-global-protections-against-existential-risks/> accessed 03 May 2023.
[1081] Nindler, 'The United Nation's Capability' (n 54).
[1082] Ord, *Precipice* (n 16).
[1083] Wilson, 'Minimizing' (n 151).
[1084] Farquhar and others, 'Existential Risk' (n 61).
[1085] Posner, *Catastrophe* (n 90).

	mechanism that links the regulatory principles to project funding. The compliance mechanisms can take the form of a clause providing that member states contributing to the project has a "…right to cease contributions and receive an automatic injunction against the project…" if the organisation (CERN) fail to comply with the agreement's safety measures. [1086]

5.5.3 The ASP and Distinct Interventions

The ASP may underpin legal interventions aiming to safeguard humanity's future. However, the ASP does not need to empower all interventions in the same way. For example, recall that it might be possible to partially mitigate x-risk by supporting organisations such as the FHI and CSER. In this context, the ASP can potentially influence funding allocation and how much resources ought to be devoted to organisations that strive to safeguard humanity's future relative to organisations that do not share this goal.

In contrast, when it comes to legal interventions, the ASP may, for example, be used to push the argument that the x-risk landscape warrants significantly more attention from legislators and, e.g., the UN.[1087] Another obvious way in which the ASP can empower legal interventions is by influencing the formulation of proportionate obligations. For instance, it seems reasonable to demand that actors take more precautions whenever the stakes are exceptionally high.

Some authors listed in Tables 5.1 and 5.2 recognise that the stakes are relevant for their legal interventions. In respect of Proposal (2), Voeneky writes that one can argue that actors should reduce risks in a way that mirrors their "probability and severity".[1088] In respect of Proposal (3), Wilson suggests that "[a]n international instrument on GCRs/ERs from emerging technologies should regulate the conduct of scientists because the stakes of GCRs/ERs are too high to leave to a small group of self-interested individuals."[1089] According to Villmer in relation to Proposal (7), procedural safeguards and judicial expediency matter little when applying NEPA to simple mining

[1086] Adams, 'Honey' (n 309).
[1087] Nindler, 'The United Nation's Capability' (n 54); Boyd and Wilson, 'Existential Risks to Humanity' (n 154); Ord, *Precipice* (n 16) 378.
[1088] Voeneky, 'Human Rights and Legitimate Governance' (n 305) 162.
[1089] Wilson, 'Minimizing' (n 151) 360.

operations; however, "[i]n the current uncharted scientific environment of particle accelerators, which could possibly result in global annihilation, NEPA modifications must become a paramount concern."[1090] Furthermore, regarding Proposal (8), Nindler writes that:

> …states are willing to invest a certain amount of resources in international enforcement mechanisms, when the stakes are high. Thus, the possibility of an adverse realization of an existential risk may be a sufficient incentive for states to contribute resources to an international enforcement agency for safe AI research and development and vest in it the necessary powers to enforce the provisions of an international treaty on safe AI research and development.[1091]

In support of Proposal (13), Adams says that "[t]here are at least a few reasons, given the stakes if physicists are wrong, why this area of research should not simply be left to the scientists."[1092]

Beyond the proposals above, Kohler's Prototype Regulation supports the general notion that the ASP can impact the extent of an actor's legal obligations. In a series of blogposts on the Regulatory institute's website,[1093] Kohler outlines a Prototype Regulation that concerns research and technology risks.[1094] The Prototype, which deals with different *risk classes*,[1095] is made up of 21 draft Articles.[1096] It deal with a range of issues, including risk assessment, risk classification, procedural obligations (by risk class), the empowerments and obligations of the 'Authority' responsible for the application of the Regulation, liability and whistle-blower protection. Of particular interest for the present purpose is Article 3(6). This Article can be interpreted as linking the ASP to the extent of an actor's obligations. It reads: "A particularly thorough risk

[1090] Villmer, 'Procedural Squabbling' (n 554) 336.
[1091] Nindler, 'The United Nation's Capability' (n 54) 31.
[1092] Adams, 'Honey' (n 309) 155.
[1093] The Regulatory Institute <https://www.howtoregulate.org/> accessed 03 May 2023.
[1094] Manfred Kohler, 'Research and Technology Risks: Part IV – A Prototype Regulation' *The Regulatory Institute* (3 March 2018) <https://www.howtoregulate.org/prototype-regulation-research-technology/> accessed 04 May 2023.
[1095] Manfred Kohler, 'Research and Technology Risks: Part III – Risk Classification' *The Regulatory Institute* (3 March 2018) <https://www.howtoregulate.org/classification-research-technology-risks> accessed 03 May 2023.
[1096] Kohler, 'Research and Technology Risks: Part IV' (n 1094).

assessment shall be undertaken when research or technologies might lead to the extinction of mankind ("existential risk")."

Article 3(6) of the Prototype Regulation aligns with Rees's suggestion that it might be reasonable to "…set an even more stringent threshold on the possible risk [of extinction] before sanctioning [certain] experiments."[1097] It is interesting to note that Kohler also recognises, pointing to Bostrom, that an existential catastrophe will "…stop the potential not only of billions, but trillions or even quadrillions of humans who could live over the next millions of years."[1098]

However, the Prototype does not offer much detail about this 'particularly thorough risk assessment'. Instead, and according to Article 3(11), the Prototype suggests that research or technology undertakings[1099] "…bearing an existential risk are only acceptable when they remedy another existential risk with higher likelihood." Therefore, rather than a 'particularly thorough risk assessment' it appears that the Prototype – reading Article 3(6) and 3(11) together – sets an absolute threshold for when undertakings posing existential risk are (un)acceptable. Another thing to keep in mind is that the Prototype assumes that it's possible to accurately assign probabilities to two different existential risks. Whereas it would have been interesting to get some further details on the requirements of the 'particularly thorough risk assessment', Kohler's idea that an x-risk requires special treatment, given that they are unique, aligns with the ASP.

In summary, it has been argued that the ASP should empower legal interventions mitigating risks within the x-risk landscape. It has also been proposed that the ASP will not necessarily underpin all such interventions in the same way.
To provide a more balanced view, the upcoming Part will offer arguments against the proposal that the ASP should underpin legal interventions within the x-risk landscape. Notably, the discussion below is framed to offer some arguments against the specific proposal that the ASP should influence the extent of an actor's substantive obligations. However, it is worth repeating that this is merely one way by which the ASP can

[1097] Rees, *Our Final Century* (n 226) 126.
[1098] Kohler, 'Research and Technology Risks: Part IV' (n 1094).
[1099] See Article 2 of the Prototype for Definitions.

underpin a particular legal mechanism. Nevertheless, framing the discussion from this perspective will streamline the upcoming Part.

5.6 REJECTING THE ASP'S INFLUENCE ON LEGAL INTERVENTIONS

The premise of the ASP is that an x-risk has astronomically high stakes and that an x-catastrophe is uniquely bad relative to virtually all non-existential catastrophes. The ASP can underpin legal interventions for x-risk mitigation in numerous ways. As one example, the ASP may empower broad or targeted legal interventions that promote unprecedented responsibility from alleged x-risk originators such as CERN. Nevertheless, the proposal that the ASP should underpin legal interventions for x-risk mitigation is not without flaws. It is possible to think of many reasons why someone might want to reject the proposal that the ASP should empower a legal intervention to promote unprecedented responsibility from x-risk originators. The upcoming sections will highlight some of these reasons.

5.6.1 Disagreeing with the Spirit of the ASP

The view that the destruction of the future makes an x-catastrophe uniquely bad is not universally accepted. In other words, one may, for various reasons, disagree with the essence of the ASP. What this means is that risks within the x-risk landscape are not as 'uniquely bad' as suggested in this Chapter.

The following objections to the ASP can be noted: Firstly, one might believe that humanity will, most likely, not survive for a very long time (e.g., because space colonisation will never be possible) or that the future will contain more *bad things* than *good things*. Such beliefs can support the view that the future is not likely to contain as much value as some scholars predict. Thus, these beliefs can sway someone's perceptions of what is at stake when the destruction of humanity's future is at risk. Secondly, some might highlight discounting as a reason to care less about the future.

Thirdly, a person-affecting view can sway whether and to what extent one is likely to accept the ASP.[1100]

All in all, not everyone will agree with the view that the destruction of the future makes a catastrophe uniquely bad. There is much literature justifying and rejecting the unique badness of humanity's premature extinction.[1101] This thesis does not deal with this in any detail. Instead, the thesis springs from the premise that a non-negligible probability of a vast and grand future does matter and ought to set x-risks apart as uniquely bad. This Chapter has also attempted to justify this view, to some extent, by noting the potential length and quality of humanity's future.

In short, if one disagrees with and rejects the ASP itself, then the argument that the ASP should influence legal interventions for x-risk mitigation is largely unpersuasive.

5.6.2 Rejecting the 'Link' Between the ASP and Legal Interventions

One might accept the ASP in theory but reject that it should empower legal interventions in practice by, for example, affecting the substantive obligations of x-risk originators. Different reasons can lead to such a view. Firstly, the ASP can lend itself to draconian measures. There is a risk that the corresponding obligations will become too stringent. As Rees points out, there is a hidden cost of saying no, and the most extreme precautionary policy will "…utterly paralyse science."[1102] This effect is clearly unfortunate given that scientific and technological progress, while associated with risk, is also highly beneficial and will likely continue to improve the human condition. As stated in Chapter 1, it may even enable humanity to avoid one or more x-catastrophes. While this thesis only deals with extinction risks, there are, as noted previously,

[1100] See e.g., Cody Fenwick, 'Longtermism: A Call to Protect Future Generations' 80,000 Hours, 28 March 2023) <https://80000hours.org/articles/future-generations/> accessed 04 May 2023.
[1101] See e.g., Elizabeth Finneron-Burns, 'Contractualism and the Non-identity Problem' [2016] Ethical Theory and Moral Practice 1151; Elizabeth Finneron-Burns, 'What's Wrong with Human Extinction?' [2017] Canadian Journal of Philosophy 327; Johann Frick, 'On the Survival of Humanity' [2017] Canadian Journal of Philosophy 344; Patrick Kaczmarek, 'How Much is Rule-Consequentialism Really Willing to Give Up to Save the Future of Humanity?' [2017] Utilitas 239; Beard and Kaczmarek, 'On the Wrongness of Human Extinction' (n 1028); Beckstead, 'On the Overwhelming Importance of Shaping the Far Future' (n 342) ch 4 and 5; Ord, *Precipice* (n 16) ch 2.
[1102] Rees, *Our Final Century* (n 226) 129.

different classes of existential risk.[1103] For example, surviving without reaching technological maturity constitutes a class of x-risk known as *permanent stagnation*.[1104] Draconian measures inspired by the ASP may contribute to a stagnation scenario which would be paradoxical since the ASP is inspired by x-risk-related resources.

Secondly, it is possible to suggest that it sets a dangerous precedent to allow the ASP to affect substantive obligations. If affecting obligations, it may also come to affect other types of practical measures or ideals. For instance, can the ASP back a decision to kill or omit to save millions of existing people in the name of x-risk reduction?[1105] A few lines from one of Bostrom's article might highlight why a sceptic might feel uneasy here. Suggesting that force or the threat of force might be necessary, his article reads:

> A preemptive strike on a sovereign nation is not a move to be taken lightly, but in the extreme case we have outlined – where a failure to act would with high probability lead to existential catastrophe – it is a responsibility that must not be abrogated. Whatever moral prohibition there normally is against violating national sovereignty is overridden in this case by the necessity to prevent the destruction of humankind.[1106]

Regardless of whether such a move is legal or not, why should anyone or anything stand in the way of Utopia as described Bostrom's article?[1107] To Bostrom's defence, his article reads that a pre-emptive strike is reserved for situations where a failure to act would with high probability bring about an existential catastrophe. But what should count as 'high probability' when the downside is the destruction of humanity's future?

In sum, it is possible to accept the ASP in theory but reject the practical link between this proposition and legal interventions for x-risk mitigation.

[1103] Bostrom, 'Existential Risk Prevention' (n 15) 19.
[1104] ibid 20.
[1105] For instance, Häggström accepts Bostrom's point as a pedagogical device for driving home the argument that x-risk reduction is important and that a small reduction can be worth substantial costs. Still, he outlines a scenario demonstrating that taking the point literally can lead to questionable outcomes. Häggström, *Here Be Dragons* (n 22) 240-241.
[1106] Bostrom, 'Existential Risks' (n 15).
[1107] Bostrom, 'Letter from Utopia' (n 967).

5.6.3 Counter-Intuitive

The objection in this Section is similar to the one above in that one may accept the ASP in theory but disagree that it should affect legal interventions in practice. However, whereas the Section above cited stringent obligations as a reason for such a view, this Section offers another reason for rejecting the link between the ASP and legal interventions for x-risk mitigation. It is possible to argue that the practical effects that may flow from connecting the ASP and legal mechanisms feel counter-intuitive.

More specifically, one may believe it is counter-intuitive that the corresponding obligations should be more stringent vis-à-vis x-risk than when, for instance, 10%, 50% or 70% of the population is at risk. Consider the following example. Imagine that Organisations A and B are about to undertake two separate experiments. Suppose that some critics contend that Organisation A's experiment poses a GCR. It should be recalled that it is possible to define a 'global catastrophe' as a catastrophe that would end the lives of 10% of the global population or do comparable damage.[1108] In effect, if the alleged GCR associated with Organisation A's experiment would materialise, 800 million people would die if, for the sake of simplicity, it is presumed that the global population is 8 billion. On the other hand, suppose that the experiment associated with Organisation B is said to pose an x-risk. Is it reasonable to make a persuasive case that Organisation B should take more precautions than Organisation A? What more can be demanded of the former originator that could not be demanded of the latter? Is it not sensible to suggest that equally onerous responsibilities and obligations must arise when an activity or experiment threatens to kill hundreds of millions of people?

In short, one may argue that empowering legal interventions using the ASP can lead to outcomes that feel counter-intuitive.

5.6.4 Issues Relating to the Severity Threshold

It is possible to argue that legal interventions empowered by the ASP are impractical because they may become contingent upon a particular 'harms' threshold. This ground

[1108] Cotton-Barratt and others, 'Global Catastrophic Risks' (n 150) 22.

of criticism takes us back to earlier parts of the thesis. Chapter 2 looked at the quantification of the probability of x-catastrophes with reference to Baum's article.[1109] As noted there, it is not enough to consider the probability of the initial catastrophe event (i.e., P_1). It is also necessary to consider the probability that the initial catastrophe event will cause harm which exceeds a specific harms threshold (i.e., P_2). For this thesis, the relevant harms threshold is the extinction of humanity since it defines an x-risk as the 'a risk that threatens the premature extinction of humanity.'[1110] The ASP is ultimately contingent upon this threshold since the ASP is mainly relevant to risks that threaten to destroy humanity's future via extinction.

Suppose that the ASP empowers a legal intervention to promote onerous obligations being imposed upon x-risk originators. This can lead to different problems. For instance, the risk originator may attempt to dispute the existence of such high obligations whenever it is unclear that a postulated risk, should it materialise, could bring about humanity's premature extinction. Posner considers a somewhat related idea in the context of a hypothetical scenario involving RHIC and an appellate science court.[1111] According to Posner, one objection to such a court is the challenge of configuring its jurisdiction. He writes:

> Whether its jurisdiction were defined as embracing cases that involve scientific issues, or (improbably) confined to cases that involve catastrophic risks, there would be endless haggling over whether the jurisdictional triggers were sufficiently important to the decision of a particular case to warrant channeling the appeal to the science court.[1112]

He also recognises that it is hard to formulate criteria for differentiating "…catastrophic from subcatastrophic risks" to confine the court's jurisdiction to catastrophic risks.[1113] A critic may propose that similar problems can manifest in relation to legal interventions if they are underpinned by the ASP.

[1109] Baum, 'Quantifying the Probability' (n 125).
[1110] See Table 2.1.
[1111] Posner, *Catastrophe* (n 90) 209-213.
[1112] ibid 210.
[1113] ibid 211.

A relevant observation here is that very few catastrophes seem capable to bring about the end of humanity. That is, many catastrophes – even if capable of killing a vast majority of people – may fail to qualify as existential since the survivors, of which there may be at least some, [1114] might be able to rebuild civilisation over time. In line with Baum and others, these may be called "sub-extinction catastrophes", and the fate of the post-catastrophe survivors and their ability to rebuild civilization, and to what extent, will reportedly depend on various factors such as the nature of catastrophe itself, the immediate aftermath of the catastrophe and whether agriculture (and industry) is maintained or redeveloped following the catastrophe.[1115] However, as far as the LHC Controversy is concerned, the postulated catastrophes would very likely qualify as existential since there is seemingly no chance whatsoever to recover should the disaster scenarios materialise.

To summarise, legal interventions underpinned by the ASP may become dependent upon a high 'harms' threshold. Consequently, these interventions may lose some of their relevance if there are uncertainties vis-à-vis such a threshold.

5.6.5 The Intervention is too Narrow

A legal intervention underpinned by the ASP might prove fruitless on account of being too narrow. As mentioned in the Section above, few catastrophes seem capable of bringing about humanity's end. This might mean that interventions empowered by the ASP will only apply to very few risk originators.

Furthermore, legal interventions promoting unprecedented responsibility may not easily find application when it comes to (i) natural x-risks, (ii) risks which can be classified as terror risks, or (iii) multi-risk catastrophes.[1116] Such an intervention is,

[1114] People are scattered across the planet, some of which are largely isolated from external contact for extended periods of time: "Even if a process manages to kill most of humanity, a surviving few might be able to rebuild." Cotton-Barratt, Daniel and Sandberg, 'Defence' (n 37) 272; See also Nick Beckstead, 'How Much Could Refuges Help us Recover from a Global Catastrophe?' [2015] Futures 36, 37-38.

[1115] Baum and others, 'Long-term Trajectories' (n 118) 59-67.

[1116] The thesis has borrowed the term 'multi-risk catastrophes' from Baum, 'Quantifying the Probability' (n 125).

perhaps, more applicable to, what Kuhlemann refer to as, "sexy" risks as opposed to "unsexy" ones.[1117] According to her article, sexy risks are "…neat, quick and techy."[1118] In particular, they have clear disciplinary homes (*epistemic neatness*), and we expect them to crystallise abruptly (*sudden onset*).[1119] In contrast, *epistemic messiness* and *gradual build up* are two hallmarks of unsexy risks.[1120]

In short, whereas the ASP may empower an intervention promoting unprecedented responsibility, such intervention may prove too narrow in that it will not often apply.

5.6.6 The ASP's Practical Impact is Idealistic

Some might say it is too idealistic to believe that the ASP can empower legal interventions. Such a view may be embraced for several reasons. Firstly, the lack of attention may undermine proposed interventions. Whereas x-risk mitigation will generally require a global response, many x-risks are overlooked at an international governance level. This thesis previously referred to Boyd and Wilson's finding that some x-risks have received *relatively* little attention in the publications of the UN.[1121] The limited and selective engagement with these risks may signify that many proposals, such as using the ASP as an empowerment tool, will fail to gain attraction.

Secondly, even if the lack of attention is addressed, the next issue revolves around agreeing upon and implementing solutions that can mitigate existential risk, and doing so in a way that utilises the ASP. Taking this to the international arena will, as noted in Chapter 2, intensify the issue. Many general problems (e.g., issues coming from the consent-based nature of international law and issues regarding efficient monitoring and enforcement) can delay, undermine, or prevent a theoretical approach from coming to fruition as a practical one.

[1117] Kuhlemann, 'Complexity, Creeping Normalcy and Conceit' (n 1052).
[1118] ibid 41.
[1119] ibid; See fast versus slow catastrophes in Baum, 'Quantifying the Probability' (n 125).
[1120] Kuhlemann, 'Complexity, Creeping Normalcy and Conceit' (n 1052) 42.
[1121] Boyd and Wilson, 'Existential Risks to Humanity' (n 154).

Borrowing a sentence from Kuhlemann, it is a mistake to assume that "…a solution that is imaginable is *also* technically and politically achievable *and* likely to be developed *and* implemented in time *and* work sufficiently well to adequately mitigate a catastrophic risk."[1122]

5.6.7 Summary

The view that the ASP should underpin legal interventions is not without shortcomings. Nonetheless, the objections above do not render the view futile. For each claim above, it is possible to think of counterarguments.

For example, as mentioned above, a critic may disagree with the essence of the ASP if she believes that humanity will not survive for a long time to come, or that the future will contain more bad things than good things. However, earlier parts of this Chapter can be used to challenge potential beliefs that humanity's future is destined to be short or bad. As mentioned, there is a non-negligible probability that the future will be grand, in terms of both length and quality. While there are philosophical views which reject the unique badness of an x-catastrophe, it is not necessary to unravel all these issues to decide whether x-risks deserve special treatment. Ord makes a solid remark here:

> [W]e need not resolve every philosophical issue about the value of the future in order to decide whether humanity's potential is worth protecting. For the idea that it would be a matter of relative indifference whether humanity goes extinct, or whether we flourish for billions of years, is, on its face, profoundly implausible.[1123]

As a further example, Section 5.6.4 cited uncertainties in terms of P_2 as a reason to question the link between the ASP and legal interventions for x-risk mitigation. In short, it was noted that interventions empowered by the ASP might lose their relevance if risk originators can successfully argue that it is impossible to prove that a postulated catastrophe would, with 100% certainty, bring about humanity's extinction. An analogy can be drawn to address this line of thinking. Clearly, it is possible to deem a risk unacceptable even if it is not possible to demonstrate that it will, with 100% certainty,

[1122] Kuhlemann, 'Complexity, Creeping Normalcy and Conceit' (n 1052) 36.
[1123] Ord, *Precipice* (n 16) 48.

occur in the first place. In other words, it is not necessary to establish $P_1 = 1$ before certain risks merit mitigation and regulation.

Similarly, it cannot be correct to suggest that it must be demonstrable that $P_2 = 1$ before legal interventions empowered by the ASP can find application. In other words, it cannot be correct to suggest that such mechanisms can only be relevant when it is possible to demonstrate that a catastrophe will, with 100% certainty, exceed the harms threshold. Instead, the probability that the catastrophe might (directly or indirectly) interfere with humanity's ability to survive ought to be enough to trigger the application of the ASP and interventions empowered by it.

5.8 CONCLUSION

Chapter 5 has explored the ASP, which is an essential gadget when it comes to x-risk mitigation. Whereas humanity's future may prove grand in terms of length and quality, x-risk threatens to foreclose humanity's trajectory towards such a future. A central argument of Chapter 5 is that the ASP should underpin legal interventions for x-risk mitigation. The Chapter has also examined different interventions that may be empowered by the ASP. This Chapter feeds into the next one given that the LHC-inspired intervention will utilise the ASP.

6 CHAPTER SIX: A BROAD LEGAL INTERVENTION IN RESPONSE TO THE LHC CONTROVERSY

> [T]he unprecedented power from technological progress requires unprecedented responsibility: both for the practitioners and for those overseeing them.[1124]

> Our understanding of nature is limited, and there are surely many dangers we have not yet appreciated. Due caution is appropriate.[1125]

6.1 BACKGROUND

Humanity's activities can no doubt influence the x-risk landscape. From the exacerbation of existing or creation of novel x-risks to their mitigation, humanity's decisions are crucial in moulding the risk landscape. There are no past existential catastrophes that can guide policymakers to avoid or lessen future x-catastrophes. This self-evident observation does not mean that the past cannot be used as a source of inspiration for the design of legal interventions for x-risk mitigation. The LHC Controversy—even if the probability of catastrophe is zero—is a dispute from the past that can inspire the design of a broad legal mechanism that sits and operates within the wider x-risk landscape. In effect, the LHC Controversy can serve as a sort of drafting guide for future legal interventions. Some legal scholars have used the LHC Controversy to inform their discussions. However, these discussions are not generally framed within the x-risk landscape. This has meant that wisdom from the field of existential risk studies has not been of central interest to many of those discussions.

