Title

Putting the Philosophy Back into Natural Philosophy: Ethics, Policy, and Using Philosophic Tools To Solve Scientific Problems

Author

Thresher, Ann

Publication Date

2022

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Putting the Philosophy Back into Natural Philosophy: Ethics, Policy, and Using Philosophic Tools to Solve Scientific Problems

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Philosophy

by

Ann C Thresher

Committee in charge:

Professor Craig Callender, Chair
Professor Ethan Bier
Professor Reuven Brandt
Professor Nancy Cartwright
Professor Elliott Sober

2022

The dissertation of Ann C Thresher is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California, San Diego

2022

DEDICATION

To Mum, Dad, and Bubbles.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Craig is entirely to blame for this dissertation. Not only did he convince me to come to San Diego by promising me palm trees, sun, and dolphins, but he then was foolish enough to suggest changing my dissertation topic five years into the program. It's impossible to thank him enough. He's been an incredible mentor, friend, and guide throughout my time at UCSD and I can honestly say I would never have reached this point in my career without him.

The same goes for Nancy, who has become like family to me over the last few years. I look forward to many more dinner parties, holidays, hikes, and camping trips. Also, our book eventually being published.

Both of you have taught me invaluable life skills including how to think critically, argue analytically, paddleboard (Craig), and throw switchblades (Nancy). I'm eternally grateful and look forward to passing those skills on to the next generation of philosophers.

I'd also like to thank the rest of my committee, who have been immensely supportive, letting me send them drafts, talking through ideas, and giving me incredible academic advice. I hope this dissertation does your support proud.

Chapter 2, in full, is a reprint of the material as it appears in Ethics, Policy & Environment 2020. Thresher, Ann. The dissertation author was the primary investigator and sole author of this paper.

Chapter 3, in full, has been submitted for publication of the material as it may appear in The Tangle of Science: Reliability Beyond Method, Rigour, and Objectivity, 2022, Cartwright, Nancy; Hardie, Jeremy; Montuschi, Eleonora; Soleiman, Matthew; Thresher,

Ann, Oxford University Press, 2022. The dissertation/thesis author was the primary investigator and author of this paper.

This dissertation wouldn't have been possible without the support of my incredible friends. Thanks for the grappa, good movies, bad TV, dinner parties, games, cocktails, and holidays. Particular thanks to Leo, for putting up with living with me for six ridiculously fun years. You're the best.

Thanks also to Mario, for some of the worst songs I've ever heard, paired with some of the best conversations.

And finally, thanks to my family. Dad, I'll take that 'Dr Thresher' plaque now.

VITA

2013	Honours in the History and Philosophy of Science, The University of Sydney, Sydney, Australia.
2012	Bachelor of Arts in Philosophy with Distinction, The University of Sydney, Sydney, Australia.
2012	Bachelor of Science in Physics with Distinction, The University of Sydney, Sydney, Australia.
2022	Doctor of Philosophy in Philosophy, The University of California San Diego, San Diego, USA

PUBLICATIONS

Ann C Thresher, "When Extinction is Warranted: Invasive Species, Suppression-Drives and the Worst-Case Scenario", *Ethics, Policy & Environment*, 2020.

Nancy Cartwright, Jeremy Hardie, Eleonora Montuschi, Matthew Soleiman, and Ann C Thresher, "The Tangle of Science: Reliability Beyond Method, Rigour, and Objectivity", *Oxford University Press*, Forthcoming

ABSTRACT OF THE DISSERTATION

Putting the Philosophy Back into Natural Philosophy: Ethics, Policy, and Using Philosophic Tools to Solve Scientific Problems

by

Ann C Thresher

in Philosophy

University of California San Diego, 2022

Professor Craig Callender, Chair

Here I discuss three different problems in science which can be advanced using philosophical tools. In each case I engage heavily with the science, drawing on the current state of play for the field, with the goal of providing helpful guidance to scientists and policy-makers.

The first chapter looks at the risks associated with a newly developing invasive species control method - gene-drives, arguing that we can warrant certain harms by weighing up the damages against the potential benefits we might see on a global scale. Here I use techniques and ideas from applied ethics and policy to show that certain highly-damaging species,

including European rabbits and ship rats, are viable targets for the technique despite the risk of their global extermination.

The second dives into political theory, intervening on a long and complicated debate surrounding the mechanisms behind the theory of the democratic peace. Here I use concepts like conceptual engineering, over-determination, and a newly proposed argument for the reliability of science called ‘the tangle’, developed by myself, Cartwright, Hardie, Montuschi and Soleiman, to argue that there is no single mechanism, but instead many which work together to secure the reliability of the democratic peace.