6.1.1 Aims of Chapter 6

This Chapter aims to formalise the theoretical substance of the LHC-inspired intervention. Chapter 6 is the culmination of various observations made up until this point in the thesis. It builds upon the observations made vis-à-vis the x-risk landscape, the LHC Controversy and the ASP.

[1124] Ord, *Precipice* (n 16) 207.
[1125] Kent, 'A Critical Look' (n 156) 166.

Importantly, there are three self-imposed conditions that the intervention should satisfy. The first condition is that the LHC-inspired intervention should respond to the review barriers that may affect different modes of external review in relation to the LHC Controversy. Chapter 3 highlighted these obstacles. The second condition is that the intervention should utilise the ASP. Chapter 5 dealt with the ASP in detail. The third condition is that the mechanism should, with some degree of plausibility, be able to find utility beyond the LHC Controversy. Satisfying this last condition makes the intervention broad rather than targeted in nature.

Chapter 6 answers the following sub-question: How does the LHC-inspired intervention (i) operate in theory, (ii) align with x-risk studies and, (iii) utilise the ASP? Not only is Chapter 6 fundamental in addressing this sub-question, but it is also the central cog in answering the thesis's main question. It is to be recalled that the main research question is as follows: How can the LHC Controversy inspire the design of a broad legal intervention to mitigate existential risk?

A proposed answer to this question was outlined in Chapter 1. It is useful to briefly repeat the proposed answer before going into the details. It is essentially suggested that the LHC Controversy reveals the value of a broad legal intervention that targets the reliability of the relevant x-risk assessment. This intervention is disconnected from probabilities, and it does not involve the validation of competing scientific theories. Instead, the legal mechanism involves analysis of interconnected deficiency factors, which can give weight to an argument that the x-risk assessment is not, at present, as reliable as it should be (especially given the astronomical stakes). These deficiency factors require consideration of (i) the actors who consider the relevant x-risk and (ii) the potential that they have based their conclusion on flawed information or an incomplete state of knowledge or understanding. It should be recalled that the thesis will use Johnson's meta-analysis to compartmentalise the deficiency factors that can be analysed as part of considering the reliability of the x-risk assessment.

Finally, Chapter 6 revisits the notion of an 'external reviewer'. As such, it is practical to repeat that reference to an external reviewer means any independent third party tasked with considering the reliability of the relevant x-risk assessment. This Chapter briefly considers who such a reviewer might be.

6.2 THE LHC CONTROVERSY: A SOURCE OF INSPIRATION

The LHC Controversy has been used by other scholars as a source of inspiration for their discussions. It is to be recalled that the 'LHC Controversy' is used as a collective term to mean the disagreements and uncertainty as to whether high-energy physics experiments facilitated by particle accelerators (i) pose any existential risk to humanity, (ii) pose an unacceptable x-risk to humanity, or (iii) should be delayed or suspended. As such, the LHC is not the only accelerator referred to in this Part. The RHIC is another accelerator, having been the subject of much discussion in the context of risk and humanity's premature extinction.

Legal scholars who have considered or utilised the LHC Controversy to a greater or lesser extent include Johnson,[1126] Posner,[1127] Adams,[1128] Peterson,[1129] Villmer,[1130] Wilson,[1131] Voeneky,[1132] and Beyleveld and Brownsword.[1133] These works demonstrate that the LHC Controversy has inspired legal scholars to think about different solutions to the same dispute.

Of course, authors beyond those cited above have looked at the LHC Controversy from a legal perspective. Some of these include Petrenko and McArthur,[1134] Neuwirth[1135] and Lehmann.[1136] For example, Lehmann, a trial attorney in California having allegedly "…represented more aircraft owners in U.S. litigation than any lawyer in the United States…"[1137] has written a book inspired by the LHC Controversy. For

[1126] Johnson, 'The Black Hole Case' (n 68); Johnson, 'Agencies and Science-Experiment Risk' (n 68); Johnson, 'Uncertain Risk, Science Experiments, and the Courts' (n 68); Johnson, 'Judicial Review' (n 68).
[1127] Posner, *Catastrophe* (n 90); Richard Posner, 'Public Policy Towards Catastrophe' in Nick Bostrom and Milan Ćirković (eds), *Global Catastrophic Risks* (OUP 2008).
[1128] Adams, 'Honey' (n 309).
[1129] Peterson, 'The "Sancho" Effect' (n 468).
[1130] Villmer, 'Procedural Squabbling' (n 554).
[1131] Wilson, 'Minimizing' (n 151).
[1132] Voeneky, 'Human Rights and Legitimate Governance' (n 305).
[1133] Beyleveld and Brownsword, 'Emerging Technologies, Extreme Uncertainty, and the Principle of Rational Precautionary Reasoning' (n 448) 55-56.
[1134] Petrenko and McArthur, 'Between Same-Sex Marriages and the Large Hadron Collider' (n 584).
[1135] Neuwirth, 'Law and Magic' (n 613).
[1136] Lehmann, *No Canary in the Quanta* (n 94).
[1137] ibid 116.

example, referring to negligence,[1138] he writes that the level of care owed by Y is correlative with the risk. If the activity may cause very grave harm, then the level of care is also very high. In effect, if the destruction of Earth is at stake, absolute care is required.[1139] Lehmann also proposes an absolute ban on scientific experiments failing the 'Hadron Test' – a hypothetical test devised by the author.[1140] To understand this test, one should note that Lehmann's corresponding 'Hadron Rule' is that no scientist "…should ever be allowed to undertake, any experiment, of any sort, which risks the very existence of our home world, to any extent."[1141] Another theme throughout his book is that the scientists running the experiment cannot reliably assess the risk associated with their own project.[1142]

One might recall that Adams's intervention was classified as a targeted intervention in Chapter 5 because it concerns the regulation of experimental particle physics.[1143] In contrast, Posner's book reveals a potential broad legal intervention. Posner discusses the RHIC throughout his book. In discussing some "highly tentative"[1144] measures for reducing catastrophic risks, he writes:

> Congress should consider enacting a law that would require all scientific research projects in specified areas, such as…experimental high-energy physics, to be reviewed by a federal catastrophic-risks assessment board and forbidden if the board found that the project would create an undue risk to human survival.[1145]

Other broad interventions inspired by the LHC Controversy include Peterson's proposal that a new science court should be created to deal with cases like the *Sancho case*.[1146] Such a court should, in his view, deal with the scientific merits of the dispute. According to Peterson, "…courts must be equipped to independently examine the

[1138] ibid 33-34.
[1139] ibid 34.
[1140] ibid 14.
[1141] ibid.
[1142] ibid 1-5.
[1143] Adams, 'Honey' (n 309).
[1144] Posner, *Catastrophe* (n 90) 216.
[1145] ibid 221.
[1146] Peterson, 'The "Sancho" Effect' (n 468).

merits of competing scientific claims."[1147] He continues by saying that "…courts must, to the extent possible, *become* the experts."[1148]

In contrast, Johnson's core premise is, as might be recalled from Chapter 4, that a generalist court could undertake a meta-analysis to deal with the situation.[1149] For instance, in 2016 he writes that "[c]ourts need not understand the science as scientists do to gauge the reliability of the scientists' risk-assessment work in the aggregate."[1150] Johnson supports this claim by referring to one of his earlier works, namely the Black Hole Case Article.[1151] The accompanying footnote in the 2016 article states that the Black Hole Case Article is about how courts can provide meaningful judicial review of science-intensive risk assessments despite the subject matter being obscure to generalist judges.[1152]

When comparing Johnson's and Peterson's approach, an underlying question is brought to the frontline. Namely, should legal interventions inspired by the LHC Controversy direct an external reviewer to focus on the scientific merits of the Controversy or something else? The phrase 'something else' can mean anything apart from the scientific merits of the disputants' arguments.

Legal proposals owing their inception to the LHC Controversy can clearly be more or less conditional upon the external reviewer's knowledge of theoretical and experimental particle physics. In particular, a targeted intervention will likely require more subject matter expertise than a broad intervention. Take Adams's article as an example. His sketch of a regulatory framework draws inspiration from the 'risk mitigating measures' in Plaga's paper.[1153] There is no need to consider the merits of the specific measures here. The point is this: Adams's proposal is conditional upon a

[1147] ibid 310.
[1148] ibid.
[1149] Johnson, 'The Black Hole Case' (n 68); Johnson uses the term 'generalist courts' in Johnson, 'Judicial Review' (n 68) 84, and 'generalist judges' in Johnson, 'Agencies and Science-Experiment Risk' (n 68) 527.
[1150] ibid 584.
[1151] Johnson, 'The Black Hole Case' (n 68).
[1152] Johnson, 'Agencies and Science-Experiment Risk' (n 68) 584.
[1153] Rainer Plaga, 'On the Potential Catastrophic Risk From Metastable Quantum-Black Holes Produced at Particle Colliders' (2008) <https://doi.org/10.48550/arXiv.0808.1415> accessed 04 May 2023.

deeper level of understanding of theoretical and experimental particle physics. The same is not true when considering, for example, Johnson's approach. As another example, the same scientific understanding is arguably not required to insist upon and implement "…a code of conduct that requires scientists to monitor their own ethical and professional conduct as well as the ethical and professional conduct of their peers and supervisors."[1154] This quote is from Wilson's article. His prototype treaty on GCR/ERs is not geared towards high-energy physics experiments, but he makes his statement in a section that highlights how the scientists at CERN "…were able to be their own risk assessors…"[1155]

The discussion above demonstrates that the LHC Controversy can inspire one to think about different legal responses to the same fundamental problem. However, and as seen throughout this thesis, the LHC Controversy has also been noted by scholars active within the x-risk community.[1156] This Chapter will further consider some of these works below.

6.3 DEFICIENCY FACTORS AND RELIABILITY

As previously explained, this thesis argues for a broad legal intervention which gives an external reviewer a principled basis upon which to base a finding that an endeavour should not presently be allowed to proceed since—despite the review barriers[1157] or a combination therefore—there are reasonable doubts as to the reliability of the work assessing the x-risk linked to this endeavour.

[1154] Wilson, 'Minimizing' (n 151) 361.
[1155] ibid 360-361.
[1156] Leslie, *The End of the World* (n 101) 8-9 and 123-127; Rees, *Our Final Century* (n 226); Ord, *Precipice* (n 16) 325; Baum, 'Quantifying the Probability' (n 125); Cotton-Barratt, Daniel and Sandberg, 'Defence' (n 37) 274; Beard, Rowe and Fox, 'An Analysis and Evaluation of Methods Currently Used to Quantify the Likelihood of Existential Hazards' (n 124); Bostrom, 'The Vulnerable World Hypothesis' (n 465) 462; Bostrom, 'Existential Risks' (n 15); Kent, 'A Critical Look' (n 156); Toby Ord, Rafaela Hillerbrand and Anders Sandberg, 'Probing the Improbable: Methodological Challenges for Risks with low Probabilities and High Stakes' [2010] Journal of Risk Research 191; Häggström, *Here Be Dragons* (n 22) 198-200; Bostrom and Ćirković, 'Introduction' (n 12) 18-19; Torres, *Morality, Foresight, and Human Flourishing* (n 17) 89-90.
[1157] Chapter 3 considered these barriers.

Unlike some of the proposals outlined in the Part above, this intervention does not concern the scientific merits of the relevant safety assessment. As such, the LHC-inspired intervention does not involve the external reviewer validating conflicting scientific theories. Furthermore, the intervention does not involve cost-benefit analysis and is largely disconnected from probabilities of harm. Instead, the LHC-inspired intervention invites the reviewer to analyse a set of interconnected deficiency factors which can give credence to a concern that the scientific work assessing existential risk(s) is not, at present, as reliable as it should be.

The upcoming sections expands upon these deficiency factors and justify why it can be useful to analyse them as part of evaluating the reliability of the x-risk assessment. Importantly, Johnson's meta-analysis[1158] is used to compartmentalise these deficiency factors and to structure the discussion below. It should be recalled that Johnson's meta-analysis revolves around (i) Defective Theoretical Groundings, (ii) Faulty Scientific Work, (iii) Credulity and Neglect, and (iv) Bias and Influence.

It is to be recalled that Johnson does not situate his discussions within the study of existential risk. In fact, one of his observations is contrary to a somewhat cardinal principle within the field of x-risk. Namely, existential risk scholars are often concerned with humanity's future and vast potential. As such, extinction is generally regarded as uniquely bad, even if no one is ultimately left to mourn such a tragedy after the fact. This is not an essential theme in Johnson's works. For example, at one point he writes: "[M]aybe the downside of a particle-accelerator disaster that destroys the planet—assuming it is quick—is nothing."[1159] Johnson says this after considering how death itself is not a redressable injury under American tort law absent special circumstances.[1160] However, as will become apparent below, Johnson's meta-analysis aligns well with observations that circulate within the field of x-risk studies. This is also one reason behind the decision to use the meta-analysis to compartmentalise the deficiency factors.

[1158] Chapter 4 outlined Johnson's meta-analysis.
[1159] Johnson, 'The Black Hole Case' (n 68) 883.
[1160] ibid.

6.3.1 Reviewing the Science: Defective Theoretical Groundings

In the context of Johnson's article, the category of *defective theoretical groundings* (the first analytical method) invites a generalist judge to analyse "…the possibility that the science underlying the exclusions of disaster…may ultimately be wrong."[1161] This involves the big picture stuff: Macro-scale scientific error.[1162] It centres around giving weight to the possibility that the science supporting the finding that the relevant endeavour is safe may ultimately prove wrong or incomplete.

Taking this to the context of the LHC-inspired intervention, the first deficiency factor that an external reviewer can analyse as part of assessing the reliability of the x-risk assessment is the notion of 'defective theoretical groundings.' It is part of the analytical exercise involving the external reviewer to consider the potential that the risk assessors have based their conclusion on flawed information or an incomplete state of knowledge or understanding. Analysis of this first deficiency factor involves appreciation that new insights may come to light over time and that these can reveal cracks in scientific theory regarded as indisputable today.

Understandably, some may be sceptical of the gist of this analytical exercise. They may argue that it undermines scientific rigour by evading the substantive issues that should dominate the dispute. Opponents may contend that it gives credence to a person's argument that the reliability of the x-risk assessment is questionable because we do not know what we do not know. It creates an imbalance in terms of the burden of proof and standard of proof.

Johnson recognises that an unrefined version of the defective theoretical groundings argument can hinder any experiment from commencing.[1163] Indeed, it is necessary to consider if an external reviewer can analyse the first general deficiency factor in a principled way. The Black Hole Case Article promotes a principled application by

[1161] ibid 887.
[1162] ibid 886.
[1163] ibid 888.

emphasising, in particular, the "temporal dimension of scientific theory."[1164] Johnson writes:

> The longer a theory persists, the more confidence it deserves. Conversely, if theory is new, it should be afforded less confidence. A matter related to the longevity of theory is the pace of relevant theoretical work. If a theory is the direct subject of back-and-forth papers arguing its merits, the theory deserves, for the time being, less confidence.[1165]

The LHC Controversy demonstrates the relevance of analysis of the first deficiency factor and the potential that the risk assessors have based their conclusion on imperfect information or an incomplete state of knowledge or understanding. As both Johnson's[1166] and other works demonstrate, "[t]he history of particle-collider safety assurances contains a quick succession of flip-flops on theory that necessitated rethinking prior conclusions."[1167] Chapter 4 highlighted that this observation was made by others before Johnson formalised his argument in the Black Hole Case Article. More specifically, Chapter 4 referred to Johnson's blog posts on *PrawfsBlawg*. In commenting on one of these posts, one person wrote: "The idea that scientists are adequately careful and knowledgeable about risks…is refuted by the history of 'safety factor' arguments for colliders."[1168]

[1164] ibid.

[1165] ibid.

[1166] The following paragraph in Johnson's article is telling of this: "The LHC/black-hole issue provides a good example of quickly-shifting theory that might cause one to doubt the reliability of safety assurances: In 1999, the argument for accelerator safety rested on the conclusion that, under prevailing theory, accelerators for the foreseeable future lacked the power to create black holes. Not long afterward, theorists showed that if new theory about the existence of hidden dimensions in the universe turned out to be correct, black holes "will be produced." A new safety case was fashioned on the basis that black holes would evaporate because of Hawking radiation. Then a few years later, that argument was abandoned after a respected theorist called black-hole evaporation into question. With CERN's safety rationale eroded, a new safety case was fashioned for the LHC in 2008, on the eve of the collider's launch. This new safety rationale followed a branching logic to conclude that black holes were not a danger since under some assumptions black holes could be ruled out based on empirical observations of certain white-dwarf stars, and under other assumptions, black holes would grow too slowly to constitute a threat." (Citations omitted) See Johnson, 'Uncertain Risk, Science Experiments, and the Courts' (n 68) 370.

[1167] Johnson, 'The Black Hole Case' (n 68) 888.

[1168] See James Blodgett's comment in the comment section to Johnson, 'Scientocracy and the need for Judicial Process' (n 881).

The back-and-forth narrative dominating the safety arguments is even present in Larsen's book, which seeks to rationalise the (undue) fears of particle accelerator experiments as a matter of failed communication.[1169] Her book features a general discussion on the safety assessments and associated scientific papers relating to the LHC and RHIC.[1170] Despite Larsen's stance, her overview of the safety assessments and related papers actually reinforces the idea that Johnson conveys. Consider, for example, the following observation from her book:

> Responding to these new predictions and the potential for renewed public concerns, CERN issued a lengthy safety study in 2003 while the LHC was under construction. All suggested catastrophe scenarios brought up in the RHIC report were revisited in light of advances in theoretical knowledge, and it was acknowledged that if space has more than three dimensions, microscopic black holes might be produced at the LHC. However, the report reaffirmed that the Hawking mechanism would destroy such objects before they could begin to pose a threat.[1171]

Her book subsequently highlights how additional criticism after the 2003 safety study possibly led CERN to commission a second report in 2008, which excluded x-risk because, among other things, the Hawking radiation mechanism would destroy any black holes produced at the LHC before they posed any threat.[1172] Larsen continues: "The result was [then] a flurry of papers confirming and disputing the results, especially those related to mini black holes and Hawking radiation."[1173] In short, the science underlying the exclusion of the relevant x-risks was subject to much back and forth debate.

The overall point is that the newness of the underlying scientific theory, the lack of historical robustness of relevant safety arguments, the gravity of past upsets within the field, the pace at which new ideas are introduced, and the frequency by which old ones are challenged can all give weight to assertations that the assessment excluding risk is potentially based upon defective scientific theory. While this theme can (and should) be extracted beyond the field of high-energy physics, Larsen's book reminds

[1169] Larsen, *Particle Panic!* (n 265).
[1170] ibid 91-111.
[1171] ibid 100.
[1172] ibid 100-101.
[1173] ibid 101.

us that "[t]here are other potentially unpredictable incidents…including the discovery of new types of particles or perhaps even new physical principles."[1174] On this theme, one may also note the discovery highlighted in the recent BBC article '*Shock Result in Particle Experiment Could Spark Physics Revolution*'.[1175] Furthermore, Ord, Hillerbrand and Sandberg remind us that "[t]he history of science contains numerous examples of how generally accepted theories have been overturned by new evidence or understanding, as well as a plethora of minor theories that persisted for a surprising length of time before being disproven."[1176]

All in all, the discussion above suggests there may be scope for an external reviewer to consider the possibility that the risk assessors have based their conclusions on flawed information or an incomplete state of knowledge or understanding. This analytical exercise is not defeated by the scientific or probabilistic uncertainties that may surround the dispute. In fact, overwhelming uncertainty *may* count in favour of the external reviewer because it might give credence to the argument that the x-risk assessment is not presently as reliable as it should be.

How does the first deficiency factor relate to x-risk research and vice versa? Firstly, scholars within the field of x-risk studies have, in line with the notion of defective theoretical groundings, reasoned that it is necessary to keep an open mind, accept the limits of humanity's current knowledge and be humble and mindful of the present generation's brief and early place in humanity's overall history.[1177] Torres even suggests that some concepts may forever lie beyond a human's cognitive space.[1178]

Secondly, philosophical views and other x-risk-related observations can support the position that an external reviewer should reasonably give credence to the potential that new insights may come to light over time and in the future. More specifically, x-

[1174] Larsen, *Particle Panic!* (n 265) 105.
[1175] Pallab Ghosh, 'Shock Result in Particle Experiment Could Spark Physics Revolution' *BBC* (7 April 2022) <https://www.bbc.co.uk/news/science-environment-60993523> accessed 04 May 2023.
[1176] Miscounting human chromosome number and estimates of the age of the Earth are given as examples. Ord, Hillerbrand and Sandberg, 'Probing the Improbable' (n 1156) 198.
[1177] Chapter 5.
[1178] "[T]here could be existential risks associated with certain high-powered physics experiments that require a theory X to specify, but understanding theory X requires a series of concepts A, B and C that lie outside our cognitive space…." Torres, *Morality, Foresight, and Human Flourishing* (n 17) 89-90.

risk-related research can legitimise the view that the policymakers should expand their timeframe and accept that there may be potential limits on humanity's current insights. It might be recalled that Chapter 5 considered the ASP. Among other things, this concept denotes that the present generations only occupy a brief period in humanity's overall history. For example, a 100-year period is relatively brief when giving weight to humanity's long-term future. So too is 1,000 years. The claim that a 100-year or 1,000-year period is brief is relative. It is relative to the vast time scales that should be given weight if one accepts the ASP and the ideas that sway x-risk research and mitigation.

How, then, can these observations support analysis of the first deficiency factor? In particular, the observations can affect the interpretation and understanding of concepts such as longevity, time, newness and so on. These concepts may all be employed when analysing the first deficiency factor. Naturally, there may be disagreement about the exact interpretation of these concepts and how much the ASP should influence them. However, giving weight to the ASP and accepting x-risk-related observations can make the first analytical approach more potent. Consider, for example Johnson's reference to the test of time:

> Even if accepted at first by the scientific community, there would seem to be no guarantee that its layers of argument will stand the test of time—despite the unequivocal nature in which such arguments have been posited.[1179]

Johnson refers to the "layers of argument" within some of the papers supporting the safety of the LHC. The broader point here is that the ASP and x-risk-related observations is relevant to and can sway the notion of anything resembling a "test of time". Any test of time for *existential* risk mitigation must clearly differ from the same test for any other risk not threatening to foreclose humanity's future.

Of course, analysis of defective theoretical groundings is not about the mere passage of time. It is about many associated factors – such as the durability of past safety arguments, the gravity of prior surprises within the field and the speed at which new

[1179] Johnson, 'The Black Hole Case' (n 68) 890.

ideas are exchanged – all of which relate to the wider argument that the science underlying the safety assurances is potentially flawed.

This section will be rounded off by two quotes from Toby Ord. The first one comes from a paragraph on the blog *Practical Ethics*, which is connected to the Philosophy Faculty at Oxford University. It reads:

> The problem is that the calculations [concerning the probability that the LHC will destroy humanity's future] don't consider that the physical theories they are using could themselves be incorrect. For example, a hundred and twenty years ago, the scientific consensus held that Newtonian mechanics was the ultimate physical theory. If they had to calculate the chance that an experiment could lead to the curving of space and time, they would have said there was no chance at all. Indeed they would have also calculated that there was no chance of modern electronics or lasers existing, since both are impossible classically. They would have been at least as certain of this as the directors of the LHC are, and they would have gotten it wrong. We could be in just such a situation and with the highest possible stakes at risk.[1180]

The second quote is from Ord's more recent book, *the Precipice*, where he states:

> In some cases, scientists confidently assert that it is impossible for the experiment to cause a disaster or extinction. But even core scientific certainties have been wrong before: for example, that objects have determinate locations, that space obeys Euclid's axioms, and that atoms can't be subdivided, created or destroyed. If pressed, the scientists would clarify that they really mean it couldn't happen without a major change to our scientific theories. This is sufficient certainty from the usual perspective of seeking accurate knowledge, where 99.9 percent certainty is more than enough. But that is a standard which is independent of the stakes. Here the stakes are uniquely high and we need a standard that is sensitive to this.[1181]

In short, the reliability of the x-risk assessment can reasonably be questioned by examining whether the assessment might rely on defective theoretical groundings. As Ord's quotes indicate, there is no reason to uncritically accept that the current state of knowledge is absolute, especially not when a postulated risk threatens to destroy humanity's future. Instead, the present state of knowledge and current scientific

[1180] Toby Ord, 'These are not the Probabilities you are Looking For' (*Practical Ethics*, 15 April 2008 <http://blog.practicalethics.ox.ac.uk/2008/04/these-are-not-the-probabilities-you-are-looking-for/> accessed 04 May 2023.
[1181] Ord, *Precipice* (n 16) 161.

theories might be subject to change as new insights come to light over time. The deficiency factor outlined in this Section can allow an external reviewer to undertake a principled review. Some relevant aspects were noted above, such as the newness of the relevant theory underpinning the x-risk assessment and the pace at which new and old ideas are introduced and challenged, respectively. For example, in relation to the LHC Controversy, an external reviewer would have been in a good position to note the "…flurry of papers confirming and disputing the results…"[1182] as a relevant indication the relevant risk assessment might not have been as reliable as it should have been given the astronomical stakes.

6.3.2 Reviewing the Science: Faulty Scientific Work

In the context of Johnson's work, the second analytical method invites a generalist judge to give weight to the potential for smaller-scale errors that may have caused the risk assessors to reach inaccurate conclusions.[1183] Whereas defective theoretical groundings is all about big picture flaws, 'faulty scientific work' is about micro-scale errors, including miscalculations, relying on flawed arguments and the like. Analysis of this second deficiency factor involves appreciation of the possibility that the data the scientists (and risk assessors) rely upon is ultimately inaccurate.[1184]

Within the remits of the LHC-inspired intervention, the second deficiency factor also falls within the remits of the external reviewer analysing whether the risk assessors might have based their conclusion on flawed information or an incomplete state of knowledge or understanding.

"The point of considering the potential for small-scale error in scientific work…" writes Johnson, "…is understanding that scientists' conclusions about risk cannot be taken at face value. Their assurances must be discounted by the probability of their own error."[1185] He appreciates that critics may perceive this line of scrutiny as unfair.[1186] If

[1182] Larsen, *Particle Panic!* (n 265) 101.
[1183] Johnson, 'The Black Hole Case' (n 68) 890; Johnson, 'Uncertain Risk, Science Experiments, and the Courts' (n 68) 371.
[1184] See Johnson, 'The Black Hole Case' (n 68) 890.
[1185] ibid 891.
[1186] ibid.

X claims that Y may have made a mistake and that it is, therefore, sensible to discount the value of Y's assurance, surely it is X's responsibility to expose Y's mistake. This view is undoubtedly reasonable when Y's work is confined to academia or the potential mistake cannot contribute to any harm or damage. However, this stance is not necessarily reasonable in all situations. For example:

> [T]he issue of LHC risk is not a purely academic debate—and the same rules do not apply. In the context of a policy debate or legal dispute where scientific arguments are used to justify a real-world course of action that is allegedly dangerous, it is improper to take all scientific arguments on their own terms without subjecting them to a higher level of scrutiny.[1187]

Johnson categorises the notion of Faulty Scientific Work into groups, including flawed arguments, flawed calculations and errors in models, assumptions, and conceptual thinking.[1188] Some of these draw a lot of inspiration from a paper that will be considered in a lot of detail momentarily.[1189]

Ultimately, an external reviewer may give weight to the possibility that a relevant x-risk assessment contains fundamental errors that weaken the conclusion that an endeavour is safe. For instance, in his 2021 article, Johnson suggests that relevant considerations include the layers of assumptions employed, the complexity of the calculations and the sources of data.[1190]

It is possible to remain sceptical that analysis of the second deficiency factor will lead to anything concrete. For example, searching and accounting for potential errors in scientific work may become a circular task.[1191] Upon having undertaken the first round of analysis (or accepting the possibility that there may be hidden flaws), the reviewer may end up exactly where it started. That is, how can it be ascertained that there are no lasting flaws or errors? In turn, should the reviewer continuously discount the subsequent assurance(s) by the chance that there may be other hidden flaws? A circular analysis will devalue and undermine all assurances and x-risk assessments.

[1187] ibid 891.
[1188] ibid 892-896.
[1189] Ord, Hillerbrand and Sandberg, 'Probing the Improbable' (n 1156).
[1190] Johnson, 'Uncertain Risk, Science Experiments, and the Courts' (n 68) 373.
[1191] Häggström, *Here Be Dragons* (n 22) 187-188 and 198-201.

Another critique is that giving weight to potential mistakes can lead to an unsustainable precautionary approach. As written by Häggström: "Should I use my kitchen knife to sharpen my pencil? The laws of physics as we know them suggest that such an action involves no risk of destroying our planet, but what if they are badly flawed?"[1192] While Häggström raises a legitimate concern, the argument as expressed with reference to the pencil example is perhaps exaggerated. For one, and in contrast to high-energy accelerators, there are, as noted by some scholars, no credible mechanisms for destroying the world by sharpening one's pencil with a kitchen knife.[1193] Those worried about the LHC destroying Earth (and humanity's future) could at least, as noted by Sandberg, "…point to papers suggesting that strangelets, small black holes and vacuum decay were theoretically possible."[1194] Furthermore, it is possible to note that some actions, such as using a knife to sharpen a pencil, dropping the pencil or knife, or simply breathing, do not come close enough to pushing outside the "envelope of nature"[1195] to merit serious concern. Finally, there is a distinction between looking for and giving weight to potential flaws on a whim and doing the same in relation to a specific x-risk assessment. There are also some practical considerations. Some activities – such as using a knife in the manner described by Häggström – cannot reasonably be suppressed in any event.

The notion of 'Faulty Scientific Work' in Johnson's paper draws much inspiration from a (then) draft paper by Ord, Hillerbrand and Sandberg. We can even suggest that what the second deficiency factor deals with ought to be called the "Ord-Hillerbrand-Sandberg-style analysis."[1196] Their paper – *Probing the Improbable: Methodological Challenges for Risks with Low Probabilities and High Stakes* – has been published in the Journal of Risk Research[1197] since being cited in the Black Hole Case Article. The

[1192] ibid 200.
[1193] Ord, Hillerbrand and Sandberg, 'Probing the Improbable' (n 1156) 203.
[1194] Anders Sandberg, 'Cool Risks Outside the Envelope of Nature' (*Andart II*, 22 October 2014) <https://aleph.se/andart2/risk/existential-risk-risk/cool-risks-outside-the-envelope-of-nature/> accessed 04 May 2023.
[1195] As written by Sandberg: "[I]f you do something that is within the envelope of what happens in the universe normally and there are no observed super-dangerous processes linked to it, then this activity is likely fine." ibid.
[1196] This term is borrowed from Häggström, *Here Be Dragons* (n 22) 200.
[1197] Ord, Hillerbrand and Sandberg, 'Probing the Improbable' (n 1156).

thesis has previously come across Ord and Sandberg many times, and scholars within the x-risk circle have echoed the assertion from their 2010 paper.[1198]

What, then, is their article saying? In short, "[t]he basic message of this paper…" write Ord, Hillerbrand and Sandberg "…is that any scientific risk assessment is only able to give us the probability of a hazard occurring conditioned on the correctness of its main argument."[1199] In effect, the relevant risk assessment will not reveal the actual probability of the hazard – e.g., existential catastrophe – occurring.[1200] Instead, the assessment only conveys the probability of the existential catastrophe occurring, provided that the assessment's argument is correct (i.e., it does not contain any critical flaws).[1201] But as mentioned above, the assessment may contain hidden flaws, and this might not necessarily be so uncommon.[1202] As such, when dealing with low-probability risks—such as the vast majority of risks within the x-risk landscape—the probability that the assessment contains a major flaw may actually be much greater than the stipulated probability of the catastrophe occurring.[1203] The above can be clarified with reference to an example.

Presume that a group of scientists have been enlisted to assess the probability that endeavour Z will cause an x-catastrophe. After thorough investigation, the scientists conclude their report with an estimate that the (subjective) probability of an existential catastrophe from Z is one in a trillion.[1204] This estimate is comforting at face value and most individuals will likely agree that this is properly classified as a low-probability risk. However, the probability that the scientists have reached an incorrect estimate – due to, e.g., miscalculations, flawed arguments, and unsound theory – is potentially much greater than one in a trillion. The risk of x-catastrophe may thus be a lot higher than

[1198] Bostrom, 'Existential Risk Prevention' (n 15) 16; Bostrom, 'The Vulnerable World Hypothesis' (n 465) 462; Ord, *Precipice* (n 16) 198-199; Torres, *Morality, Foresight, and Human Flourishing* (n 17) 89; Häggström, *Here Be Dragons* (n 22) 187-188 and 199-201.
[1199] Ord, Hillerbrand and Sandberg, 'Probing the Improbable' (n 1156) 203.
[1200] ibid 192.
[1201] ibid.
[1202] Ord, Hillerbrand and Sandberg highlight that flawed arguments are rather common, and that this can be assessed by, inter alia, consider the proportion of academic papers that, after publication, are retracted. ibid 193-194.
[1203] ibid 192.
[1204] Ord, Hillerbrand and Sandberg use one in a billion as an example. ibid 192.

the given estimate. Thus, it might be sensible for the external reviewer to "…adjust for this by using a higher number…"[1205] than the one in the scientists' initial report.

As highlighted by Ord, Hillerbrand and Sandberg, to calculate the true probability of catastrophe, the following equation should be used:

$$P(X)=P(X|A)P(A)+P(X|\neg A)P(\neg A).$$

This formula is clarified in the following footnote.[1206] In short, to calculate the true probability of catastrophe, it is necessary to estimate the probability that the (i) scientists' argument is sound (this is, of course, not easy) and (ii) catastrophe occurs should the scientists' argument not be sound (again, this is not easy).[1207]

It is useful to reiterate that the core idea to note is that the scientists' estimates should not be taken at face value because the probability that their estimate is inaccurate – owing to hidden flaws – might be much greater than the reported probability of the x-catastrophe. In effect, giving weight to and analysing the potential for micro-scale errors is not an empty prospect. In fact, there are examples of flaws in the argument that was crucial in the assessment considering the safety of RHIC.[1208] This is highlighted in further below in this Chapter.

Previous Chapters have noted that supporters claim that experimental particle physics poses no x-risk whatsoever. They claim that the occurrence of the x-catastrophe is *virtually* impossible. Recall Baum's equation by which this can be expressed like this: $P_1=0$. Nevertheless, an argument of 'impossibility' (whether in relation to experimental particle physics or beyond) does not render the Ord-Hillerbrand-Sandberg-style

[1205] Ord, *Precipice* (n 16) 199.

[1206] In this equation, X represents the catastrophe occurring; A represents the scientists' argument being sound. We can read "|" to mean "given". I.e., $P(X|A)$ is the probability of X given A. Notably, and as explained above, this is what the initial report gives us. I.e., the probability of the catastrophe occurring provided that the argument supporting the report is sound. We can read "¬" to mean "the argument is unsound". I.e., $P(X|\neg A)$ is the probability of the catastrophe occurring, given that the argument is unsound. In effect, to get the 'true' P(X), we should consider the probability that the argument is sound and the probability of catastrophe, given that the argument is not sound. ibid 192-193.

[1207] ibid.

[1208] ibid 200.

analysis inapplicable. This is because a claim that the catastrophe is impossible might be based on a flawed argument just as a claim that the probability of catastrophe is, for instance, one in a trillion.[1209] Hence, the impossibility claim does no more than suggesting that the occurrence of the x-catastrophe is impossible, provided that the argument supporting the claim is sound ($P(X|A)$).

Ord, Hillerbrand and Sandberg recognise that (i) their analysis is more useful for low-probability risks, and (ii) the stakes must be "very high" to warrant the additional analysis.[1210] Whereas all low-probability risks will be subject to the methodological issue above, the stakes will arguably dictate whether the Ord-Hillerbrand-Sandberg-style analysis is justified.[1211] Certainly, the level of precaution that their analysis speaks to can be rather extreme. On the point of the stakes, the authors briefly explore some ideas inherent within the ASP. Mentioning the destruction of the environment, death of all living humans, and the loss of all generations yet to come, they write: "It is worth noting that [the] loss of all future generations (and with it, all of humanity's potential) may well be the greatest of the three, but a comprehensive assessment of these stakes is outside the scope of this paper."[1212] Chapter 5 dealt with this aspect in some detail and argued that the destruction of humanity's future is certainly the aspect which renders x-catastrophes uniquely bad compared to non-extinction catastrophes.

Ord, Hillerbrand and Sandberg use the risk evaluations from the LHC Controversy as a test case to apply their analysis to the alleged x-risks associated with experimental particle physics.[1213] They write:

> While the arguments for the safety of the LHC are commendable for their thoroughness, they are not infallible. Although the report considered several possible physical theories, it is eminently possible that these are all inadequate representations of the underlying physical reality. It is also possible that the models of processes in the LHC or the astronomical processes appealed to in the cosmic ray argument are flawed in an important way. Finally, it is possible that there is a calculation error in the report.[1214]

[1209] ibid 193.
[1210] ibid 194.
[1211] ibid; Ord, *Precipice* (n 16) 396.
[1212] Ord, Hillerbrand and Sandberg, 'Probing the Improbable' (n 1156) 201.
[1213] ibid 198-202.
[1214] ibid 201.

Ord, Hillerbrand and Sandberg go on to say that "…the current safety report should not be the final word in the safety assessment of the LHC."[1215] The authors also suggest that the lessons from the LHC test case should be drawn upon when other high-stakes low-probability risks are assessed.[1216] Their paper ends with some "very general remarks" on how argument flaws might be avoided when assessing risks with high stakes.[1217] Particularly relevant for this thesis, they write:

> R]eproducibility appears to be the most effective way of removing many of these errors. By having other people replicate the results of calculations independently, our confidence in them can be dramatically increased. By having other theories and models *independently* predict the same risk probability, our confidence in them can again be increased, as even if one of the arguments is wrong the others will remain. Finally, we can reduce the possibility of unconscious bias in risk assessment through the simple expedient of splitting the assessment into a 'blue' team of experts attempting to make an objective risk assessment and a 'red' team of devil's advocates attempting to demonstrate a risk, followed by repeated turns of mutual criticism and updates of the models and estimates (Calogero 2000). Application of such methods could in many cases reduce the probability of error by several orders of magnitude.[1218]

This thesis will return to the notion of unconscious bias below. However, before doing so, it is necessary to consider how the notion of the deficiency factor considering in this Section relates to x-risk research and vice versa.

Firstly, Johnson's argument that a generalist court should analyse the potential for Faulty Scientific Work is, as noted above, derivative of insights uncovered by scholars within the x-risk scene (i.e., Ord, Hillerbrand and Sandberg). In effect, it is possible to view an invitation to scrutinise the second deficiency factor as giving weight, at law, to the Ord-Hillerbrand-Sandberg-style analysis. Furthermore, x-risk scholars continue to highlight the applicability of the argument provided by Ord, Hillerbrand and Sandberg. This means that existential risk research can justify and support legal interventions giving weight to potential small-scale errors in scientific work or x-risk assessments.

[1215] ibid 202.
[1216] ibid.
[1217] ibid 203.
[1218] ibid.

Secondly, like the first deficiency factor, analysing the potential for Faulty Scientific Work holds a strong connection to the passage of time. This is because time may subject the relevant argument or x-risk assessment to a 'natural' Ord-Hillerbrand-Sandberg-style analysis. That is, time will enable more people to independently consider and possibly reveal enduring flaws in a relevant x-risk assessment or supporting paper. It is reasonable to believe that there is a higher chance of uncovering critical flaws if a paper or x-risk assessment is available and subject to, for example, ten years of open scrutiny than if the same is only available and subject to one year of scrutiny. On this note, Chapter 3 emphasised that the claimants in the *Sancho complaint* sought to secure additional time for just such a purpose. Paragraph 15 in the complaint reads:

> Plaintiffs and their associates are…capable of reviewing and analyzing such safety reviews for flaws or errors. Plaintiffs and some of their associates have filed in support of this complaint various affidavits detailing some of the safety flaws and ethical flaws in safety review currently evidenced. Plaintiffs and their associates require a minimum of four to six months time to review the *LSAG Safety Review*, as well as the relevant scientific literature, in order to determine whether defendants' most recent pending *LSAG Safety Review* is once against fundamentally flawed…[1219]

Whereas it is possible to doubt the claimants' ability to undertake this exercise, their underlying ambition to check for and uncover fundamental mistakes is not unreasonable. This is precisely the point discussed exposed above.

In 2004, Kent noted calculation mistakes in some papers dealing with the safety of RHIC.[1220] As examined later, Tegmark and Bostrom also uncovered a flaw in a significant paper speaking to the safety of RHIC.[1221] Findings like this can reinforce the belief that a relevant x-risk assessment is not, at present, sufficiently reliable on the basis that it might contain (other critical) flaws.

[1219] *Sancho complaint* (n 460) 5.
[1220] Kent, 'A Critical Look' (n 156).
[1221] Max Tegmark and Nick Bostrom, 'How Unlikely is a Doomsday Catastrophe?' [2005] Nature 754, 754. For an extended version of this paper see <https://arxiv.org/abs/astro-ph/0512204> accessed 04 May 2023.

It is useful to repeat that the current argument is that analysis of the second deficiency factor is linked to the passage of time. As with the first deficiency factor above, the ASP also speaks to the interpretation of time in relation to the second factor. In particular, it can sway how many years a given paper, report or x-risk assessment should be subject to a 'natural' Ord-Hillerbrand-Sandberg-style analysis. This natural process can be supplemented with an orchestrated analysis whereby selected experts or qualified persons are – in line with Calogero's suggestion – allocated to a "blue" and "red" team with the ultimate goal of finding and refuting flaws. Ultimately, if critical flaws are uncovered relatively frequently[1222] following such orchestrated exercise or more naturally, then there is merit in not pushing ahead too quickly with the endeavour.

Thirdly, due to their nature, risks within the x-risk landscape are suitable for the analysis promoted by Ord, Hillerbrand and Sandberg. These risks are generally classified as low-probability. In addition, they have astronomically high stakes. Whereas both features are relevant, the thesis will concentrate on the stakes below. The stakes (and the ASP) can justify analysis of the second deficiency factor and the high level of precaution flowing from the Ord-Hillerbrand-Sandberg-style analysis. Beyond justifying the general remits and cautionary approach of this analytical exercise, the stakes (and ASP) can also influence the more nuanced application of the analysis. For example, the ASP can speak to the interpretation of time, which, in turn, relates to how long it is reasonable to subject a paper or assessment to a natural or orchestrated Ord-Hillerbrand-Sandberg-style analysis. As another example, the ASP can also influence the "…greatest acceptable value of [$P(X)$]…"[1223] following the analysis that has taken us one step closer to the true probability. Of course, a critic may argue that it is justifiable to look at the second deficiency factor without reference to the ASP. In other words, there will likely be those who think the Ord-Hillerbrand-Sandberg-style analysis is justifiable when a risk threatens to kill, say, 1,000 persons. While this view is reasonable, this line of thinking has been addressed in Chapter 5.

In sum, an external review can reasonably evaluate the potential that the x-risk assessment is based on flawed information or an incomplete state of knowledge or

[1222] I leave open the question of what constitute relatively frequently in this context.
[1223] Häggström, *Here Be Dragons* (n 22) 200.

understanding. For instance, calculation mistakes were noted by Kent in the context of the LHC Controversy. In addition, as seen further below, Tegmark and Bostrom uncovered a flaw grounded in the observation selection effect. Had these mistakes been teased out and discovered in time, an external reviewer might have been in a better position to question the reliability of the x-risk assessment.

6.3.3 Reviewing the scientists and their organisation: Credulity and Neglect

The third and fourth deficiency factor is concerned with the scientists (i.e., the risk assessors) rather than the scientific work (i.e., the risk assessment). In other words, this involves analysis of the actors who consider the relevant x-risk.

Despite being about the assessors, the thesis contends that the third and fourth deficiency factor predominantly operates to feed into the broader concern that the scientific work assessing the relevant risk is not, at present, as reliable as it ought to be. In other words, it is argued that the material value in reviewing the scientists (and their organisations) lies in elevating potential concerns about the scientific work.

As for Credulity and Neglect in the context of Johnson's article, when considering whether there are serious questions for the sake of an interim injunction request, a generalist judge may assess "…to what extent there is a possibility for scientists to be too ready to believe a rosy view of potential dangers (credulity) and to not properly pay attention to signals that safety has not been adequately assured (neglect)."[1224]

In essence, the third factor that an external reviewer can analyse as part of assessing the scientific work's reliability relates to psychological and sociological factors.[1225] It is useful to divide the exposition between psychological and social factors.

Considering the former, scientists are not immune to irrationality that can cloud and affect their judgment when it comes to risk. Johnson's articles highlight how different

[1224] Johnson, 'Uncertain Risk, Science Experiments, and the Courts' (n 68) 373.
[1225] ibid; Johnson, 'The Black Hole Case' (n 68) 897.

cognitive biases – including confirmation bias,[1226] myopia bias[1227] and availability bias[1228] – may lead to credulity and neglect. The thesis has previously outlined that different psychological factors are relevant in influencing how we think about x-risk.[1229]

As for social factors, an organisation's culture and practices and the scientific community's broader norms, are all examples of elements that may affect the assessors' judgments about the risks associated with their endeavour.[1230] For example, it is possible to consider whether the scientists can freely vent safety concerns, whether the organisation is under tight deadlines which may compromise the quality of the x-risk assurances, and whether the organisation and its people are open to insights from outsiders. The works by several scholars dealing with the LHC Controversy can lend support to the proposition that these aspects merited consideration vis-à-vis Brookhaven National Laboratory and CERN.[1231]

It is possible to find at least one immediate and clear connection between x-risk research and the third general deficiency factor. This connection is manifested in the writings of some x-risk scholars who highlight the relevance of unconscious biases when thinking about and assessing existential risk.