The third and final chapter points out a worrying trend of scientists and policy-makers thinking that the potential harms of science arise almost exclusively out of the application of research, rather than the research itself. Using tools from analytic philosophy I dive into the harms research can cause, proposing a new categorisation of these harms for use when creating moratoriums and bans. This categorisation system is then use on a real world case - research into the ‘gay gene’ - to show how it can help determine the most effective ways to mitigate harms via controlling what science can study.

Chapter 1

Introduction

Philosophers can and should be scientists, and scientists can and should be philosophers. The two do not exist independently, and attempting to separate them does harm to both projects. Indeed, it is only by acknowledging the philosophical questions embedded in science, and the need for philosophy to be grounded in scientific fact, that we can make progress on some of the most difficult questions facing both fields. How should we respond to climate change? Why trust astronomical models? What new technologies ought we to pursue? How do we avoid bias in our theories? Why does science work? What are the fundamental building blocks of the universe? Why does time have a direction? How should we divide up biological species, phylum, and families? What constitutes self-awareness? What types of data are admissible as evidence? These, and many more, make up some of the most difficult and philosophically interesting problems at the forefront of science, problems which can be advanced using philosophic techniques.

Unfortunately it is a serious failing of both modern science and modern philosophy that the two are seen as so separate by many practitioners. This failure, in turn, means that many of the useful tools of both have been disregarded by the other and many problems that can be solved, or at least advanced, have sat unresolved.

Given this, I am firmly of the belief that one of the most important things philosophers of science can do is roll up their sleeves, wade in, and attempt to give clear and scientifically informed practical advice to help resolve the conflicts that arise within, and out of, science. This thesis comprises three papers which do just that, looking at different areas of science that can benefit from science-oriented philosophic intervention. My goal in each case is to give pragmatic guidance to scientists, philosophers, and, in some cases, policy-makers, to try and clean up ongoing questions and problems within the field.

The first paper, **When Extinction is Warranted: Invasive Species, Suppression-Drives and the Worst-Case Scenario**, is narrow in scope, providing practical advice on how to use a new and potentially dangerous technology. In concert with scientists, sociologists, and philosophers at UC San Diego, I delve into the ethical questions of invasive species control, and new genetic techniques that could prove to be a silver bullet for the problem. Working with the specifics of gene-drive technology, I provide guidance on when and where we are warranted in using it, and under what conditions we are able to circumnavigate one of the serious risks of gene-drives - the accidental global extinction of a species. This paper draws on philosophy work surrounding risk, including work on dominance arguments and the precautionary principle, to help navigate the difficult question of when we are justified in applying scientific research which could either save the planet, or destroy it.

The second paper, **Conceptually Engineering a Theory of the Democratic Peace**, is aimed at solving an ongoing scientific debate - why don't democracies go to war?

It looks at the confusing tangle of theories surrounding one of the most successful guiding laws of political science - democratic peace theory and attempts to show how, using conceptual engineering and other tools provided by philosophy, we can make sense of a difficult problem and come to a clear solution. Here, I argue that what is seen as a number of competing theories is, in fact, a series of mechanisms working in tandem to secure a reliable scientific law. In doing so, I use work from philosophy on conceptual engineering, ballung concepts, generics, and overdetermination arguments, as well as a novel account of scientific reliability proposed by Cartwright, Hardie, Montuschi, Soleiman, and myself.

Finally, in the last paper in this thesis, **How Research Harms**, I use the tools of philosophy to categorise and analyse the harms of scientific research. Here I argue that there is a broad bias in science and policy where research is seen as either neutral or beneficial, and that to prevent science from causing harm we ought to focus on controlling it's application. That is, we allow for research into nuclear physics, but ban building nuclear bombs. This, I argue, is a dangerous sentiment, and one which means that we often fail to acknowledge or capture the harms that occur at the research stage. As part of this I categorise and clarify a classification scheme for research harms with the intent of providing both scientists and policy-makers with the language to help articulate these harms. For each harm I go in-depth with an example from science to help demonstrate what they look like *in situ*, concluding with a case study demonstrating how these categories can help us identify and implement appropriate moratoriums on harmful research.

Chapter 2

When Extinction is Warranted: Invasive Species, Suppression-Drives, and the Worst-Case Scenario

Abstract

Most current techniques to deal with invasive species are ineffective or have highly damaging side effects. To this end suppression drives based on clustered regularly interspaced short palindromic repeats (CRISPR/Cas9) have been touted as a potential silver bullet for the problem, allowing for a highly focused, humane and cost effective means of removing a target species from an environment. Suppression-drives come with serious risks, however, such that the precautionary principle seems to warrant us not deploying this technology. The focus of this paper is on one such risk – the danger of a suppression-drive escaping containment and wiping out the target species globally. Here, I argue that in most cases

this risk is significant enough to warrant not using a gene-drive. In some cases, however, we can bypass the precautionary principle by using an approach that hinges on what I term the 'Worst-Case Clause'. This clause, in turn, provides us with a litmus test that can be fruitfully used to determine what species are viable targets for suppression-drives in the wild. Using this metric in concert with other considerations, I suggest that only three species are currently possible viable targets – the European rabbit, ship rat and Caribbean Tree Frog.

2.1 Introduction

Invasive species are now considered the top factor in animal extinction rates worldwide, threatening to destabilise wide swathes of the ecosystem [22, 2]. Moreover, the crisis is accelerating [129, 60], largely due to globalisation and increased inter-bio-region trade [62, 63], and now poses significant threats to biodiversity, ecosystem functionality, human health, and the economy [129, 62, 116].

Given this, addressing the invasive species crisis is a high-priority goal for most governments and conservation groups. New Zealand, for example, has recently announced its 'predatory free 2050' campaign aimed at eliminating invasive mammals that are decimating the local environment.

Unfortunately, most current techniques are ineffective or have highly damaging side effects [51, 136, 8]. New tools, then, are a priority in a battle we are losing on a global scale [22, 2]. To this end genetics, and recombinant technologies in particular, have drawn considerable attention in the last few decades. Amongst these, gene-drives based on clustered regularly inter-spaced short palindromic repeats (CRISPR/Cas9) have been touted as a potential silver bullet for the problem, allowing for a highly-focused, humane, and cost effective means of removing a target species from an environment [144, 115]. Gene-drive technology comes with a number of risks, however, risks that currently make it the subject of intense international debate over whether to implement a global moratorium banning use of the technology for germ-line editing [3]. Appeals to something like the precautionary principle are common - arguing that we should not implement a technology which has an unknown potential for catastrophic effects despite the benefits it may promise.

Of these risks, three have come to dominate the dialogue - the possibility of unforeseen genetic changes due to the editing itself, the chance of the driven gene jumping to, spreading in, and adversely affecting another non-target species, and the fact that in the extreme gene-drives that aim to suppress a species in a specific range have the potential to significantly impair or even wipe out the target species globally. This last risk, which is a necessary implication of an effective suppression-drive, is extremely difficult to minimise - quarantine of bio-control techniques has a notoriously high fail rate. Indeed, as I will argue in section 3, given the release of a suppression-drive we should treat it as a certainty that it will escape to native ranges. The goal of this paper is to examine this third risk in depth, arguing that while we might think the precautionary principle prevents us from deploying technologies with the potential to wipe out entire species, there are some invasive species which seem to warrant the use of suppression-drives, even in worst case scenarios where the subsequent global extinction of the species is certain. The force of this argument rests on an application of the dominance argument, allowing us to sidestep concerns about unknown probabilities and factors by using a methodology which is robust against such factors. Section 5 explores this in some depth, arguing that by using a version of a dominance argument called the

Worst-Case Clause we can arrive at a litmus test by which we can make at least preliminary conclusions as which species seem to be viable targets for suppression-drive technologies.

2.2 Gene Editing, Suppression-Drives, and CRISPR/Cas9

Gene editing is a broader term for a technique of which gene drives are a subset. The term in general refers to any process by which DNA is inserted, deleted, modified, or replaced in a living organism, and can be performed in a number of ways. CRISPR/Cas9 has fast come to dominate the field, however, partly because it can be precisely targeted, and partly because it so cheap. Previous methods could cost upwards of $5000 per modification, while CRISPR/Cas9 can be as little as $30. The technology relies on an ability of bacteria that allows them to quickly modify themselves to resist viruses. Taking advantage of this, researchers introduce a Cas9 nuclease complex with a guiding RNA into a cell, and thereby are able to precisely slice genomes and insert new code [25].

While there are many potential applications for this technology, ranging from curing diseases to optimising crop yields, in this paper we'll be focusing on one particular implementation known as a suppression-drive. Gene drives in general utilise what are referred to as 'selfish' genes - genes that have an inheritance probability greater than the normal 50% that develops as a result of segregation of the chromosome pairs during meiosis [6]. The selfish genes discussed here achieve greater than 50% inheritance by copying themselves from one chromosome into its partner chromosome, changing an individual heterozygous for a gene to one that is homozygous. Because the selfish gene is now present on both members of a chromosome pair, all offspring, rather than only half, get a copy. The gene, in turn, copies itself onto the offspring's partner chromosome and so forth, resulting in the rapid spread of the selfish gene through a population [17]. This class of gene drives occurs naturally, but CRISPR/Cas9 techniques allow for the production of highly efficient 'synthetic' drives that can in theory prevent, reduce, or alter the reproduction of any individual which has it, thereby suppressing a species' population fecundity and viability. The most common ways in which this is done generally involve either altering the driven individuals to have only male offspring, rendering non-viable any embryo which inherits the gene from both parents, or rendering sterile any female which inherits the gene from both parents.