As noted by Ord, the scientist's given subjective probability of catastrophe from an experiment – even if stipulated to be very low – may be too high, "[e]specially when there are biases and selection effects leading even conscientious scientists to systematically underestimate the risks…."[1232] In addition, the thesis has illustrated that

[1226] This bias can lead to the inclination to process information to support one's pre-existing beliefs and hypotheses. ibid 899-901; Yudkowsky, 'Cognitive Biases' (n 372) 98-99.
[1227] This bias can lead to the tendency and inclination to undervalue the future, favouring short-term rewards. Johnson, 'Uncertain Risk, Science Experiments, and the Courts' (n 68) 373.
[1228] This bias can lead to the tendency and inclination to believe that events that are more easy to recall are more probable and vice versa. ibid.
[1229] See Chapter 2.
[1230] Johnson, 'Uncertain Risk, Science Experiments, and the Courts' (n 68) 374; Johnson, 'The Black Hole Case' (n 68) 897-904.
[1231] Posner, *Catastrophe* (n 90); Kent, 'A Critical Look' (n 156); Wilson, 'Minimizing' (n 151); Peterson, 'The "Sancho" Effect' (n 468); Calogero, 'Might a Laboratory Experiment Destroy Planet Earth?' (n 446); Johnson, 'The Black Hole Case' (n 68); Johnson, 'Agencies and Science-Experiment Risk' (n 68); Johnson, 'Uncertain Risk, Science Experiments, and the Courts' (n 68); Johnson, 'Judicial Review' (n 68).
[1232] Ord, *Precipice* (n 16) 377.

many authors having considered the LHC Controversy highlight the problem of lack of objectivity in the risk assessment activities.

The upcoming section will outline the final deficiency factor that an external reviewer may seek to analyse as part of assessing the reliability of scientific work. After that, the thesis will consider the synergy between the deficiency factors and how this can ultimately contribute towards a broad legal intervention for x-risk mitigation.

6.3.4 Reviewing the Scientists and Their Organisation: Bias and Influence

In the context of Johnson's article, Bias and Influence concerns the potential for a conflict of interest and lack of independence amongst those who vouch for an endeavour's safety.[1233] In his 2021 article, Johnson writes: "The relevant questions here revolve around to what extent the risk assessors are independent of the organizations and scientific communities whose risk is being assessed."[1234] Taking this line of analysis to the courtroom as proposed by Johnson, a generalist judge can analyse whether "serious questions" exist for the purpose of an interim injunction request on the basis that, for instance, the risk assessment work was produced by persons (directly or indirectly) affiliated with the organisation seeking to go ahead with the project. A judge may also wish to consider to what extent the report's conclusion can affect the assessors' own research agenda and output, career prospects and status within the scientific community.[1235]

The fourth deficiency factor that an external reviewer can analyse as part of assessing the reliability of a scientific work concerns the potential existence of a conflict of interest and lack of independence amongst the risk assessors. Such conflict of interest is troubling because it creates a risk of bias in the risk assessment process. If an organisation and its scientists have a preferred outcome – such as the finding that a given project is safe – then it is desirable to ensure that those scientists play little to no role in assessing the safety of that project. In short, it is appropriate to ensure that

[1233] Johnson, 'The Black Hole Case' (n 68) 904-907; Johnson, 'Uncertain Risk, Science Experiments, and the Courts' (n 68) 374-375.
[1234] ibid 374.
[1235] ibid 374-375; Johnson, 'The Black Hole Case' (n 68) 904-907.

persons with no personal interest or stake in the project's future undertake the relevant safety and risk assessment process. Such personal interest or stake may take many forms, such as career and financial aspirations, a search for prestige and intellectual curiosity, and pressure to supply the answer that the community expects one to find.

In his article on the "killer strangelet" scenario, Kent writes that debates concerning hypothetical risky experiments should involve "…experts with no stake in the experiments under consideration."[1236] Beyond the works produced by Johnson – one of which reads that "[t]he public record of the [LHC] controversy reveals for plaintiffs an embarrassment of riches in arguments to show bias among the assessors…"[1237] – many scholars have voiced concern about the lack of objectivity in the risk assessment process and works that speak to the safety of the RHIC and LHC experiments.[1238] Of course, it is possible to highlight here that the only persons capable of reviewing the safety of the experiments (i.e., the physicists) are the very same people who have a stake in ensuring the future of experimental particle physics.

The final aspect to consider is how analysis of the fourth deficiency factor relates to x-risk and vice versa. There is arguably no obvious connection between this deficiency factor and x-risk research apart from the general contention that those assessing existential risk from any given endeavour should not have any personal stake in the project's future. As stated in *the Precipice*, one of the governance problems vis-à-vis risks from experiments creating unprecedented conditions relates to "…various biases and conflicts of interest, where the very people whose jobs (or whose colleagues' jobs) depend on a verdict are responsible for deciding that verdict."[1239]

Part 6.4 will collate the sections above and argue that there is a synergy between the deficiency factors and that the value of this synergy is to elevate a potential concern

[1236] Kent, 'A Critical Look' (n 156) 157.
[1237] Johnson, 'The Black Hole Case' (n 68) 904.
[1238] Posner, *Catastrophe* (n 90); Rees, *Our Final Century* (n 226); Bostrom and Ćirković, 'Introduction' (n 12); Calogero, 'Might a Laboratory Experiment Destroy Planet Earth?' (n 446); Lehmann, *No Canary in the Quanta* (n 94); Peterson, 'The "Sancho" Effect' (n 468); Wilson, 'Minimizing' (n 151); Adams, 'Honey' (n 309); Ord, Hillerbrand and Sandberg, 'Probing the Improbable' (n 1156); Ord, *Precipice* (n 16) 377-378.
[1239] ibid 378.

that the work assessing an x-risk is, at present, not as reliable as it ought to be. There are two prongs to this claim. The first one concerns *reliability*. Part 6.4 will deal with that. The second prong concerns the interpretation of *at present*, and Part 6.5 consider this point in more detail.

6.4 RELIABILITY

The deficiency factors can be assessed in tandem to give credence to a concern that the scientific work assessing x-risk linked to an endeavour is not, at present, as reliable as it should be. It is suggested that there is a natural synergy between the deficiency factors and that the value of analysing the third and fourth factor lies in supporting an analysis of the first and second deficiency factor.

For example, consider how the third deficiency factor intersects with the first and second one. As mentioned before, the third one concerns the potential for innocent mistakes flowing from psychological and social factors. Relevant here is the potential for cognitive biases that can strengthen a concern that the x-risk assessment is not as reliable as it should be. For instance, the potential for *confirmation bias* may lead the scientists and assessors to fall victim to the tendency to search for information that supports their finding that the project is safe. In turn, this can impact the quality and reliability of the scientific work and assessment. Confirmation bias is said to be even more problematic in cases of a 'hot' (as opposed to a 'cold') form of confirmation bias,[1240] such as in the LHC Controversy context where the physicists are personally invested in the future of experimental particle physics. In effect, they may ignore or undervalue relevant information as well as omit to give weight to past (and potential future) upsets within the scientific field. Another example revisited below in relation to the LHC and RHIC is *observation-selection bias*, whereby the assessors rely on, among other things, Earth's continued existence to underestimate the probability of dangerous outcomes or to exclude them altogether.

[1240] A hot form of confirmation bias refers to situations where a person's belief is emotionally charged. A cold form refers to situations where the belief is logical. See Yudkowsky, 'Cognitive Biases' (n 372) 99.

These, and other biases, such as *expectation bias*,[1241] *overconfidence bias*,[1242] *availability bias*,[1243] *bystander effect*,[1244] *myopia bias*,[1245] and *present bias*,[1246] can seemingly more or less inadvertently result in (i) the assessors omitting to be mindful of the limits of humanity's current knowledge and our brief and early place in humanity's overall history, and (ii) the x-risk assessment being more likely to contain hidden flaws and erroneous probability estimates. It is to be appreciated that (i) is relevant for the first deficiency factor (big picture flaws) and (ii) is relevant for the second (smaller errors such as miscalculations).

In short, various cognitive biases can affect the reliability of the x-risk assessment. Contrariwise, the scientific work itself can possibly influence the potential for cognitive biases. It is reasonable to believe that the more complex and layered the x-risk assessment and the more scope for the scientists' own judgements, the higher the chance that biases can affect those judgments. On the other hand, straightforward arguments in favour of an endeavour's safety will arguably be "…more resistance to [unconscious] biases…"[1247] Indeed, it is difficult to imagine that unconscious biases, no matter how strong in effect, can sway any assessor or scientist to believe that 20 / 2 = 25 or that life on Earth began a decade ago.

The potential for unconscious biases is not the only relevant aspect when analysing the third deficiency factor. Instead, the assessors' organisation is a target for

---

[1241] "Particularly in these cases, having an expected or desired outcome in mind while setting up a model makes one vulnerable to expectation bias: the tendency to reach the desired answer rather than the correct one." See Ord, Hillerbrand and Sandberg, 'Probing the Improbable' (n 1156) 198.

[1242] *Overconfidence* involves persons subjectively believing their judgment is better than the objective actuality. See Torres, *Morality, Foresight, and Human Flourishing* (n 17).

[1243] This is a tendency where people rely too much on information they can more easily recall. This can affect judgements about the likelihood of risk. See Yudkowsky, 'Cognitive Biases' (n 372) 92-93; Sunstein, *Averting Catastrophe* (n 281) 10.

[1244] This is the tendency where larger groups of persons are less likely to respond to and act in emergencies individually and collectively. "Being part of a group reduces individual responsibility." See Yudkowsky, 'Cognitive Biases' (n 372) 109-111.

[1245] This is the tendency of people place more value on the prevention of immediate losses than on the avoidance of future losses. See David Dana, 'A Behavioral Economic Defense of the Precautionary Principle' [2003] Northwestern University Law Review 1315, 1324-1326.

[1246] "Many people suffer from "present bias": they focus on today and tomorrow, and not on the long term." Sunstein, *Averting Catastrophe* (n 281) 17.

[1247] "…No amount of bias or cultural filtering would, for example, cause someone to believe 2 + 2 = 5." See Johnson, 'Uncertain Risk, Science Experiments, and the Courts' (n 68) 374.

examination too. The organisation's culture, traits and practices may display weaknesses that can impact the reliability of the x-risk assessment. Organisational weaknesses can arguably operate independently (e.g., by punishing whistle-blowers) or in conjunction with unconscious biases (e.g., by strengthening overconfidence bias and confirmation bias). To demonstrate the potential role played by organisational traits, the findings of the Columbia Accident Investigation Board concerning the space shuttle *Columbia* disaster can be considered.[1248] The Report devotes an entire chapter to the organisational causes of the accident and suggests that NASA's organisational culture and structure were as much to blame for the accident as the physical failures leading to the destruction of the Space Shuttle.[1249] For example, it reads:

> The organizational causes of this accident are rooted in the Space Shuttle Program's history and culture, including the original compromises that were required to gain approval for the Shuttle, subsequent years of resource constraints, fluctuating priorities, schedule pressures…Cultural traits and organizational practices detrimental to safety were allowed to develop, including: reliance on past success as a substitute for sound engineering practices (such as testing to understand why systems were not performing in accordance with requirements); organizational barriers that prevented effective communication of critical safety information and stifled professional differences of opinion; lack of integrated management across program elements; and the evolution of an informal chain of command and decision-making processes that operated outside the organization's rules.[1250]

The Columbia Accident Investigation Board records that the employees were overconfident[1251] and that unjust optimistic organisational thinking affected the decision-making.[1252] Another observation is the reported shortcomings of effective communication of safety issues and taking into account minority opinions.[1253] For example, the Board emphasises how organisations with a strong safety culture "…generally acknowledge that a leader's best response to unanimous consent is to play devil's advocate and encourage an exhaustive debate."[1254] In contrast, NASA's

[1248] The Columbia Accident Investigation Board, 'The Columbia Accident Investigation Board Report' (volume 1, 2003) <https://history.nasa.gov/columbia/CAIB_reportindex.html> accessed 07 May 2023.
[1249] ibid 177.
[1250] ibid, 9.
[1251] ibid, 177 and 192.
[1252] ibid 177 and 181.
[1253] ibid 192.
[1254] ibid.

organisational culture encouraged silence and a pattern of ignoring minority opinions.[1255] Looking at the Board's quote, one might recall Calogero's discussion regarding the value of having a blue and red team trying to prove/disprove that certain experiments – such as those facilitated by the RHIC and LHC – are dangerous.[1256]

Turning back to the matter at hand, the organisational culture and structure can affect the reliability of the risk assessment. One can also believe that different organisational cultures can foster some of the cognitive biases that have been mentioned above. For example, *overconfidence bias*, *confirmation bias* and *bystander effect* can seemingly flourish in an organisation that facilitates a culture like the one that the Columbia Accident Investigation Board uncovered in their Report.

To summarise, analysis of the third deficiency factor can cut to the reliability of the x-risk assessment showing the natural synergy between the third and the first and second deficiency factors.

As another example, there is a synergy between the fourth deficiency factor and the third. The fourth factor largely concerns the existence of a conflict of interest. The point is that the existence of such a conflict may intensify, for example, the potential for unconscious biases and render the effect of those biases stronger. As noted above, a hot form of confirmation bias is more likely when all risk assessors have a personal stake in the project's future. A hot form of such bias is "…larger in effect and more resistant to change."[1257] This brief example shows the synergy between the fourth and third deficiency factors. As another example, the existence of a conflict of interest also relates to *expectation bias*[1258] and *present bias*.[1259] In essence, it seems reasonable to believe that objectivity amongst the risk assessors can help to guard against the potential for biases and the flourishing of negative organisational traits when it comes

[1255] ibid.
[1256] Calogero, 'Might a Laboratory Experiment Destroy Planet Earth?' (n 446) 193 and 199.
[1257] Yudkowsky, 'Cognitive Biases' (n 372) 99.
[1258] The existence of a conflict of interest will likely affect and intensify the want to reach the desired result (i.e., that the endeavour is safe).
[1259] A conflict of interest will likely affect and intensify the tendency to omit the relevance and value of the future.

to risk assessment. This assertation is supported by Ord, Hillerbrand and Sandberg. In echoing Calogero's 'blue and red team' proposal, they write:

> [W]e can reduce the possibility of unconscious bias in risk assessment through the simple expedient of splitting the assessment into a 'blue' team of experts attempting to make an objective risk assessment and a 'red' team of devil's advocates attempting to demonstrate a risk followed by repeated turns of mutual criticism and updates of the models and estimates…[1260]

In sum, the existence of a conflict of interest may increase the risk of innocent mistakes and amplify the effect of those mistakes. Of course, unconscious biases and organisational traits can affect the reliability of the x-risk assessment without the existence of a conflict of interest. Likewise, the existence of a conflict of interest is problematic irrespective of its bearing on the third deficiency factor. This statement brings us to the more direct synergy between the fourth and the first and second deficiency factors.

A conflict of interest may influence what the risk assessors reveal and whether they intentionally omit to flag certain aspects that may delay or halt their project. A conflict of interest may also influence the assessors' to mainly become concerned with not alarming the public.[1261] It may even lead to attempts to obscure scientific analyses in the spiriting of managing public relations.[1262] Furthermore, in certain extreme situations, a conflict of interest may influence the assessors to fake results and lie to create an impression that their endeavour is safe. One may note that Johnson's 2021 article examines the potential for "Fraud, Lies, and Faked Results".[1263]

The first and second deficiency factors are also interrelated. One will recall that the first one is about macro-scale errors, whereas the second factor is about micro-scale errors. However, the potential for macro-scale errors is arguably an example of a potential hidden flaw that can compromise the x-risk assessment and its underlying

[1260] Ord, Hillerbrand and Sandberg, 'Probing the Improbable' (n 1156) 203.
[1261] See e.g., Calogero, 'Might a Laboratory Experiment Destroy Planet Earth?' (n 446) 201-202.
[1262] ibid.
[1263] Johnson examines the potential for Fraud, Lies, and Faked Results separately from the fourth meta-tool. See Johnson, 'Uncertain Risk, Science Experiments, and the Courts' (n 68) 375-377.

arguments. In other words, if humanity's *current* state of knowledge and understanding of the scientific theory underlying the assessment is defective or incomplete, then the assessment's argument(s) can be viewed as being subject to a hidden flaw.[1264]

Nevertheless, it remains helpful to distinguish the first and second factors. This is because the first one is more intertwined with recognising the limits of humanity's *current* knowledge and its relatively early place in humanity's overall history. In contrast, the second deficiency factor can be viewed as more closely connected to the risk assessment work itself. The distinction can further be outlined by highlighting that analysis of the second deficiency factor can uncover critical flaws even if the scientific theory underlying the x-risk assessment is solid. For example, suppose that theory X underlying safety argument Y is entirely sound. Despite this, analysis of the second deficiency factor can still uncover other hidden flaws in argument Y. Examples include calculation errors, inaccurate probability estimates and, failing to account for observer selection effects when quantifying risk.

To conclude, this Section has suggested that there is a synergy between the deficiency factors. It considered that an external reviewer can analyse these factors in tandem to uncover whether there is reasonable concern about the work's *reliability*. Having considered the idea of *reliability*, the next Part turns to the interpretation of *at present*.[1265]

6.5 AT PRESENT: THE VISTA OF TIME

The notion of *at present* should be interpreted against the backdrop of the ASP. As seen in Chapter 5, the ASP underscores humanity's overall history and the significance of the far future. This concept demands an expansion of one's time frame when dealing with x-risk, and it can support the view that a 100-year period is relatively brief when placed in the context of humanity's long-term future. As such, it is not unreasonable to make a case for delaying some endeavours for, say, 100 years,

[1264] I.e., the flaw springs from the faulty or critically incomplete understanding of the scientific theory.
[1265] Indeed, there are two crucial aspects to the statement that the synergy can elevate a potential concern that the scientific work assessing an existential risk linked to project X is, *at present*, not as *reliable* as it ought to be.

provided that (i) one accepts the ASP, and (ii) there is an underlying justification for the delay in the first place. The underlying justification can flow from a concern about the *reliability* of the x-risk assessment.

There is a good reason for suggesting that the ASP should influence the interpretation of *at present* and the reasonable length of any potential delay. That reason has to do with the deficiency factors and their connection to the passage of time. More specifically, the longer the delay, the more potent these factors become in supporting an investigation on whether the x-risk assessment is or is not reliable. In this way, the deficiency factors can serve as the cause and cure for any potential delay.

For example, the first deficiency factor can elevate a concern that the x-risk assessment is not as reliable as it should be since it is supported by a relatively new theory sitting at the centre of considerable debate and disagreement. This finding can be used to support a delay of some project, like a scientific endeavour. In this way, the first deficiency factor can function as a cause for a delay. However, it can also serve as a cure for the delay. If the scientific endeavour, following concerns, is delayed for 100 years, and if the (formerly) new theory becomes more entrenched and less disputed during the delay, that may boost the reliability of the x-risk assessment. In turn, this may suggest that the project can go ahead despite outstanding uncertainties.

As another example, the second deficiency factor can elevate a concern about the assessment's reliability since the assessment does not account for hidden flaws or because numerous flaws have already been uncovered since the work became available to open scrutiny. This finding can encourage a delay in the project—this is the second deficiency factor as a cause for a delay. Delaying a project for a century may, in turn, subject the risk assessment to a natural and orchestrated Ord-Hillerbrand-Sandberg-style analysis. If the x-risk assessment withstands 100 years of open security without the unearthing of significant flaws in its argument, then that is telling of the reliability of the assessment. In this way, the second deficiency factor can serve as a cure for the delay. Notably, the "…RHIC had been running for five years on

the strength of a flawed safety report, before Tegmark and Bostrom noticed and fixed this gap in the argument."[1266]

The idea of delaying a scientific endeavour for a century may be considered irrational and unscientific by some. Others may attack this line of thinking as promoting an undue obstacle to humanity's intellectual curiosity and scientific and technological progress. There is also the tentative reference to a century above: Why not one year, ten years, or 10,000 years? In other words, how much should the ASP affect the duration of any delay?

Whereas these concerns are understandable, it is also absurd to accept a condition where every scientific endeavour can proceed as of right, especially when it is linked to risk and the potential destruction of humanity's future. As written by Kent, it is necessary to "…accept that more stringent catastrophe risk criteria might indeed delay or preclude at least some interesting future experiments."[1267] He continues:

> One cannot defensibly adopt a mindset that requires that every interesting experiment must be carried out, and that sees every risk analysis as an exercise in justifying this foregone conclusion.[1268]

When safety concerns have been raised, the mere prospect of delaying some endeavours cannot be viewed as unreasonable in and by itself. Instead, the more pressing critique will expectedly revolve around the delay's *duration* (or its basis) and not the prospect of delaying some endeavours per se. Nevertheless, before condemning the duration, one should remember that *delaying* an endeavour in response to safety concerns is not the same as *banning* it. The former option is preferable to the latter. That being said, the delay must be more than symbolic. It should help address the underlying concerns in a manner responsive to the stakes and uncertainties associated with a debate such as the LHC Controversy. The delay's length may help with this prospect, especially if the justification for the postponement is grounded in the deficiency factors as outlined above.

[1266] Ord, Hillerbrand and Sandberg, 'Probing the Improbable' (n 1156) 200.
[1267] Kent, 'A Critical Look' (n 156) 166.
[1268] Ibid.

As previously suggested, the duration of the delay is meaningful when it comes to these deficiency factors because they hold a connection to the passage of time. A substantial delay means that the decision whether the endeavour can proceed can be taken at such a time when the science and scientific work underlying the exclusion of the risk has matured – i.e., its reliability is less uncertain – and stood the test of time when 'standing the test of time' is seen in the context of humanity's overall history owing to the ASP. In essence, it may be necessary to require an alleged x-risk originator to show that the relevant safety assurances are fortified to an extent, and over a period of time, that is worthy of the existential risk that the assessments seek to reject or quantify.

The prospect of delaying some endeavours for centuries has been raised before. Writing on *Practical Ethics* in 2008, Sandberg proposed that the benefits associated with the LHC might outweigh the extinction risks (presumed in his post to be one in a billion) associated with running the machine.[1269] Sandberg's post attracted many responses, including one by Ord targeting CERN's hastiness to push ahead with the collisions. His response reads:

> It is clear that the CERN team either haven't addressed many of the important issues, or haven't been transparent about it. Either way, the experiments should not presently go ahead. We can gain almost all the same benefits at reduced risks if we have a decade or five to seriously think about the issues, collect more data on the intersection of relativity and quantum mechanics, and put together a safety report worthy of the stakes. A serious reconsideration of the safety of the LHC, followed by the appropriate action is better than a blanket ban, or proceeding with our current level of ignorance.[1270]

Whereas Ord discussed delaying some endeavours – such as particle accelerator experiments – for up to 50 years, his idea lacks two essential elements. The first one is a principled justification for the delay in the first place. This justification can perhaps come from an external review analysing the deficiency factors so as to elevate a concern about the reliability of the x-risk assessment despite the review barriers

[1269] Anders Sandberg, 'Extinction Risks and Particle Physics: When Are They Worth it?' *(Practical Ethics*, 29 March 2008) <http://blog.practicalethics.ox.ac.uk/2008/03/extinction-risks-and-particle-physics-when-are-they-worth-it/> accessed 04 May 2023.

[1270] See Toby Ord's comment in the comments section to Sandberg, 'Extinction Risks and Particle Physics' (n 1269).

outlined in Chapter 3. The other missing element is an underlying basis for the duration of the delay. This is where the ASP comes into play. This concept can give credence to the length of the delay.

Let's come back to the notion of delaying the LHC for 50 years. This idea is not all that unreasonable when looking at, in particular, the history of the safety assurances vis-à-vis the RICH and LHC. As demonstrated before, a significant theme in the LHC Controversy was "…a repeating pattern of retreat and fortification in arguing the case for the safety of particle colliders. Arguments that are initially offered as unsusceptible to doubt [were] quietly abandoned when weaknesses [were] exposed. The new arguments are offered with the same sense of resolute conviction."[1271] Moreover, the benefits flowing from the LHC were not and are not imminently necessary to improve humanity's overall condition or for its self-preservation. The situation would have been different if the LHC experiments (knowingly) contributed to an existential trade-off.[1272]

Over a decade later, in his 2020 book, Ord reiterates the general theme seen in his above-mentioned response, albeit more broadly and beyond the context of high-energy physics experiments. He reasons that the growth of humanity's *power* (and technological progress) outstrips the growth of its *wisdom* and that it may be useful to slow down the former relative to the latter.[1273] His book reads:

> I think that a more patient and prudent humanity would indeed try to limit this divergence. Most importantly, it would try to increase its wisdom. But if there were limits to how quickly it could do so, it would also make sense to slow the rate of increase in its power—not necessarily putting its foot on the brake, but at least pressing more lightly on the accelerator.[1274]

Ord does not advocate for a ban or the abandonment of certain areas of technology or endeavours. Instead, his claim centres around accepting that humanity might not be ready for certain technologies or endeavours "…until meeting a given standard."[1275]

[1271] Johnson, 'The Black Hole Case' (n 68) 836.
[1272] The thesis uses the term 'existential trade-off' to describe a situation where the taking of x-risk (A) may help humanity mitigate x-risk (B) that is potentially more likely than risk (A).
[1273] Ord, *Precipice* (n 16) 206.
[1274] ibid.
[1275] ibid 207.

Such a standard can take many different shapes. Ord provides an example in his book: "[N]o nuclear technologies until we've had a hundred years without a major war."[1276] Another example of a standard might be to delay a particular experiment until, for instance, 100 years have passed without any grave upset in the scientific theory underlying the relevant x-risk assessment. Yet another standard might revolve around a holistic analysis of the deficiency factors. If it is considered that there is reasonable doubt as to the reliability of the x-risk assessment following an external investigation, the relevant standard is not met; the effect being that humanity is deemed not to be ready for whatever technology or endeavour the assessment concerns. Contrariwise, if there is no reasonable doubt following the investigation, the standard is met, and a delay is not justified.

One can also think of a requirement that revolves around not pushing ahead with certain experiments until it is possible to confirm, or observe, that these experiments are not "…going outside of the envelope of what we think is going on in the universe."[1277] Such a standard might consequentially be met if the scientists and their organisations can prove they are doing something which already "…happens in the universe normally and [that] there are no observed super-dangerous processes linked to it…"[1278] If such a standard cannot be met, it may be sensible to delay the experiment until observations of natural occurrences and analogous 'natural experiments' coupled with humanity's continued survival can contradict the x-risk associated with the endeavour. Yet, such a standard is not watertight because it fails to account for a significant observation selection effect. The next Sub-section (6.5.1) considers this aspect in relation to the RHIC and LHC. The upcoming Sub-section is also relevant in thinking about how a significant RHIC assessment contained a noticeable flaw.

6.5.1 Particle Accelerators and Observation Selection Effect Bias

CERN would have been in a relatively strong position to meet a standard built around proving that the LHC does not do anything that does not already occur elsewhere in the universe. As the Organisation writes: "Whatever the LHC will do, Nature has

[1276] ibid.
[1277] Anders Sandberg, 'Cool Risks Outside the Envelope of Nature' (n 1194).
[1278] ibid.

already done many times over during the lifetime of the Earth and other astronomical bodies."[1279] This observation, coupled with the fact that the Earth (and other celestial bodies) still exists, is used as an argument to exclude the possibility of any dangerous outcomes.[1280] As written regarding the formation of synthetic black holes, for example, "[t]he fact that the Earth and Sun are still here rules out the possibility that cosmic rays or the LHC could produce dangerous charged microscopic black holes."[1281] As already mentioned, this argument is not impermeable as it omits to account for observation selection effects. Max Tegmark (professor of physics at MIT) and Bostrom conveyed this message in discussing the risks associated with the RHIC.[1282]

A bit of context is helpful before highlighting their argument. Just like the LSAG report on the safety of the LHC, an earlier study regarding the disaster scenarios at the RHIC highlighted that natural experiments (i.e., collisions) have already occurred many times in the Universe.[1283] This observation was used as "empirical evidence" to exclude some of the dangers associated with the RHIC experiments. The study states:

> We know of two domains where empirical evidence tells us that cosmic ray collisions [occurring naturally in the Universe] have not produced strangelets with disasterous consequences: first, the surface of the Moon, which has been impacted by cosmic rays for billions of years, and second, interstellar space, where the products of cosmic ray collisions are swept up into the clouds from which new stars are formed. In each case the effects of a long-lived, dangerous strangelet would be obvious, so dangerous strangelet production can be bounded below some limit.[1284]

Nevertheless, Tegmark and Bostrom's paper highlights that it is not prudent to idly rely on Earth's continued existence and its history of having survived for billions of years to argue that the catastrophic event is impossible or to infer that it is exceptionally rare.[1285] This is because such an argument does not account for "…the observation selection effect…that precludes any observer from observing anything other than that

[1279] CERN, 'The Safety of the LHC' (n 127).
[1280] ibid.
[1281] ibid.
[1282] Tegmark and Bostrom, 'How Unlikely is a Doomsday Catastrophe?' (n 1221) 754.
[1283] Jaffe and others, 'Review of Speculative "Disaster Scenarios" at RHIC' (n 533).
[1284] ibid.
[1285] Tegmark and Bostrom, 'How Unlikely is a Doomsday Catastrophe?' (n 1221) 754.

their own species has survived up to the point where they make the observation."[1286] Our planet, and the observer, may simply have been exceptionally lucky up until the observation, even if the "…average cosmic neighbourhood is typically sterilized…" relatively frequently.[1287] In light of their argument, Tegmark and Bostrom considered the risk associated with accelerators in a manner that was not affected by the relevant selection-observation bias.[1288] In other words, they accounted for this bias in attempting to quantify the risk from particle accelerator experiments. Whereas they conclude that the risk remains "reassuringly small",[1289] their paper demonstrates why pointing to analogous natural experiments coupled with humanity's continued survival is not a watertight safety argument.

In addition, thinking back to the Ord-Hillerbrand-Sandberg-style analysis, Tegmark and Bostrom's paper demonstrate that the key study on RHIC by Jaffe and others contained a noticeable flaw. This fact did not elude Ord, Hillerbrand and Sandberg writing that "[t]his is an example of a demonstrated flaw in an important physics risk argument (one that was pivotal in the safety assessment of the RHIC)."[1290] While the flaw was ultimately discovered, the RHIC had nonetheless been up and running for five years based on a flawed safety report.[1291]

6.6 SUMMARISING THE BROAD LEGAL INTERVENTION

This Part will summarise how the various observations made until this point in the Chapter can be used as the foundation for a broad legal intervention for x-risk mitigation. Part 6.6 is divided into two main sections: Section 6.6.1 on the theoretical substance of the intervention and Section 6.6.2 on its practical application.

The focus is on the theoretical substance because, as mentioned in Chapter 1, "…what matters at this stage is the theoretical building blocks and the function and

[1286] ibid.
[1287] ibid.
[1288] ibid.
[1289] ibid.
[1290] Ord, Hillerbrand and Sandberg, 'Probing the Improbable' (n 1156) 200.
[1291] ibid.

logic behind the legal intervention as inspired by the LHC Controversy." Before proceeding, it is useful to summarise the essence of the LHC-inspired mechanism:

> The broad legal intervention in response to the LHC Controversy centres around giving an external reviewer a principled basis upon which to base a finding that some endeavours should not *presently* be allowed to go ahead since, and notwithstanding the review barriers, there are reasonable doubts as to the reliability of the work assessing an existential risk linked to these endeavours.

6.6.1 Theoretical Substance

This Chapter has argued that an external reviewer can analyse a set of interconnected deficiency factors in evaluating the reliability of scientific work assessing an x-risk associated with an endeavour such as experimental particle physics. The deficiency factors require analysis of (i) the humans and organisations who considered the postulated x-risk and (ii) the potential that they have based their conclusion on flawed information or an incomplete state of knowledge. Analysis of the deficiency factors coupled with the ASP can be used as a theoretical basis for a broad legal intervention seeking to reduce the likelihood of an existential risk. It is helpful to summarise some key observations made above before introducing some further directions for the LHC-inspired intervention.

Firstly, the deficiency factors correlate with observations circulating within the field of x-risk research. Secondly, it was argued that the material value of the synergy between these factors is to elevate a potential concern that the x-risk assessment is not, at present, as reliable as it should be. The thesis divided the examination of this two-prong statement between 'reliability' and 'present'. Thirdly, and regarding the first prong of the statement above, it was argued that the deficiency factors can be analysed in tandem to target the reliability of the assessment. An elevated concern about reliability can justify a delay, even if the review barriers outlined in Chapter 3 render other external review methods, such as cost-benefit analysis and validation of competing scientific claims, problematic. Fourthly, and regarding the second prong of the statement, it was argued that the ASP should affect the interpretation of at present. In turn, this can set the standard for the duration of the delay. Fifthly, it was mentioned

that the deficiency factors hold a connection to the passage of time, which means that they combine well with the ASP and vice versa.

The LHC-inspired intervention can be imbued with further substance by adding certain directions, limitations, and general parameters. Cross-reference will also be made to the x-risk landscape because the sub-sections below attempt to track where, within this landscape, the legal mechanism may and may not find application.

(i) Beyond the Context of Preliminary Injunction Requests

The thesis has utilised Johnson's meta-analysis to compartmentalise the deficiency factors that an external reviewer can analyse to consider the present reliability of an x-risk assessment. Notably, Johnson places his meta-analysis within the preliminary injunction context. Chapter 4 explored that some individuals have reasonably questioned this court-centric approach. In any event, the meta-analysis can, according to Johnson, permit a generalist judge to assess whether there are "serious questions" to merit such interim remedy.

The language of "serious questions" within the preliminary injunction context (as far as the US and UK legal systems are concerned) should be discarded for the purpose of the LHC-inspired intervention. This is because the LHC-inspired intervention should operate beyond the context of preliminary injunction requests. Therefore, rather than a generalist judge considering whether there are "serious questions" to justify an interim injunction, the appropriate external reviewer should, analyse the deficiency factors to consider if the x-risk assessment is, at present, sufficiently reliable. Alternatively, taking inspiration from the language employed by Wilson, one may suggest that an external reviewer should consider whether there are "reasonable grounds for concern"[1292] that the assessment is not, at present, sufficiently reliable.

[1292] Wilson, 'Minimizing' (n 151) 355.

(ii) Not Limited to the Judiciary

In line with the Sub-section above, the LHC-inspired intervention should not necessarily be limited to the judiciary. Section 6.6.2 below will deal with this matter in more detail. For now, it suffices to say that the legal mechanism can, and should, be seen as disconnected from litigation. That is not to say that the judiciary should play no role in relation to the LHC-inspired intervention. However, it is proposed that the issue as to whether an x-risk assessment is sufficiently reliable should be addressed long before an endeavour is close to completion, launch or real-work application, avoiding a situation where private individuals are driven to challenge the alleged x-risk originator in court.

(iii) A Component Within an International Mitigation Scheme

The LHC-inspired intervention may itself form part of a comprehensive (and preferably international) mitigation scheme. For instance, it may feature in a novel international treaty designed to mitigate x-risk from scientific endeavours. Alternatively, given the abundance of international treaties, there may be scope to realise the intervention by amending an existing treaty.

The key point is that the LHC-inspired intervention may itself form part of an even broader mitigation scheme. For example, a novel treaty designed to mitigate existential may consist of various interrelated regulatory mechanisms,[1293] with the LHC-inspired intervention working together with other regulatory mechanisms. An example of this can be provided by revisiting Wilson's treaty on GCRs and x-risks. In considering the implementation and use of the precautionary principle, Wilson writes that a body of experts may, in some instances, decide whether the available information displays reasonable grounds for concern "…that a certain risk exceeds whatever level is deemed acceptable."[1294] Ultimately, such a finding could "…trigger a requirement that states impose certain measures to regulate this technology or even prohibit it until there is further research of the risks."[1295] He also suggests that the

[1293] ibid; Kohler, 'Research and Technology Risks: Part IV' (n 1094).
[1294] Wilson, 'Minimizing' (n 151)354-355.
[1295] ibid 355.

treaty could require the advocates to refute the reasonable grounds for concern before being allowed to continue development or applying the relevant technology.[1296]

Wilson's proposal can be linked to the LHC-inspired intervention. To begin with, his suggestion outlined immediately above works alongside other mechanisms within the framework treaty. Similarly, as mentioned, the LHC-inspired intervention can sit within a wider mitigation scheme, such as a treaty. Another noteworthy aspect is that the LHC-inspired intervention may offer a more comprehensive and principled basis for Wilson's proposal above. This is because the 'body of experts' that he refers to may use the LHC-inspired intervention in deciding whether there is a reasonable ground for concern. If there is, this may justify the temporary prohibition that Wilson highlights. One can also see that Wilson's treaty mechanism allows the endeavour's advocates to refute the reasonable ground for concern. The LHC-inspired intervention may play a part here too. As noted above in this Chapter, analysis of the deficiency factors can function as the cause and cure for a delay.

Since the LHC-inspired intervention can form part of a more comprehensive mitigation scheme, it can operate holistically and alongside other regulatory components. For example, this thesis has previously discussed the concept of an existential trade-off, and it might be that the LHC-inspired intervention can work alongside an existential trade-off clause. Chapter 4 highlighted Kohler's Prototype Regulation. This Prototype stipulates, among other things, that research or technology undertakings "…bearing an existential risk are only acceptable when they remedy another existential risk with higher likelihood."[1297] The thesis has previously questioned this proviso because it hinges on the quantification of the probability of two separate x-risks when doing the same in respect of merely one such risk can be highly challenging. This point aside, what is essential for present purposes is that the LHC-inspired intervention may operate alongside other regulatory components, such as some rendition of an existential trade-off clause.

[1296] ibid.
[1297] Article 3(11); Kohler, 'Research and Technology Risks: Part IV' (n 1094).

As another example of the LHC-inspired intervention working alongside other legal mechanisms, Kohler's Prototype enlists an administration ("the Authority") to apply the Regulation.[1298] The Prototype Regulation imbues the Authority with various functions,[1299] empowerments[1300] and obligations.[1301] Aligning with the LHC-inspired intervention, Article 8(1)(c) of the Prototype empowers the Authority "to temporarily stop an undertaking or subject it to conditions in view of further investigating the related risks…"[1302] The LHC-inspired intervention may even offer the principled basis for the Authority's empowerment to issue long-term bans in cases of significant uncertainty because of review barriers similar to those examined in Chapter 3.

Finally, the broad legal intervention may form part of a mitigation scheme subjecting some risk originators to an existential risk-assessment obligation. However, the LHC-inspired intervention is not, per se, intended to function as the trigger for compelling x-risk originators to undertake such an assessment. Instead, the intervention deals with the reliability of the assessment itself.

In sum, the LHC-inspired intervention can form part of an overarching mitigation scheme dealing with both substantive and procedural aspects of x-risk mitigation.

(iv) Universal < Broad Legal intervention

While the LHC-inspired intervention is classified as broad in nature, it is not intended to serve as an all-purpose intervention that can mitigate x-risk from all hazards within the risk landscape. For example, there may be little value in applying the LHC-inspired intervention in relation to the mitigation of x-risks associated with nuclear warfare, a

[1298] ibid.
[1299] E.g., the Authority may authorise certain undertakings before they can proceed (Article 6); the Authority shall create a registry for research and technology undertakings (Article 14); it shall create an alert portal allowing people to anonymously submit information about undertakings (Article 15). ibid.
[1300] Article 8. For instance, the Authority is empowered to request information, temporarily or permanently stop certain undertakings and subject those undertakings to conditions. ibid.
[1301] Article 9. For instance, the Authority shall investigate undertakings that may pose a risk and ensure that all staff members are independent and have no personal stake in the undertaking they investigate. ibid.
[1302] Article 8(1)(c). ibid.

synergy of multiple man-made catastrophes or 'natural' x-risks associated with, for instance, asteroid impacts, the death of the Sun or Universe.

Naturally, the intervention can apply within the context of the LHC Controversy, which is the archetypical example of a situation where the review barriers made various modes of risk analysis difficult. However, it is necessary to ask whether these review barriers can, in some shape or form, manifest in the future and beyond experimental particle physics so that the LHC-inspired intervention can find application beyond the concerns associated with particle accelerators like the LHC or RHIC. If the answer is no, the intervention is not a broad one. Instead, it would be a *targeted* intervention.

There are reasons for believing that the LHC-inspired intervention can find application beyond the concerns associated with particle accelerators like the LHC and so correctly classify as broad. It is possible to support this from different vantage points.

(i) The LHC-inspired legal intervention may apply to situations involving future unknown unknowns. These unknown unknowns may (or may not) manifest as future scientific endeavours not involving particle accelerators. Admittedly, this argument may be seen as a weak attempt to justify an intervention by pointing to things that may or may not transpire in the future. Nevertheless, unknown unknowns constitute a credible hazard within the x-risk landscape.[1303] An intervention that might apply to certain unknown unknowns is valuable, especially when appreciating that most current x-risks were unknown unknowns a few centuries ago.

(ii) Examples from the recent past may also suggest that the LHC-inspired intervention can find application beyond the concerns associated with machines like the LHC and RHIC. The first example concerned the Manhattan Project. As noted in earlier Chapters, some physicists involved in developing the bomb began worrying the Trinity Test could potentially ignite the atmosphere and thereby cause an x-catastrophe.[1304]

[1303] See Chapter 2.

[1304] Hans Bethe, 'Ultimate Catastrophe?' [1976] Bulletin of the Atomic Scientists 36, 36; Yuen Yiu, 'The Fear of Setting the Planet on Fire with a Nuclear Weapon' *Inside Science* (15 July 2020) <https://insidescience.org/manhattan-project-legacy/atmosphere-on-fire> accessed 04 May 2023; Leiss, *The Doom Loop in the Financial Sector, and Other Black Holes of Risk* (n 16) 25-28; Häggström,

A secret scientific report (LA-602) which is now declassified was produced on the possibility of this happening.[1305] It might be recalled that some have observed that LA-602 might be the first-ever quantitative risk assessment of human extinction.[1306] Whereas the authors of LA-602 viewed the ignition-catastrophe as very unlikely, they concluded the report by saying that "…the complexity of the argument and the absence of satisfactory experimental foundations makes further work on the subject highly desirable."[1307] It is possible to uncover some traits within this risk scenario that support the claim that the LHC-inspired intervention could have seen some application to the Manhattan Project.

Firstly, even though many experts worried right up until the test, the physicists' x-risk assessments did not benefit from external review.[1308] Secondly, the atomic bomb was only made possible owing to the (then) recent discovery of nuclear fission. Thirdly, and as seen above, LA-602 recognises the complexity of its own argument and further work on the matter as being desirable. Fourthly, the Trinity Test and ignition possibility (x-risk one) was not needed to offset another x-risk. Whereas one may view Nazi Germany's victory as a potential non-extinction x-risk,[1309] Germany had surrendered before the Trinity Test, and Japan was near defeat.[1310] With this in mind, and using the LHC-inspired intervention, an external reviewer could theoretically have decided that the Trinity Test was not (at that time) to go ahead since there were reasonable doubts about the *reliability* of the underlying x-risk assessment.

As written by Ord: "Given the weak conclusions of the [LA-602] report, the inability to get external review, and the continuing concerns of eminent scientists, there was a strong case for simply delaying, or abandoning, the test."[1311] Had the Trinity Test been delayed, a greater and more entrenched theoretical understanding would have strengthened the reliability of the x-risk assessment. As would the availability of

Here Be Dragons (n 22) 198-199; Baum, 'The Great Downside Dilemma' (n 22); Miller and Felton, 'The Fermi Paradox; Bayes' (n 22) 50.

[1305] Konopinski, Marvin, and Teller, 'Ignition of the Atmosphere with Nuclear Bombs' (n 23).

[1306] Bostrom, 'Existential Risk Prevention' (n 15) 29; Yudkowsky, 'LA-602 vs. RHIC Review' (n 24).

[1307] Konopinski, Marvin, and Teller, 'Ignition of the Atmosphere with Nuclear Bombs' (n 23).

[1308] Ord, *Precipice* (n 16).

[1309] As noted in Chapter 1, Bostrom's definition of an x-risk covers non-extinction scenarios.

[1310] Baum, 'The Great Downside Dilemma' (n 22).

[1311] Ord, *Precipice* (n 16) 93.

computers that could have aided the physicists' calculations,[1312] the longevity of the underlying argument and absence of uncovered flaws, and more opportunities for external review to reduce the effect of various biases. Of course, the above is theoretical given, among other things, the secrecy that surrounded the Manhattan Project and the lack of any established external reviewer.

The CUORE experiment is another example from the recent past suggesting that the LHC-inspired legal intervention may find some application beyond the concerns associated with the LHC and RHIC machines. The CUORE experiment involved cooling a copper vessel with a volume of one cubic meter and weight of ~400 kg to -273.144 degrees Celsius in 2014.[1313] This experiment was first to "…cool a mass and a volume of this size to a temperature this close to absolute zero (0 Kelvin). The cooled copper…was the coldest cubic meter in the universe for over 15 days."[1314] In contrast to the LHC Controversy, there was (and has been) far less discussion about the potential x-risk associated with the CUORE experiment. Nevertheless, Sandberg[1315] and Miller and Felton[1316] have considered this experiment within the context of x-risk. The main concern, if any, is linked to the fact that copper does not get this cold in nature.[1317] Thus, and setting it aside from the LHC collisions, the COURE experiment goes outside "the envelope of nature".[1318]

As with the Manhattan Project above, it is necessary to consider if there are any factors that can support the claim that the LHC-inspired intervention could have seen application in relation to the CUORE experiments. Unlike with the ignition of the atmosphere and the synthetic black hole catastrophes, there is seemingly no risk assessment dealing with the potential x-risks associated with the CUORE experiment. As such, and as a matter of practicality, the LHC-inspired intervention could not have been utilised to assess whether the underlying x-risk assessment was sufficiently reliable for present purposes. The lack of such an assessment may convey an

[1312] ibid 92.
[1313] Interactions, 'CUORE' (n 257).
[1314] Ibid.
[1315] Anders Sandberg, 'Cool Risks Outside the Envelope of Nature' (n 1194).
[1316] Miller and Felton, 'The Fermi Paradox; Bayes' (n 22) 51.
[1317] ibid.
[1318] Anders Sandberg, 'Cool Risks Outside the Envelope of Nature' (n 1194).

important lesson from the CUORE collaboration: Namely, a risk assessment might be warranted whenever scientists plan to do things or create conditions that do not happen or exist naturally in the universe.

Admittedly, implementing this lesson too strictly might be impractical because it is likely possible for you, as the reader, at this very moment, to do something that does not normally happen in the universe. It is possible to give an example of this by drawing inspiration from Miller and Felton's article.[1319] The example features in the following footnote to avoid straying too far away from the essential point here.[1320]

Getting back on track, and for the sake of argument, suppose that the scientists involved in the CUORE collaboration had assessed if cooling the cooper vessel so close to absolute zero – thereby transcending the envelope of nature – could have caused an x-catastrophe. That is, presume that they had produced an x-risk assessment.

It is impossible to tell how such an alternative reality would have played out. However, it seems plausible to assume that the assessors would have concluded that the CUORE experiment posed no x-risk in light of current theoretical understanding and the present state of humanity's knowledge. In turn, the reliability of their assessment could have been weakened or reinforced by subjecting it to the LHC-inspired legal intervention. For example, the assessment would likely have mentioned the lack of theories suggesting that cooling the cooper vessel according to plan posed any danger. This specific argument is appropriately evaluated by analysing the themes represented by the first deficiency factor. This involves an external reviewer giving weight to the possibility that a theory linking the CUORE experiment to a dangerous

[1319] Miller and Felton, 'The Fermi Paradox; Bayes' (n 22).

[1320] The following example is an adaptation of Miller and Felton's thought experiment. Suppose you produce a random character string on your keyboard with ~ 50 letters, symbols and numbers. An attempt might look something like this: "/.v.;,efpojgertt-32113-=3ergjcreogvnroin2g58h." One can assume this is the first time this character string has appeared within Earth's neighbourhood, which Miller and Felton define as "…everything within a hundred million light-years of our solar system." In effect, one should ask whether a risk assessment is warranted since this character string would, but for your act, not be produced elsewhere within Earth's neighbourhood and thereby take us outside the envelope of nature? One will remember that the LHC-inspired legal intervention is not designed to trigger an existential risk-assessment obligation. As such, it is beyond the scope of this thesis to deal with this issue. ibid 51.

process might come to light as humanity's knowledge of the laws of physics continues to mature. However, as argued above, the analysis of the first deficiency factor must be conducted in a principled manner. In other words, the reviewer should not simply give credence to an argument that there might be future discoveries linking the experiment to x-risk. A more principled analysis of the first deficiency factor may, as previously suggested, come from assessing, inter alia, the newness of the underlying scientific theory, the lack of historical robustness of relevant safety arguments, the gravity of past upsets within the field, the pace at which new ideas are introduced, and the frequency by which old ones are challenged.

How does a principled application work when there are no theories linking the endeavour to dangerous processes? The longstanding lack of such theories – and the continued lack after an open and external review – would arguably strengthen the reliability of the x-risk assessment. In this way, and regarding the first deficiency factor, the LHC-inspired intervention would arguably have supported the reliability of the x-risk assessment associated with the CUORE collaboration. The same is not true had the assessors justified their safety conclusion based on a new theory that displaced past theories linking the endeavour to x-risk. Of course, it is useful to recall that the first deficiency factor should be examined in tandem with the other factors. For example, drawing on the second deficiency factor, an *open* and *external* review of the risk assessment for a reasonable period would – save for discovering flaws during that period – have strengthened the assessment's reliability. Contrariwise, drawing on the third and fourth deficiency factors, a lack of independence and questionable organisational culture fuelling biases that may impact the assessment's reliability would have weakened the conclusion, notwithstanding the lack of past theories linking existential risk to the cooling experiment.

(iii) The ability to deconstruct the LHC-inspired legal intervention might also suggest that it can find application beyond concerns linked to the LHC and RHIC. This point relates to the synergy between the deficiency factors. In essence, if it is impractical to analyse a given deficiency factor, the others might still be assessed in considering the reliability of the assessment. In fact, analysing the themes represented by a single deficiency factor can seemingly be valuable. The third deficiency factor provides a clear example. Uncovering the existence of a conflict of interest amongst the risk

assessors can, by and of itself, be reason enough to question the reliability of an x-risk assessment. Unconscious biases (the fourth factor) can also elevate a concern about the assessment's reliability (irrespectively of the existence of a conflict of interest). Thus, both the third and fourth tools may function in isolation across most risk scenarios where someone assesses whether an activity, project, or scientific endeavour, be it in relation to AGI, nanotechnology or geoengineering, poses an x-risk. In essence, deconstructing the intervention may give it further reach.

This Sub-section has extended three overarching points that may support the claim that the LHC-inspired legal intervention can, with some degree of plausibility, find application and utility beyond the context of the LHC Controversy and so classify as a broad intervention. The upcoming sub-sections will continue to add a few suggested directions, limitations and parameters to add substance to the LHC-inspired intervention.

(v) No Automatic Ban

The LHC-inspired intervention should not be viewed as a precursor to an automatic delay. On the contrary, the intervention should enable an external reviewer to conclude that the endeavour can go ahead if there are no reasonable concerns about the x-risk assessment's reliability. As seen above, an x-risk assessment connected to the CUORE experiment could potentially have satisfied the 'reliability' standard imposed by the intervention despite the experiment transcending the envelope of nature.

(vi) Anthropogenic Risks & Terror v Error

It is to be appreciated that the LHC-inspired legal intervention applies to anthropogenic rather than natural x-risks. Nevertheless, natural x-risks might indirectly relate to the intervention advanced in this thesis. For example, an endeavour might be exempted by offsetting a natural x-risk even though, following the use of the LHC-inspired intervention, there are reasonable doubts about the reliability of the x-risk assessment associated with this endeavour. Furthermore, within the anthropogenic scene, the LHC-inspired mechanism is intended to cover error-type rather than terror-type risks.

(vii) Theoretical Building Blocks

It is worth recalling that this thesis will not formalise the LHC-inspired intervention by proposing specific draft clauses or sections, the design of an international treaty, or the like. Instead, what matters is the theoretical building blocks and the function and logic behind the intervention. It is early days for broad legal interventions for x-risk mitigation, and the LHC-inspired intervention might not prove fruitful.

The better option is to make the case that there is scope to build upon the theoretical foundations examined in this thesis. For example, it might be possible to consider how the LHC-inspired intervention can be taken further and how it can be expressed in the form of draft clauses. Alternatively, it is possible to explore the synergy between the legal intervention and other regulatory mechanisms. Further research might also examine whether the LHC-inspired intervention already fits within some existing treaty or regulatory framework.

Section 6.6.1 above has attempted to imbue the LHC-inspired intervention with further substance by highlighting directions, limitations and parameters. Section 6.6.2 will move on to consider the intervention's practical application. However, it will not deal with this aspect in detail. Instead, the discussion on the practical application is confined to a concise examination of the institutional framework within which the LHC-inspired intervention might operate. The decision not to focus on this aspect in detail stems from the fact that the essential contribution of this thesis concerns the theoretical substance underpinning the LHC-inspired intervention.

It is to be recalled that Chapter 4 critically examined Johnson's court-centric approach. It also compared judicial review versus interagency regulation in the context of risk originating from agencies.[1321] That discussion is relevant in the present context of thinking about the most suitable institution within which the LHC-inspired intervention might be levied. Significantly, the discussion from Chapter 4 may support to the suggestion that the intervention might work better if utilised by a regulatory agency rather than the judiciary.

[1321] See Section 4.5.1.

6.6.2 Practical Application: Institutional Framework

The LHC-inspired legal intervention should preferably be applied at an international governance level. Some form of international governance is the obvious choice, given the nature of x-risks. Despite the challenges in agreeing to, realising, and implementing the LHC-inspired intervention at an international governance level, this Section focuses on who the external reviewer might be.

It is the external reviewer that may utilise the legal mechanism to evaluate whether, notwithstanding the review barriers, there is reasonable doubt as to the reliability of the x-risk assessment. Ultimately, the endeavour or project which is assessed can justifiably be delayed if the review uncovers that the assessment is not presently as reliable as it should be.

The current discussion can be fleshed out by taking inspiration from scholars who have, to some extent, dealt with the institutional framework for x-risk mitigation. To this effect, the sub-sections below discuss the practical application under the following headings: An international court, a novel international agency, and an existing international organisation.

(i) Courts

One tentative option is for the judiciary to assume the role of the external reviewer. Some scholars have proposed the possibility of domestic or international courts dealing with the mitigation of x-risks. For example, Wilson writes that an international treaty could either:

> (1) require states to establish domestic 'science courts' that are equipped to consider alleged GCRs/ERs arising from emerging technologies…or (2) create an international court that enables citizens to submit disputes regarding GCRs/ERs from emerging technologies, much like the right of European citizens to submit disputes to the European Court of Human Rights.[1322]

[1322] Wilson, 'Minimizing' (n 151) 358.

Wilson further suggests that judges in scenarios (1) and (2) should be scientifically literate lawyers with an ability to understand the science.[1323] Similarly, Peterson proposes a new science court in light of the LHC Controversy.[1324] Ord also highlights the option of resorting to a scientifically literate judge who, following arguments for and against a given endeavour, can decide whether the experiment should "…proceed at this time, or whether it should be delayed until a better case can be made for it."[1325] However, albeit without explanation, Ord supposes that "very few" experiments would be denied following this process.[1326] Finally, as mentioned a few times above, Johnson also gears his discussion towards the judiciary. Nevertheless, in contrast to scientifically literate judges, he places generalist judges in the spotlight.[1327]

In sum, the judiciary might assume the role of the external reviewer and employ the LHC-inspired intervention to gauge whether the x-risk assessment is as reliable as it should be. Given the nature of the legal intervention, the judge need not be scientifically literate, even though it might help. More important is an insight and appreciation of the skills that the field of existential risk studies has accrued, borrowed, and formalised.

(ii) A Novel International Agency

Another option entails establishing a new international agency that assumes the role of the external reviewer. Posner has written about such an option in the context of the institutional framework for dealing with catastrophic risks. For example, he considers the possibility of creating an International Environmental Protection Agency[1328] or International Bioweaponry Agency.[1329] Another option that Posner highlights is to ensure that technologically advanced nations enact a law requiring certain scientific

[1323] ibid.
[1324] Peterson, 'The "Sancho" Effect' (n 468).
[1325] Ord, *Precipice* (n 16) 378.
[1326] ibid.
[1327] Johnson, 'The Black Hole Case' (n 68); Johnson, 'Agencies and Science-Experiment Risk' (n 68); Johnson, 'Uncertain Risk, Science Experiments, and the Courts' (n 68); Johnson, 'Judicial Review' (n 68).
[1328] Posner, *Catastrophe* (n 90) 216-218.
[1329] ibid 218-221.

projects to be reviewed by a Catastrophic-Risk Assessment Board.[1330] These examples may be of little direct relevance to this thesis. However, they underscore the theoretical possibility of creating an international agency that may use the LHC-inspired intervention to mitigate x- risk in some situations.

Furthermore, Wilson notes the prospect of concluding a treaty that grants a body of experts the authority to manage existential risks.[1331] According to Wilson, the Environmental Protection Agency of the United States can potentially function as a model for such a body.[1332] Furthermore, and in the context of x-risk from AI, Nindler outlines the potential establishment of an international enforcement agency for safe AI R&D. Such an enforcement agency, he writes, can "…be an integral and extremely important part of the implementation and enforcement regime of a potential future treaty on the issue."[1333]

In sum, the LHC-inspired intervention might be utilised by some novel international agency made up of scientists, lawyers, x-risk scholars, and national government representatives. The agency may, hence, act as an external reviewer and evaluate whether the x-risk assessment is presently as reliable as it should be. Whereas this thesis will not explore the configuration of such a hypothetical agency, one may wonder how the agency might become involved in assessing the reliability of the assessment in the first place.[1334] Kohler's Prototype Regulation can perhaps shine some light on this question.[1335] For instance, the establishing treaty may impose obligations on certain companies, research institutions and individuals (i.e., x-risk originators) to assess risks before proceeding with proposed undertakings. The relevant risk originator may be required to inform the agency, seek relevant authorisation, and register their project in the agency's database.[1336] The agency might then use the LHC-inspired intervention as part of the authorisation process.

[1330] ibid 221.
[1331] Wilson, 'Minimizing' (n 151) 355-357.
[1332] ibid 355.
[1333] Nindler, 'The United Nation's Capability' (n 54) 31.
[1334] The judiciary will become involved when cases are brought to court.
[1335] Kohler, 'Research and Technology Risks: Part IV' (n 1094).
[1336] Manfred Kohler, 'Regulating Research and Technology Risks: Part I – Research Risks' *The Regulatory Institute* (10 September 2017) <https://www.howtoregulate.org/regulating-research-technology-risks-part-i-research-risks/#more-248> accessed 04 May 2023.

(iii) An Existing International Organisation

The third option entails using the expertise and framework of an existing international organisation. For instance, the UN has been referred to by scholars who have considered the institutional framework for x-risk mitigation.[1337] Boyd and Wilson write that member nations can ensure that there is a body at the UN that directly addresses specific x-risks such as those from, among other things, AI and biotechnology.[1338] An example of this approach may involve the WHO monitoring biotechnology-related x-risks.[1339] Beyond risk-specific bodies, Boyd and Wilson suggest that there can be an overarching body at the UN dealing with x-risk as a unique risk category. This overarching body might assess and "…recommend which risks justify greater or lesser immediate resources to assist analysis or mitigation measures."[1340] Finally, they note the UN might be involved with mitigation by protecting future generations.[1341]

The LHC-inspired intervention will not necessarily require an entirely new agency or organisation. Instead, it might find some application under, for example, the auspices of the UN. Noteworthy, the UN has expressly discussed the notion of x-risk in some settings. For one, in 2021, the UN published Our Common Agenda.[1342] This report "…presents the Secretary-General's vision on the future of global cooperation through an inclusive, networked, and effective multilateralism."[1343] It contains different proposals directly applicable to x-risk mitigation. It notes the presence of x-risks, the warranted effort to identify such risks,[1344] and the need to ensure long-term thinking through a Futures Laboratory which, among other things, facilitates the reporting on catastrophic risks and strengthens preparedness for such risks [1345] In short, mitigating existential risk under the auspices of the United Nations may not be too far-fetched.

[1337] Nindler, 'The United Nation's Capability' (n 54); Wilson, 'Minimizing' (n 151) 350; Posner, *Catastrophe* (n 90) 221; Boyd and Wilson, 'Existential Risks to Humanity' (n 154) 2308-2310.
[1338] ibid 2308-2309.
[1339] Wilson, 'Minimizing' (n 151) 350.
[1340] Boyd and Wilson, 'Existential Risks to Humanity' (n 154) 2309.
[1341] ibid.
[1342] UN, 'Our Common Agenda' (n 34).
[1343] ibid.
[1344] ibid 65.
[1345] ibid 7 and 45.

(iv) Summary

This Section has considered the institutional framework within which the LHC-inspired intervention might be levied. To this effect, the thesis considered different options for an external reviewer, including an international agency tasked with x-risk mitigation and protecting humanity's future, as part of which the LHC-inspired intervention might assume partial relevance. Nonetheless, making this happen in practice is easier said than done. There are many obstacles to the fruition of such an agency, including political willingness, a reluctance to surrender sovereignty, financial constraints, enforcement and monitoring issues, uncertainties involving the institution's more precise configuration, jurisdiction, powers and responsibilities. As observed by Ord:

> [I]t is very unclear at this stage what forms [international institutions focusing on x-risk mitigation] should take. This includes questions of whether the change should be incremental or radical, whether institutions should be advisory or regulatory, and whether they should have a narrow or broad set of responsibilities. Our options range from incremental improvements to minor agencies, through to major changes to key bodies such as the UN Security Council, all the way up to entirely new institutions for governing the most important world affairs.[1346]

Despite practical constraints, establishing an institutional framework for x-risk mitigation does not seem entirely theoretical. For example, there is a growing academic community focusing on x-risks and humanity's long-term future;[1347] there has been express recognition of this risk category and the significance of the future amongst policymakers at the national[1348] and international governance level;[1349] and it is possible to observe the establishment of committees for the future in various countries.[1350] As written in Our Common Agenda:

> At the national level, some countries have established committees for the future or future generations commissioners who advise governments and public bodies on the effects of present decisions on people in the future. Other States could establish similar mechanisms, building on these good practices. At the multilateral level, a growing number of Member States

[1346] Ord, '*Precipice* (n 16) 200.
[1347] Section 1.1.2.
[1348] The Government Office for Science, 'Innovation: Managing Risk, Not Avoiding It' (n 33) chp 10.
[1349] UN, 'Our Common Agenda' (n 34).
[1350] Jones, O'Brien, Ryan, 'Representation of Future Generations' (n 390).

> and advocates have proposed options to represent succeeding generations in the United Nations system, including through a Commissioner or Ombudsperson for Future Generations, a Commission of Global Guardians for the Future, or a repurposed Trusteeship Council.[1351]

The baseline view adopted in this thesis is that an international agency seems to be well-suited to apply the LHC-inspired intervention. This preliminary view is based on the nature of the LHC-inspired intervention and the institutional flexibility that may be realised through establishing an agency. Ultimately, it seems advantageous that the external reviewer can draw upon "…an organized body of thinking about…existential risks…"[1352] This is because, among other things, the deficiency factors that should be analysed as part of applying the LHC-inspired mechanism align with x-risk research.

6.7 THE LEGAL INTERVENTION AND THE THREE SELF-IMPOSED CONDITIONS

This thesis applies three self-imposed conditions that the LHC-inspired intervention should satisfy. Part 6.7 summarises how the broad legal intervention for x-risk mitigation, as examined in this Chapter, satisfies these conditions.

Firstly, the intervention should respond to the review barriers that may complicate different modes of external review in the LHC Controversy context. These obstacles were extrapolated in Chapter 3 by examining the *Sancho case*, which came about from an injunction request. This first condition is central to the LHC-inspired intervention because the review barriers constitute a crucial element of what can be learnt from the Controversy. The LHC-inspired intervention satisfies the first condition because the intervention's utility is not contingent upon the scientific merits of the claims, quantifiable probabilities of x-catastrophe or safety reassurances advanced by (potentially) self-interested and biased experts. This approach sets the LHC-inspired intervention apart from proposals where a court or reviewer is invited to focus on the merit of the underlying scientific arguments, undertake cost-benefit analysis or

[1351] UN, 'Our Common Agenda' (n 34) 45.
[1352] Yudkowsky, 'Cognitive Biases' (n 372) 112.

determine the acceptability of risk based on proposed subjective probabilities of x-catastrophe.

Secondly, the LHC-inspired intervention should use the ASP because it responds to risks that threaten to destroy humanity's future. In particular, the ASP may justify more onerous obligations being imposed on x-risk originators. This argument flows from the premise that the nature of the risk (including its severity) should influence the precautions demanded of a risk originator. However, this is not the primary way in which the LHC-inspired intervention utilises the ASP. Instead, for the purpose of the LHC-inspired legal mechanism, the ASP is connected to the very essence of the deficiency factors. As already mentioned, the deficiency factors hold a connection to the passage of time.

Thirdly, the LHC-inspired intervention should—with some degree of plausibility—be able to find utility beyond the concerns associated with machines including the LHC and RHIC. As examined above in relation to the Manhattan Project and CUORE experiment, there is some scope for the intervention to apply beyond the context of the LHC Controversy. However, it is recognised that the intervention proposed in this thesis is narrower than some other broad legal interventions that could operate within the x-risk landscape. However, a future research question may involve considering how the LHC-inspired intervention might be adapted to expand its remits and scope.

6.8 CONCLUSION

Chapter 6 has built upon observations throughout this thesis to formalise the theoretical substance for a broad legal intervention for x-risk mitigation. The intervention derives from the LHC Controversy. It centres around giving an external reviewer a principled basis to conclude that some endeavours should not presently go ahead since, notwithstanding the review barriers or some credible combination thereof, there are reasonable doubts about the reliability of an x-risk assessment.

The conclusion as to the x-risk assessment's reliability can come from an analysis of interrelated deficiency factors. This Chapter has used Johnson's meta-analysis to

compartmentalise the deficiency factors. These factors do not involve consideration of the scientific merits or the probability of existential catastrophe. Instead, they require analysis of (i) the humans and organisations who consider the relevant x-risk and (ii) the possibility that they have based their conclusion on flawed information or an incomplete state of knowledge or understanding.

7 CHAPTER SEVEN: CONCLUSION

This thesis has explored how the LHC Controversy can inspire the design of a broad legal intervention for existential risk mitigation (the LHC-inspired intervention). This objective is reflected in the thesis's central research question, which is formulated in the following way: How can the LHC Controversy inspire the design of a broad legal intervention to mitigate existential risk? The motivation for this research objective and question originated from a preliminary view that the past can be indicative when thinking about future strategies for x-risk mitigation. In effect, the thesis has looked at the LHC Controversy as a drafting guide for the law's response to certain risks within the x-risk landscape.

Whether or not the objective risk of catastrophe is zero, the LHC Controversy can offer valuable insights for x-risk mitigation beyond the context of experimental particle physics. For example, and aligning with the findings in this thesis, Bostrom and Ćirković have noted the possibility that (i) the relevant safety argument might be flawed, (ii) the experts who have assessed the risk scenario might be consciously or unconsciously biased, and (iii) the experts might have formed part of the same professional community who wants the experiment to proceed.[1353] These sorts of insights are relevant irrespective of whether or not particle accelerator experiments can trigger an existential catastrophe. The insights are also applicable beyond the field of experimental particle physics. This is why Chapter 1 claimed that "the value of this thesis is not contingent on the correctness of the assertion that running the LHC, the RHIC, or any other accelerator can cause an existential catastrophe."

The thesis has examined several interrelated themes to answer the main research question. Among other things, it examined (i) the wider x-risk landscape and (ii) selected aspects of the LHC Controversy focusing on the *Sancho case*. Firstly, the x-risk landscape was examined because it is within this landscape that the broad legal intervention will sit and operate. However, as explained in Chapter 6, the LHC-inspired intervention is not intended to function as an all-purpose mechanism mitigating risk from all hazards within the x-risk landscape. Despite this caveat, the bird's-eye view

[1353] Bostrom and Ćirković, 'Introduction' (n 12) 19.

of the wider risk landscape in Chapter 2 has been instrumental in investigating the opportunities and limits that may affect the LHC-inspired intervention.

Secondly, the LHC Controversy was examined to uncover certain difficulties or 'review barriers' that an external reviewer may face if invited to address the disagreements and ambiguity about whether particle accelerator experiments pose an existential risk. An overview of the x-risk landscape and the LHC Controversy indicated the value of a legal intervention disconnected from probabilities of risk and the adjudication of competing scientific theories aiming to support or refuse the relevant x-risk. Instead, the thesis has advanced an intervention that revolves around analysing interconnected deficiency factors. Broadly speaking, these require analysis of (i) the actors who considered the relevant x-risk and (ii) the possibility that they have based their conclusion on imperfect information or an incomplete state of knowledge. The resulting analysis can—despite the existence of the review barriers outlined in Chapter 3—indicate that the x-risk assessment is not, at present, sufficiently reliable. The effect of this might be that the endeavour linked to an x-risk should not presently proceed.

The theoretical underpinnings of the LHC-inspired intervention have been supported by considering x-risk-related and legal literature dealing with, among other things, x-risk and the LHC Controversy. Notably, the thesis used Johnson's so-called meta-analysis to compartmentalise the deficiency factors. More specifically, and as explored in Chapter 6, the deficiency factors that might be analysed as part of assessing the reliability of the risk assessment can conveniently be compartmentalised into the following groups: (i) Defective Theoretical Groundings; (ii) Faulty Scientific Work; (iii) Credulity and Neglect; (iv) Bias and Influence. In rationalising and examining the legal intervention at great lengths, Chapter 6 ultimately answered the main research question. It summarised the LHC-inspired intervention as follows:

> The broad legal intervention in response to the LHC Controversy centres around giving an external reviewer a principled basis upon which to base a finding that some endeavours should not *presently* be allowed to go ahead since, and notwithstanding the review barriers or a combination thereof, there are reasonable doubts as to the *reliability* of the scientific work assessing existential risk linked to these endeavours.

A vital aspect of this thesis has been to formulate a legal mechanism which satisfies three self-imposed conditions. These conditions were crucial in steering the thesis towards a legal intervention that aligned with the main research aim and question. The self-imposed conditions were as follows: Firstly, the LHC-inspired intervention should be capable of circumventing the review barriers that may thwart distinct modes of external review in relation to the LHC Controversy. Secondly, it should utilise the ASP. Finally, the intervention should, with some degree of plausibility, find application beyond the context of the LHC Controversy context. Each of these conditions was, to some extent, the subject matter of a separate chapter in the thesis.[1354]

Ultimately, the LHC-inspired intervention satisfies these conditions. The first one has been addressed by advancing a mechanism that is disconnected from probabilities and competing scientific theories about the relevant x-risk. In other words, promoting an analysis of the interrelated deficiency factors enables an external reviewer to consider the reliability of the relevant assessment despite the review barriers that characterised the LHC Controversy.

The second condition has been addressed since the LHC-inspired intervention can utilise the ASP. For one, the ASP underpins the very process of analysing the deficiency factors. As outlined in Chapter 6, two deficiency factors—namely, (i) Defective Theoretical Groundings and (ii) Faulty Scientific Work—hold a strong connection to the passage of time. Both factors require consideration of concepts such as longevity, time, newness and so on. The ASP, which demands expansive thinking across time, can affect the interpretation and understanding of these concepts. Furthermore, the ASP can sway the length of any delay flowing from a finding that an endeavour should not presently go ahead since the underlying x-risk assessment is unreliable. It is also possible to attempt to use the ASP to justify stringent obligations being imposed on actors who allegedly create an x-risk. However, as demonstrated in this thesis, a legal mechanism (including the LHC-inspired intervention) can utilise the ASP in alternative ways.

[1354] Chapter 3 dealt with the first condition. Chapter 5 dealt with the second condition. Chapter 2 and Chapter 6 dealt with the third condition.

The final condition has also been addressed. Chapter 6 highlighted examples from the recent past which indicate that analysis of the deficiency factors might prove useful beyond the LHC Controversy. This was examined with reference to the Manhattan Project and the CUORE experiment. The prospect of future x-risk from unknown unknowns and the possibility of deconstructing the LHC-inspired intervention also support the view that the intervention might find application beyond the concerns associated with machines like the LHC and RHIC.

Existential risks are increasingly associated with various speculative hazards traceable to anthropogenic processes. Policymakers should do more to protect humanity's far future by mitigating x-risk. Looking to the past can offer an avenue for examining the law's response to certain x-risks that may arise in the future. In line with this approach, this thesis has considered and formulated a potential broad legal intervention for x-risk mitigation by considering what can be learnt from the disagreements and uncertainty as to whether high-energy physics experiments facilitated by particle accelerators (i) pose any x-risk to humanity, (ii) pose an unacceptable x-risk to humanity, or (iii) should be delayed or suspended.

BIBLIOGRAPHY

‘Physicists and the Bomb’ (2015) 11(201) Nature Physics <https://doi.org/10.1038/nphys3287> accessed 26 April 2023

Abraham K and Merrill R, ‘Scientific Uncertainty in the Courts’ [1986] Issues in Science and Technology 93

Adams F, ‘Long-term Astrophysical Processes’ in Nick Bostrom and Milan Ćirković (eds), *Global Catastrophic Risks* (OUP 2008)

Adams S, ‘"Honey I Blew Up the World!"? One Small Step Towards Filling the Regulatory "Black Hole" at the Intersection of High-Energy Particle Colliders and International Law’ [2009] Georgia Journal of international and Comparative Law 131

Althaus D and Gloor L, ‘Reducing Risks of Astronomical Suffering: A Neglected Priority’ (Center on Long-Term Risk 2016) <https://longtermrisk.org/reducing-risks-of-astronomical-suffering-a-neglected-priority/#link_ajs-fn-id_4-3131> accessed 24 April 2023

Amerasinghe C, *Principles of the Institutional Law of International Organizations* (2nd edn, Cambridge University Press 2005)

American Airlines Newsroom, ‘Return of the Boeing 737 MAX to Service’ (18 November 2020) <http://news.aa.com/news/news-details/2020/Return-of-the-Boeing-737-MAX-to-service-OPS-DIS-11/default.aspx> accessed 27 April 2023

Annas G, ‘Crimes Against the Human Species (‘Type II Crimes Against Humanity’ Explained), in Britta van Beers, Luigi Corrias and Wouter Werner (eds), *Humanity across International Law and Biolaw* (Cambridge University Press 2014)

Armstrong S and Sandberg A, ‘Eternity in Six Hours: Intergalactic Spreading of Intelligent Life and Sharpening the Fermi Paradox’ [2013] 1

Avin S, Wintle B C, Weitzdörfer J, Ó hÉigeartaigh S, Sutherland W J and Rees M J, ‘Classifying Global Catastrophic Risks’ [2018] Futures 20

Barfield W and Williams A, ‘Law, Cyborgs, and Technologically Enhanced Brains’ (2017) 2(1) Philosophies <https://doi.org/10.3390/philosophies2010006> accessed 03 May 2023

Baum S D, Armstrong S, Ekenstedt T, Häggström O, Hanson R, Kuhlemann K, Maas M M, Miller J D, Salmela M, Sandberg A, Sotala K, Torres P, Turchin, and Yampolskiy R V, ‘Long-term Trajectories of Human Civilization’ [2019] Foresight 53

Baum S, ‘The Great Downside Dilemma for Risky Emerging Technologies’ (2014) 89(12) Physica Scripta <https://iopscience.iop.org/article/10.1088/0031-8949/89/12/128004> accessed 20 April 2023

——, ‘Quantifying the Probability of Existential Catastrophe: A reply to Beard et al.’ (2020) 123 Futures <https://doi.org/10.1016/j.futures.2020.102608> accessed 24 April 2023

Baum S, Timothy Maher and Jacob Haqq-Misra, ‘Double Catastrophe: Intermittent Stratospheric Geoengineering Induced by Societal Collapse,’ [2013] Environment Systems & Decisions 168

Baumann T, ‘S-risks: An Introduction’, (Center for Reducing Suffering, 2017) <https://centerforreducingsuffering.org/intro/> accessed 03 May 2023

Bean D, Burns A and Parry I, *Injunctions* (13th edn, Sweet & Maxwell 2018)

Beard S and Holt L, ‘What are the Biggest Threats to Humanity?’ BBC (15 February 2019) <https://www.bbc.co.uk/news/world-47030233> accessed 20 April 2023

Beard S and Kaczmarek P, ‘On the Wrongness of Human Extinction’ [2019] Argumenta 85

Beard S, Rowe T and Fox J, ‘An Analysis and Evaluation of Methods Currently Used to Quantify the Likelihood of Existential Hazards’ (2020) 115 Futures <https://doi.org/10.1016/j.futures.2019.102469> accessed 24 April 2023

Beckstead N, ‘On the Overwhelming Importance of Shaping the Far Future’ (DPhil thesis, Rutgers University 2013)

——, Will we Eventually be Able to Colonize Other Stars? Notes From a Preliminary Review’ (Future of Humanity Institute, 22 June 2014) <https://www.fhi.ox.ac.uk/will-we-eventually-be-able-to-colonize-other-stars-notes-from-a-preliminary-review/> accessed 03 May 2023

——, ‘How Much Could Refuges Help us Recover from a Global Catastrophe?’ [2015] Futures 36

Beiker S, ‘Legal Aspects of Autonomous Driving’ [2012] Santa Clara Law Review 1145

Berggren C, ‘Good Things on the Rise: The One-Sided Worldview of Hans Rosling’ Kvartal (20 September 2018) <https://kvartal.se/artiklar/bra-saker-pa-uppgang-roslings-varldsbild-ar-ensidigt-positiv/> accessed 03 May 2023

Bethe H, ‘Ultimate Catastrophe?’ [1976] Bulletin of the Atomic Scientists 36

Beyleveld D and Brownsword R, ‘Emerging Technologies, Extreme Uncertainty, and the Principle of Rational Precautionary Reasoning’ [2012] Law, Innovation and Technology 35

Boggio A and Romano C, ‘Freedom of Research and the Right to Science, from Theory to Advocacy’ in Simona Giordano (ed), *The Freedom of Scientific Research: Bridging the Gap between Science and Society* (Manchester University Press 2019)

Bostrom N and Ćirković M, ‘Introduction’ in Nick Bostrom and Milan Ćirković (eds), *Global Catastrophic Risks* (OUP 2008)

Bostrom N, ‘Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards’ (2002) 9 Journal of Evolution and Technology

——, ‘Astronomical Waste: The Opportunity Cost of Delayed Technological Development’ [2003] Utilitas 308

——, ‘Letter from Utopia’ [2008] Studies in Ethics, Law, and Technology 1

——, ‘Existential Risk Prevention as Global Priority’ [2013] Global Policy 15

——, *Superintelligence: Paths, Dangers, Strategies* (OUP 2014)

——, ‘The Vulnerable World Hypothesis’ [2019] Global Policy 455

——, ‘Ethical Issues in Advanced Artificial Intelligence’ <https://www.fhi.ox.ac.uk/wp-content/uploads/ethical-issues-in-advanced-ai.pdf> accessed 25 April 2023

Bouwhuis S, 'The International Law Commission's Definition of International Organizations' [2012] International Organizations Law Review 451

Boyd M and Wilson N, ‘Existential Risks to Humanity Should Concern International Policymakers and More Could Be Done in Considering Them at the International Governance Level’ [2020] Risk Analysis 2303

Braun-Munzinger P and others, ‘SPC Report on LSAG Documents’ (SPC Panel 2008) <http://cdsweb.cern.ch/record/1113558/files/cer-002766289.pdf> accessed 28 April 2023

Burt G, ‘Five Ways Particle Accelerators Have Changed the World (Without a Higgs Boson in Sight)’ The Conversation (8 February 2016) <https://theconversation.com/five-ways-particle-accelerators-have-changed-the-world-without-a-higgs-boson-in-sight-54187> accessed 26 April 2023

Calogero F, ‘Might a Laboratory Experiment Destroy Planet Earth?’ [2000] Interdisciplinary Science Reviews 191

Cartlidge E, ‘Law and the End of the World’ Physics World (02 February 2010) <https://physicsworld.com/a/law-and-the-end-of-the-world/> accessed 28 April 2023

Centre for the Study of Existential Risk, ‘About us’ <https://www.cser.ac.uk/about-us/> accessed 20 April 2023

CERN, ‘A New Schedule for the LHC and its Successor <https://home.cern/news/news/accelerators/new-schedule-lhc-and-its-successor> accessed 27 April 2023

CERN, ‘Contribute to Society’ <https://home.cern/about/what-we-do/our-impact> accessed 26 April 2023

CERN, ‘Facts and Figures About the LHC’ <https://home.cern/resources/faqs/facts-and-figures-about-lhc> accessed 20 April 2023

CERN, ‘Fundamental Research’, <https://home.cern/about/what-we-do/our-research> accessed 28 April 2023

CERN, ‘Future Circular Collider’ <https://home.cern/science/accelerators/future-circular-collider> accessed 27 April 2023

CERN, ‘High-Luminosity LHC’ <https://home.cern/science/accelerators/high-luminosity-lhc> accessed 27 April 2023

CERN, ‘International Collaboration Publishes Concept Design for a Post-LHC Future Circular Collider at CERN’ <https://home.cern/news/press-release/accelerators/international-collaboration-publishes-concept-design-post-lhc> accessed 27 April 2023

CERN, ‘Is the Large Hadron Collider Dangerous?’ <https://home.cern/resources/faqs/large-hadron-collider-dangerous> accessed 28April 2023

CERN, ‘LHC Season 2: Facts & Figures’ <https://run2-13tev.web.cern.ch/background/lhc-season-2-facts-figures> accessed 1 February 2020

CERN, ‘LHC: The Guide’ (CERN-Brochure-2017-002-Eng 2017) 15 <https://home.cern/sites/home.web.cern.ch/files/2018-07/CERN-Brochure-2017-002-Eng_0.pdf> accessed 28 April 2023

CERN, ‘Origins’ <https://timeline.web.cern.ch/origins>

CERN, ‘Our Mission’ <https://home.cern/about/who-we-are/our-mission> accessed 27 April 2023

CERN, ‘Our People’ <https://home.cern/about/who-we-are/our-people> accessed 26 April 2023.

CERN, ‘The History of CERN’ <https://timeline.web.cern.ch/taxonomy/term/89> accessed 28 April 2023

CERN, ‘The Large Hadron Collider’ <https://home.cern/science/accelerators/large-hadron-collider> accessed 20 April 2023

CERN, ‘The Safety of the LHC’ <https://home.cern/science/accelerators/large-hadron-collider/safety-lhc > accessed 28 April 2023

CERN, ‘The Standard Model’ <https://home.cern/science/physics/standard-model> accessed 28 April 2023

CERN, ‘Where Did it all Begin?’ <https://home.cern/about/who-we-are/our-history> accessed 28 April 2023

CERNCOURIER, ‘Procurement at the Forefront of Technology’ (31 August 2018) <https://cerncourier.com/a/procurement-at-the-forefront-of-technology/> accessed 28 April 2023

Ćirković M, ‘Resource Letter: PEs-1: Physical Eschatology’ [2002] American Journal of Physics 122

——, ‘Forecast for the Next Eon: Applied Cosmology and the Long-Term Fate of Intelligent Beings’ [2004] Foundations of Physics 239

Ćirković M, Sandberg A and Bostrom N, ‘Anthropic Shadow: Observation Selection Effects and Human Extinction Risks’ [2010] Risk Analysis 1495

Cogen M, 'Membership, Associate Membership and Pre-Accession Arrangements of CERN, ESO, ESA, and EUMETSAT' [2012] International Organizations Law Review 145

Conselice C, Wilkinson A, Duncan K and Mortlock A, ‘The Evolution of Galaxy Number Density at z < 8 and its Implications’ (2016) 830(83) The Astrophysical Journal <https://iopscience.iop.org/article/10.3847/0004-637X/830/2/83> accessed 24 April 2023

Cotton-Barratt O and Ord T, ‘Existential Risk and Existential Hope: Definitions’, (Future of Humanity Institute – Technical Report #2015-1) <http://www.fhi.ox.ac.uk/Existential-risk-and-existential-hope.pdf> accessed 24 April 2023

Cotton-Barratt O, Daniel M and Sandberg A, ‘Defence in Depth Against Human Extinction: Prevention, Response, Resilience, and Why They All Matter’ [2020] Global Policy 271

Cotton-Barratt O, Farquhar S, Halstead J, Schubert S and Snyder-Beattie A, ‘Global Catastrophic Risks’ (Global Challenges Foundation 2016)

<https://globalchallenges.org/wp-content/uploads/2019/07/Global-Catastrophic-Risk-Annual-Report-2016.pdf> accessed 24 April 2023

Crawford J, *Brownlie's Principles of Public International Law* (8th edn, OUP 2012)

Currie A and Ó hÉigeartaigh S, 'Working Together to face Humanity's Greatest Threats: Introduction to the Future of Research on Catastrophic and Existential Risk' [2018] Futures 1

Dafoe A and Russell S, 'Yes, We are Worried About the Existential Risk of Artificial Intelligence' (2016) MIT Technology Review <https://www.technologyreview.com/s/602776/yes-we-are-worried-about-the-existential-risk-of-artificial-intelligence/> accessed 25 April 2023

Dahlgreen W, 'Inside the Mind of an Optimist', YouGov (13 May 2015) <https://yougov.co.uk/topics/politics/articles-reports/2015/05/13/optimists-about-world> accessed 03 May 2023

——, 'Chinese People are Most Likely to Feel the World is Getting Better' YouGov (05 Jan 2016) <https://yougov.co.uk/topics/lifestyle/articles-reports/2016/01/05/chinese-people-are-most-optimistic-world> accessed 03 May 2023

Dana D, 'A Behavioral Economic Defense of the Precautionary Principle' [2003] Northwestern University Law Review 1315

Daniel M, 'S-risks: Why they are the Worst Existential Risks, and how to Prevent them' (Center on Long-Term Risk 2017) <https://longtermrisk.org/s-risks-talk-eag-boston-2017/> accessed 24 April 2023

Davey T, '55 Years After Preventing Nuclear Attack, Arkhipov Honored With Inaugural Future of Life Award' (Future of Life Institute 2017) <https://futureoflife.org/2017/10/27/55-years-preventing-nuclear-attack-arkhipov-honored-inaugural-future-life-award/> accessed 24 April 2023

——, ‘Benefits & Risks of Biotechnology’ (Future of Life Institute 2018) <https://futureoflife.org/background/benefits-risks-biotechnology/> accessed 25 April 2023

Davies J, 'Admissibility of Scientific Evidence in Courts' [2005] Medicine and Law 243

Dredge S, ‘Artificial Intelligence will Become Strong Enough to be a Concern, says Bill Gates’, The Guardian (29 January 2015) <https://www.theguardian.com/technology/2015/jan/29/artificial-intelligence-strong-concern-bill-gates> accessed 25 April 2023

Duff R A, ‘Criminalizing Endangerment’ [2005] Louisiana Law Review 941

Ellis J and others, ‘Review of the Safety of LHC Collisions’ (LHC Safety Assessment Group 2008) <http://lsag.web.cern.ch/LSAG-Report.pdf> accessed 27 April 2023

Ellis J, ‘The LHC is Safe’ (CERN Colloquium, 14 August 2008, <https://cdsweb.cern.ch/record/1120625> accessed 24 April 2023

Etchells P, ‘Declinism: Is the World Actually Getting Worse?’ The Guardian (16 January 2015) <https://www.theguardian.com/science/head-quarters/2015/jan/16/declinism-is-the-world-actually-getting-worse> accessed 03 May 2023

Etzioni A and Etzioni O, ‘Should Artificial Intelligence be Regulated?’ (2017) 33(4) Issues in Science and Technology <http://issues.org/33-4/perspective-should-artificial-intelligence-be-regulated/> accessed 06 November 2020

Farquhar S, Halstead J, Cotton-Barratt O, Schubert S, Belfield H and Snyder-Beattie A, ‘Existential Risk: Diplomacy and Governance’ (Global Priorities Project 2017) <https://www.fhi.ox.ac.uk/wp-content/uploads/Existential-Risks-2017-01-23.pdf> accessed 24 April 2023

Fenwick C, ‘Longtermism: A Call to Protect Future Generations’ 80,000 Hours, 28 March 2023) <https://80000hours.org/articles/future-generations/> accessed 04 May 2023

Ferguson J, ‘Scientific Inquiry and the First Amendment’ [1979] Cornell Law Review 639

Finneron-Burns E, ‘Contractualism and the Non-identity Problem’ [2016] Ethical Theory and Moral Practice 1151

——,, ‘What’s Wrong with Human Extinction?’ [2017] Canadian Journal of Philosophy 327

Fisher R, ‘The Moments that Could have Accidentally Ended Humanity’ BBC (19 February 2021) <https://www.bbc.com/future/article/20210217-the-moments-that-we-could-have-destroyed-humanity> accessed 20 April 2023

——, ‘The Perils Of Short-Termism: Civilisation’s Greatest Threat’ BBC Future (14 December 2020) <https://www.bbc.com/future/article/20190109-the-perils-of-short-termism-civilisations-greatest-threat> accessed 27 April 2023

Frick J, ‘On the Survival of Humanity’ [2017] Canadian Journal of Philosophy 344

Future of Humanity Institute, ‘Research Areas’ <https://www.fhi.ox.ac.uk/research/research-areas/> accessed 20 April 2023

Ghosh P, ‘What next for the Large Hadron Collider?’ BBC (5 April 2015) <https://www.bbc.co.uk/news/science-environment-32106990> accessed 28 April 2023

——, ‘Shock Result in Particle Experiment Could Spark Physics Revolution’ BBC (7 April 2022) <https://www.bbc.co.uk/news/science-environment-60993523> accessed 04 May 2023

Giddings S and Mangano M, ‘Astrophysical Implications of Hypothetical Stable TeV-scale Black Holes’ (2008) 78)(3) Physical Review D <http://dx.doi.org/10.1103/PhysRevD.78.035009> accessed 28 April 2023

Greshko M, ‘Elusive Elementary Particle Finally Caught Decaying—Get the Facts’ National Geographic (29 August 2018) <https://www.nationalgeographic.com/science/2018/08/higgs-boson-decay-quarks-lhc-standard-model-physics/> accessed 26 April 2023

Gruner B, 'A New Hope for International Space Law: Incorporating Nineteenth Century First Possession Principles into the 1967 Space Treaty for the Colonization of Outer Space in the Twenty-First Century' [2004] Seton Hall Law Review 299

Häggström O, *Here Be Dragons: Science, Technology and the Future of Humanity* (OUP 2016)

Hall D W and Pesenti J, ‘Growing the Artificial Intelligence Industry in the UK’ (Independent report, the Department for Digital, Culture, Media & Sport and the Department for Business, Energy & Industrial Strategy 2017) <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf> accessed 26 April 2023

Halstead J, ‘Stratospheric Aerosol Injection Research and Existential Risk’ [2018] Futures 63

Hansson S, ‘Risk’ (The Stanford Encyclopedia of Philosophy 2018) <https://plato.stanford.edu/archives/fall2018/entries/risk/> accessed 24 April 2023

Hawking S, *Brief Answers to the Big Questions* (John Murray Publishers 2018)

Hervey, T Cryer R and Sokhi-Bulley B, ‘Legal Research Methodologies in European Union & International Law: Research Notes (Part 1)’ (2007) 3 Journal of Contemporary European Research 161

Hoecke M V, ‘Legal Doctrine: Which Method(s) for What Kind of Discipline?’ in Mark Van Hoecke (ed), *Methodologies of Legal Research: Which Kind of Method for What Kind of Discipline?* (Hart Publishing 2011)

Hudson A, *Equity and Trusts* (9th edn, Routledge 2016)

Hutchinson T, ‘The Doctrinal Method: Incorporating Interdisciplinary Methods in Reforming the Law’ [2015] Erasmus Law Review 130

Institute for International Law and Justice, ‘Global Administrative Law’ <https://www.iilj.org/gal/> accessed 28 April 2023

Institute for International Law and Justice, ‘Projects’ <https://www.iilj.org/projects/> accessed 28 April 2023

Interactions, ‘CUORE: The Coldest Heart in the Known Universe’ (2014) <https://www.interactions.org/press-release/infn-cuore> accessed 25 April 2023

Jaffe R and others, ‘Review of Speculative "Disaster Scenarios" at RHIC’ [2000] Reviews of Modern Physics 1125

Johnson E, ‘Could Bad Judging Cause the Earth to Be Sucked Into a Black Hole? Maybe.’ (PrawfsBlawg, 21 October 2008) <https://prawfsblawg.blogs.com/prawfsblawg/2008/10/could-bad-judgi.html> accessed 28 April 2023

——, ‘Culture and Inscrutable Science: An Analytical Method for Preliminary Injunctions in Extreme Cases’ (PrawfsBlawg, 24 October 2008) <https://prawfsblawg.blogs.com/prawfsblawg/2008/10/culture-and-ins.html> accessed 28 April 2023

——, ‘Scientocracy and the Need for Judicial Process’ (PrawfsBlawg, 10 November 2008) <https://prawfsblawg.blogs.com/prawfsblawg/2008/11/scientocracy-an.html> accessed 28 April 2023

——, ‘The Black Hole Case: The Injunction Against the end of the World’ [2009] Tennessee Law Review 819

——, ‘Agencies and Science-Experiment Risk’ [2016] University of Illinois Law Review 527

——, ‘Judicial Review of Uncertain Risks in Scientific Research’ in Gilles Motet and Corinne Bieder (eds), *The Illusion of Risk Control: What does it Take to Live with Uncertainty?* (SpringerOpen 2017)

——, 'Uncertain Risk, Science Experiments, and the Courts' [2021] Florida State University Law Review 333

Jones E, ‘Where is Everybody? An Account of Fermi’s Question’ (Technical Report LA-10311-MS, Los Alamos National Laboratory 1985), in Anders Sandberg, Eric Drexler and Toby Ord, ‘Dissolving the Fermi Paradox’ (Future of Humanity Institute 2018) <https://arxiv.org/pdf/1806.02404.pdf> accessed 24 April 2023

Jones N, O’Brien M, Ryan T, ‘Representation of Future Generations in United Kingdom Policymaking’ [2018] Futures 153

Kaczmarek P and Beard S, ‘Human Extinction and Our Obligations to the Past’ [2020] Utilitas 199

Kaczmarek P, ‘How Much is Rule-Consequentialism Really Willing to Give Up to Save the Future of Humanity?’ [2017] Utilitas 239

Kaplan S and Garrick J, ‘On The Quantitative Definition of Risk’ [1981] Risk Analysis 11

Kemp L and Rhodes C, ‘The Cartography of Global Catastrophic Governance’ (Report of Centre for the Study of Existential Risk 2020)

Kent A, ‘A Critical Look at Risk Assessments for Global Catastrophes’ [2004] Risk Analysis 157

Kilbourne E, ‘Plagues and Pandemics: Past, Present, and Future’ in Nick Bostrom and Milan Ćirković (eds), *Global Catastrophic Risks* (OUP 2008)

Klabbers J, *An Introduction to International Organizations Law* (3rd edn, Cambridge University Press 2015)

——, *International Law* (2nd edn, Cambridge University Press 2017)

Kohler M, ‘Regulating Research and Technology Risks: Part I – Research Risks’ The Regulatory Institute (10 September 2017) <https://www.howtoregulate.org/regulating-research-technology-risks-part-i-research-risks/#more-248> accessed 04 May 2023

——, ‘Research and Technology Risks: Part III – Risk Classification’ The Regulatory Institute (3 March 2018) <https://www.howtoregulate.org/classification-research-technology-risks> accessed 03 May 2023

——, ‘Research and Technology Risks: Part IV – A Prototype Regulation’ The Regulatory Institute (3 March 2018) <https://www.howtoregulate.org/prototype-regulation-research-technology/> accessed 04 May 2023

——, ‘The Need for Global Protections Against Existential Risks’ The Regulatory Review (11 June 2020) <https://www.theregreview.org/2020/06/11/kohler-need-global-protections-against-existential-risks/> accessed 03 May 2023

Konopinski E, Marvin C, and Teller E, ‘Ignition of the Atmosphere with Nuclear Bombs’ (1946) <https://sgp.fas.org/othergov/doe/lanl/docs1/00329010.pdf> accessed 20 April 2023

Kraemer J and Gostin L, ‘The Limits of Government Regulation of Science’ [2012] 1047

Kuhlemann K, ‘Complexity, Creeping Normalcy and Conceit: Sexy and Unsexy Risks’ [2019] Foresight 35

Langbroek P and others, ‘Methodology of Legal Research: Challenges and Opportunities’ [2017] Utrecht Law Review 1

Larsen K, *Particle Panic!: How Popular Media and Popularized Science Feed Public Fears of Particle Accelerator Experiments* (Springer 2019)

Lauta K C, ‘Dysfunction and Disruption: Disaster Risks and the Law’ [2017] European Journal of Risk Regulation 83

Legal Priorities Project, ‘Legal Priorities Research: A Research Agenda’ (2021) <https://www.legalpriorities.org/research_agenda.pdf> accessed 20 April 2023

Legal Priorities Project, <https://www.legalpriorities.org/> accessed 20 April 2023

Lehmann H, *No Canary in the Quanta: Who Gets to Decide if the Large Hadron Collider is Worth Gambling Our Planet?* (Green Swan 2009)

Leiss W, *The Doom Loop in the Financial Sector, and Other Black Holes of Risk* (University of Ottawa Press 2010)

Leslie J, *The End of the World: The Science and Ethics of Human Extinction* (Routledge 1996)

Liu Hin-Yan, Lauta K C and Maas M M, ‘Governing Boring Apocalypses: A New Typology of Existential Vulnerabilities and Exposures for Existential Risk Research’ [2018] Futures 6

——, Lauta K and Maas M M, ‘Apocalypse Now? Initial Lessons from the Covid-19 Pandemic for the Governance of Existential and Global Catastrophic Risks’ [2020] Journal of International Humanitarian Legal Studies 295

Logan D, ‘Known Knowns, Known Unknowns, Unknown Unknowns and the Propagation of Scientific Enquiry’ [2009] 712

MacAskill W, ‘What is Longtermism?’ BBC Futures (8 August 2022) <https://www.bbc.com/future/article/20220805-what-is-longtermism-and-why-does-it-matter> accessed 20 April 2023

MacAskill W, *What We Owe The Future: A Million-Year View* (Oneworld Publications 2022)

MacDonald E, ‘GAL - and The End Of The World As We Know It’ (20 January 2009)<https://globaladminlaw.blogspot.com/2009/01/global-administrative-law-and-end-of.html> accessed 28 April 2023

Mack K, *The End of Everything: (Astrophysically Speaking)* (Allen Lane 2020)

Marc Cogen, An Introduction to European Intergovernmental Organizations (Ashgate Publishing 2015)

Marchant G, Allenby B and Herkert J (eds), *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem* (Springer 2011)

Martínez E and Winter C, ‘Foundations of Experimental Longtermist Jurisprudence’ in StefanMagen and Karolina Prochownik (eds), *Advances in Experimental Philosophy of Law* (forthcoming)

Martinez E and Winter C, ‘Protecting Future Generations: A Global Survey of Legal Academics’, (LPP Working Paper Series No 1-2021) <https://www.legalpriorities.org/documents/1%20-%20Protecting%20Future%20Generations.pdf> accessed 27 April 2023

Matus K, ‘Existential Risk: Challenges for Risk Regulation’ (2014) <http://www.lse.ac.uk/accounting/assets/CARR/documents/R-R/2014-Winter/Risk-and-Regulation-28-existential-risk.pdf> accessed 24 April 2023

McKinnon C, ‘Endangering Humanity: An International Crime?’ [2017] Canadian Journal of Philosophy 395

Miller J and Felton D, ‘The Fermi Paradox, Bayes' Rule, and Existential Risk Management’ [2017] Futures 44

Moore D, ‘The Preliminary Injunction Standard: Understanding the Public Interest Factor’ [2019] Michigan Law Review 939

Moynihan T, X-*Risk: How Humanity Discovered its Own Extinction* (Urbanomic 2020)

Muehlhauser L, ‘How big a deal was the Industrial Revolution?’ <https://lukemuehlhauser.com/industrial-revolution/> accessed 25 April 2023

Müller V and Bostrom N, ‘Future Progress in Artificial Intelligence: A Survey of Expert Opinion’ in Vincent Müller (eds), Fundamental Issues of Artificial Intelligence (Springer 2016)

Neuwirth R, 'Law and Magic: A(Nother) Paradox' [2014] Thomas Jefferson Law Review 139

Nindler R, ‘The United Nation’s Capability to Manage Existential Risks with a Focus on Artificial Intelligence’ [2019] International Community Law Review 5

Nouri A and Chyba C, ‘Biotechnology and Biosecurity’ in Nick Bostrom and Milan Ćirković (eds), *Global Catastrophic Risks* (Oxford University Press 2008)

Oberdiek J, ‘Towards a Right Against Risking’ [2009] Law and Philosophy 367

Ord T, ‘These are not the Probabilities you are Looking For’ (Practical Ethics, 15 April 2008 <http://blog.practicalethics.ox.ac.uk/2008/04/these-are-not-the-probabilities-you-are-looking-for/> accessed 04 May 2023

——, ‘Will We Cause Our Own Extinction? (Public Lecture, Future of Humanity Institute 2015) <https://www.cser.ac.uk/events/extinction/> accessed 24 April 2023

——, *The Precipice* (Bloomsbury Publishing 2020)

Ord T, Hillerbrand R and Sandberg A, ‘Probing the Improbable: Methodological Challenges for Risks with low Probabilities and High Stakes’ [2010] Journal of Risk Research 191

Our World in Data, https://ourworldindata.org/

Parfit D, *Reasons and Persons* (Oxford University Press 1984)

Pearce R and Barr W, *Pearce & Stevens' Trusts and Equitable Obligations* (7th edn OUP 2018)

Peterson M C, ‘The "Sancho" Effect: Why the Large Hadron Collider Won't Destroy the World, and how it Could Improve Science in the Courts’ [2014] American Bar Association 303

Petrenko A and McArthur D, ‘Between Same-Sex Marriages and the Large Hadron Collider: Making Sense of the Precautionary Principle’ [2010] Science and Engineering Ethics 591

Piccirillo L, ‘Big Science and Small Science: Reflections on the Relationship Between Science and Society from the Perspective of Physics’ in Simona Giordano (ed), *The Freedom of Scientific Research: Bridging the Gap between Science and Society* (Manchester University Press 2019)

Pinker S, ‘Is the World Getting Better or Worse? A Look at the Numbers’ (TED Talk, 21 May 2018) <https://www.youtube.com/watch?v=yCm9Ng0bbEQ> accessed 03 May 2023

Plaga R, ‘On the Potential Catastrophic Risk From Metastable Quantum-Black Holes Produced at Particle Colliders’ (2008) <https://doi.org/10.48550/arXiv.0808.1415> accessed 04 May 2023

Posner R, *Catastrophe: Risk and Response* (Oxford University Press 2004)

——, ‘Public Policy Towards Catastrophe’ in Nick Bostrom and Milan Ćirković (eds), *Global Catastrophic Risks* (OUP 2008)

Rees M, *Our Final Century: Will Civilisation Survive the Twenty-First Century?* (Arrow Books 2003)

——, ‘Denial of Catastrophic Risks’ (2013) 339(6124) Science <https://www.science.org/doi/10.1126/science.1236756> accessed 20 April 2023

——, *On the Future Prospects for Humanity* (Princeton University Press 2018)

Reinisch A, *International Organizations Before National Courts* (Cambridge University Press 2000)

Rhodes C, ‘The Future Governance of Biotechnology’ in Simon Whitby and others (eds), *Preventing Biological Threats: What You Can Do* (Bradford Disarmament Research Centre 2015)

Rodrigues R, ‘Legal and Human Rights Issues of AI: Gaps, Challenges and Vulnerabilities’ (2020) 4 Journal of Responsible Technology <https://doi.org/10.1016/j.jrt.2020.100005> accessed 03 May 2023

Roser M, ‘Most of us are Wrong About how the World has Changed (Especially Those who are Pessimistic About the Future)’ Our World in Data (27 July 2018) <https://ourworldindata.org/wrong-about-the-world> accessed 03 May 2023

——, ‘The Short History of Global Living Conditions and Why it Matters That we Know it’ Our World in Data (2020) <https://ourworldindata.org/a-history-of-global-living-conditions-in-5-charts> accessed 03 May 2023

Rosling H and Rosling O, ‘How not to be Ignorant About the World’ (TED Talk, 11 September 2014) <https://www.youtube.com/watch?v=Sm5xF-UYgdg&t=216s> accessed 03 May 2023

Rosling H, Rosling O, and Rönnlund A R, *Factfulness: Ten Reasons We're Wrong About the World – And Why Things Are Better Than You Think* (Sceptre 2018)

Rowe T and Beard S, ‘Probabilities, Methodologies and the Evidence Base in Existential Risk Assessments’ (Centre for the Study of Existential Risk 2018) <http://eprints.lse.ac.uk/89506/> accessed 25 April 2023

Royal Academy of Engineering, ‘Innovation in Autonomous Systems’ (Summary of an event held at the Royal Academy of Engineering, 2015) 7 <https://www.raeng.org.uk/publications/reports/innovation-in-autonomous-systems> accessed 26 April 2023

Rumsfeld D (DoD News Briefing, 12 February 2002) <https://archive.ph/20180320091111/http://archive.defense.gov/Transcripts/Transcript.aspx?TranscriptID=2636> accessed 25 April 2023.

Sagan C, ‘Nuclear War and Climatic Catastrophe: Some Policy Implications’ [1983] Foreign Affairs 257
——, *Pale Blue Dot: A Vision of the Human Future in Space* (Random House Publishing Group 1994)

Sandberg A, ‘Extinction Risks and Particle Physics: When Are They Worth it?’ (Practical Ethics, 29 March 2008) <http://blog.practicalethics.ox.ac.uk/2008/03/extinction-risks-and-particle-physics-when-are-they-worth-it/> accessed 04 May 2023

——, ‘Cool Risks Outside the Envelope of Nature’ (Andart II, 22 October 2014) <https://aleph.se/andart2/risk/existential-risk-risk/cool-risks-outside-the-envelope-of-nature/> accessed 04 May 2023

——, ‘What kinds of grand futures are there?’ (Andart, 5 July 2018) <http://aleph.se/andart2/space/what-kinds-of-grand-futures-are-there/> accessed 03 May 2023

——, ‘Grand Futures: How Much is there to Hope For, How Much is at Stake?’ (The Cambridge Conference on Catastrophic Risk 2020, The Centre for the Study of Existential Risk, 9 December 2020) <https://www.youtube.com/watch?v=9a5uYOoFJvE&t=6s> accessed 03 May 2023

Sandberg A and Bostrom N, ‘Global Catastrophic Risks Survey’, (Future of Humanity Institute – Technical Report #2008-1) <https://www.fhi.ox.ac.uk/reports/2008-1.pdf> accessed 26 April 2023

Sandberg A, Eric Drexler and Toby Ord, ‘Dissolving the Fermi Paradox’ (Future of Humanity Institute 2018) <https://arxiv.org/pdf/1806.02404.pdf> accessed 24 April 2023

Sandberg A, Matheny J and Ćirković M, ‘How can we Reduce the Risk of Human Extinction?’ Bulletin of the Atomic Scientists (9 September 2008) <https://thebulletin.org/2008/09/how-can-we-reduce-the-risk-of-human-extinction/> accessed 03 May 2023

Santosuosso A, Sellaroli V and Fabio E, ‘What Constitutional Protection for Freedom of Scientific Research?’ [2007] Journal of Medical Ethics 342

Scherer M, ‘Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies’ [2016] Harvard Journal of Law & Technology 353

Schneier B, ‘Resources on Existential Risk’ (Syllabus For: Catastrophic Risk: Technologies and Policies Berkman Center for Internet and Society’) <https://futureoflife.org/data/documents/Existential%20Risk%20Resources%20(2015-08-24).pdf> accessed 20 April 2023

Schrama W, ‘How to Carry out Interdisciplinary Legal Research: Some Experiences with an Interdisciplinary Research Method’ [2011] Utrecht Law Review 147

Schubert S, Caviola L and Faber N, ‘The Psychology of Existential Risk: Moral Judgments about Human Extinction’ (2019) 9 Scientific Reports <https://doi.org/10.1038/s41598-019-50145-9> accessed 03 May 2023

Schwartz A, ‘A "Dogma of Empiricism" Revisited: Daubert v. Merrell Dow Pharmaceuticals, Inc. and the Need to Resurrect the Philosophical Insight of Frye V. United States’ [1997] Harvard Journal of Law & Technology 149

Setzer J and Byrnes R, ‘Global Trends in Climate Change Litigation: 2020 Snapshot’ (2020) <https://www.lse.ac.uk/granthaminstitute/wp-content/uploads/2020/07/Global-trends-in-climate-change-litigation_2020-snapshot.pdf> accessed 28 April 2023

Singer M, ‘Jurisdictional Immunity of International Organizations: Human Rights and Functional Necessity Concerns’ [1995] Virginia Journal of International Law 53

Slovic P, ‘“If I look at the Mass I will Never act”: Psychic Numbing and Genocide’ [2007] Judgment and Decision Making 79

Snyder-Beattie A, Ord T and Bonsall M, ‘An Upper Bound for the Background Rate of Human Extinction’ (2019) 9 Scientific Reports <https://doi.org/10.1038/s41598-019-47540-7> accessed 24 April 2023

Stark F, *Culpable Carelessness: Recklessness and Negligence in the Criminal Law* (Cambridge University Press 2016)

Steele J, *Tort Law – Text, Cases and Materials* (3rd edn, Oxford University Press 2014)

Stone P and others, 'Artificial Intelligence and Life in 2030' (One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, Stanford University 2016) <https://ai100.stanford.edu/sites/default/files/ai100report10032016fnl_singles.pdf> accessed 06 November 2020

Sunstein C, *Averting Catastrophe: Decision Theory for COVID-19, Climate Change, and Potential Disasters of All Kinds* (NYU Press 2021)

Tardif A, 'How we can Benefit from Advancing Artificial General Intelligence (AGI)' Unite.AI (27 September 2020) <https://www.unite.ai/artificial-general-intelligence-agi/> accessed 26 April 2023

Tegmark M and Bostrom N, 'How Unlikely is a Doomsday Catastrophe?' [2005] Nature 754

Tegmark M, 'Benefits & Risks of Artificial Intelligence' (Future of Life Institute 2015) <https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/> accessed 25 April 2023

The Gapminder Foundation <https://www.gapminder.org/> accessed 03 May 2023

The Regulatory Institute <https://www.howtoregulate.org/> accessed 03 May 2023

The Royal Society, 'Geoengineering the Climate: Science, Governance and Uncertainty' (2009) <https://royalsociety.org/topics-policy/publications/%202009/geoengineering-climate/> accessed 25 April 2023

Thierer A, ‘The Pacing Problem and the Future of Technology Regulation’ The Mercatus Center (8 August 2018) <https://www.mercatus.org/bridge/commentary/pacing-problem-and-future-technology-regulation> accessed 27 April 2023

Todd B, ‘The Case for Reducing Existential Risks’, (80,000 Hours, October 2017) <https://80000hours.org/articles/existential-risks/> accessed 03 May 2023

Torres P, ‘Problems with Defining an Existential Risk’ (IEET 2015) <https://ieet.org/index.php/IEET2/more/torres20150121> accessed 24 April 2023

——, *The End: What Science and Religion Tell Us about the Apocalypse* (Pitchstone Publishing 2016)

——, *Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks* (Pitchstone Publishing 2017)

——, ‘Space Colonization and Suffering Risks: Reassessing the “Maxipok Rule”’ [2018] Futures 74

——, ‘Existential Risks: A Philosophical Analysis’ [2019] Inquiry 614

——, ‘Facing Disaster: The Great Challenges Framework’ [2019] Foresight 4

——, ‘Why an Existential Risk Expert Finds Hope in Humanity’s Certain Doom’ (Medium, 16 December 2019) <https://onezero.medium.com/rebelling-against-extinction-d7e112979bed> accessed 27 April 2023

Turchin A and Denkenberger D, ‘Global Catastrophic and Existential Risks Communication Scale’ [2018] Futures 27

United States Census Bureau, ‘Historical Estimates of World Population’ (Last Revised: 5 December 2022) <https://www.census.gov/data/tables/time-

series/demo/international-programs/historical-est-worldpop.html> accessed 25 April 2023

Verdirame G, ‘For China, a Legal Reckoning is Coming: We Need a new International Treaty to Deal with Risks to the Future of Humanity’ UnHerd (21 April 2020) <https://unherd.com/2020/04/for-china-a-legal-reckoning-is-coming/> accessed 03 May 2023.

Verschuuren J, ‘The State of the Netherlands v Urgenda Foundation: The Hague Court of Appeal Upholds Judgment Requiring the Netherlands to Further Reduce its Greenhouse gas Emissions’ [2019] Review of European, Comparative & International Environmental Law 94

Villmer M, ‘Procedural Squabbling Ahead of Global Annihilation: Strengthening the National Environmental Policy Act in a New Technological Era’ [2010] Florida Coastal Law Review 321

Voeneky S, ‘Human Rights and Legitimate Governance of Existential and Global Catastrophic Risks’ in Silja Voeneky and Gerald Neuman (eds), *Human Rights, Democracy, and Legitimacy in a World of Disorder* (Cambridge University Press 2018)

Vöneky S, ‘The Public International Law Perspective on Evaluating Existential Risks’ (Cambridge Conference on Catastrophic Risk 2018, The Centre for the Study of Existential Risk, 17 April 2019 <https://www.youtube.com/watch?v=O6ydGbBv2m0> accessed 26 April 2023

Voss R, ‘CERN: A global project’ (2017) 878 Journal of Physics: Conference Series <https://iopscience.iop.org/article/10.1088/1742-6596/878/1/012001> accessed 27 April 2023

Wansley M, ‘Regulation of Emerging Risks’ [2016] Vanderbilt Law Review 401

Welsch C, ‘What Have Particle Accelerators Ever Done For Us?’ Physics World (20 August 2019) <https://physicsworld.com/a/what-have-particle-accelerators-ever-done-for-us/> accessed 26 April 2023

Westin U and others, ‘Global Catastrophic Risks 2020’ (Global Challenges Foundation Annual Report 2020) <https://globalchallenges.org/wp-content/uploads/Global-Catastrophic-Risks-2020-Annual-Report.pdf> accessed 24 April 2023

WHO, ‘About WHO’ <https://www.who.int/about> accessed 25 April 2023

Wiener, J ‘The Tragedy of the Uncommons: On the Politics of Apocalypse’ [2016] Global Policy 67

Wilczek F, ‘Big Troubles, Imagined and Real’ in Nick Bostrom and Milan Ćirković (eds), *Global Catastrophic Risks* (OUP 2008)

Wilson G, ‘Minimizing Global Catastrophic and Existential Risks from Emerging Technologies Through International Law’ [2013] Virginia Environmental Law Journal 307

Wilson M W, 'Cultural Understandings of Risk and the Tyranny of the Experts' [2011] Oregon Law Review 113

Winter C, ‘The Value of Behavioral Economics for EU Judicial Decision-Making’ [2020] German Law Journal 240

Wittes B and Blum G, *The Future of Violence: Robots and Germs, Hackers and Drones: Confronting A New Age of Threat* (Basic Books 2015)

World Health Organisation, ‘WHO Coronavirus Disease (COVID-19) Dashboard’ <https://covid19.who.int/> accessed 25 April 2023

Yiu Y, ‘The Fear of Setting the Planet on Fire with a Nuclear Weapon’ Inside Science (15 July 2020) <https://insidescience.org/manhattan-project-legacy/atmosphere-on-fire> accessed 04 May 2023

Yudkowsky E, ‘LA-602 vs. RHIC Review’ (LessWrong, 19 June 2008) <https://www.lesswrong.com/posts/f3W7QbLBA2B7hk84y/la-602-vs-rhic-review> accessed 20 April 2023

——, ‘Cognitive Biases Potentially Affecting Judgment of Global Risks’ in Nick Bostrom and Milan Ćirković (eds), *Global Catastrophic Risks* (OUP 2008)