The efficacy of the technique depends on the target population and the context within which it is deployed [115], but all will, in an ideal case, cause the complete elimination within a chosen region of the species to which it is introduced.

Suppression-drives have been tested on a number of species, including mosquitoes [53] and rats [127], and current models predict that they can be highly effective at eliminating a target species. Trials in mosquitoes, for example, showed complete spread and subsequent population crash within 11 generations [73].

2.3 Suppression-Drives: Benefits, Risks, and Worries

The potential benefits of suppression-drives for controlling invasives are significant. Numerous studies have shown that the options currently on the table are insufficient for the task of dealing with the current crisis [51, 136, 8]. Examples include restrictions on the efficacy of hunting and trapping in fully eliminating most species [83, 124, 131, 94], extensive environmental damage from poisons like 1080 [65, 36], the evolution of resistances towards targeted diseases like myxomatosis [45, 12, 123, 70], inhumane suffering from both poisons and diseases [29, 109], and practical limitations on sterile male release [126, 35, 98], and Mendelian inheritance techniques [56].

In contrast, suppression-drives are often more effective, cost less, do less environmental damage, and are more humane. One benefit that is often cited is that these drives have minimal impact on the lives of the individuals in question - animals who inherit a suppression-drive live long, full lives. As such, they cause much less pain and suffering than many of the methods above. Similarly, they are kinder to the surrounding environment - there is no risk that native species will be affected, nor local waterways, pets, or humans. Suppression-drives are also often far better at reducing the numbers of a target population, especially over broader areas. The release of a drive in one small area can lead to the removal of a species as a whole from a large region. By exploiting natural migration and interaction patterns, the drive can reach areas more traditional techniques cannot, and because it doesn't kill the host like diseases, it has a much more effective transmission rate. Finally, they require much less ongoing maintenance than sterile male release or Mendelian approaches, and are easier to implement in cases where the continual mass-release of individuals is difficult or would prove devastating to the environment [56].

Downsides obviously include the fact that the drive is, in some cases, much slower than other methods - poisons and diseases are much faster at reducing a population for example - but this is counterbalanced by its theoretical effectiveness, ideally acting to eliminate rather than suppress in the long-term.

This technology comes with a number of commonly cited worries and risks, however. We can divide these into two broad categories - in-principle concerns, and risk-based concerns. In-principle concerns will include things like arguments from the inherent right or intrinsic value of the species, arguments from playing god, arguments from arrogance, and arguments from mistaken technological reliance. Risk-based concerns often brought up include arguments from ignorance, arguments from species jump via horizontal gene transfer or hybridisation, and arguments from escape. All of these play some factor in current considerations over whether we have the right to release suppression-drives into the wild.

In this paper I'm going to focus on the last of these - the argument from escape and possible subsequent global extinction for the target species. This is partially because many of the other worries have been either explored elsewhere in some depth[1] or, as with many of the pragmatic risks, because they are an acknowledged problem that will necessarily need to be eliminated before the technology could ever get anywhere close to release. Conversely, the escape of a gene-drive to non-target regions is extremely likely, no matter how much effort is put into quarantine and, given this escape, we are likely to see significant suppression or even extinction of the target species in their home-ranges - a risk that has not yet been discussed in any depth in the ethics literature.

[1] For a good discussion of playing god see Evans (2002) [43]. Intrinsic value we will return to later, but Rolston III [122] is the only positive argument I can find in favour of this position. For technological reliance both White 1967 [148], and Scott 2011 [128]provide nice overviews of how technological advances interact with environmental concerns. For risk-based concerns, work by Roberts [120], Moran [91], and Zhang [154, 153] talk about why problems like off-target effects, and horizontal gene transfer are, even today, relatively minimal concerns for the technology.

2.4 Escaping Genes

In many ways this worry is what distinguishes the invasives case from other gene drive applications. Consider, for example, the implementation of a drive into mosquito populations that prevents them from transmitting malaria. If this genetic modification left the region it was released in, such that all individuals of the target mosquito species world-wide eventually possessed this gene, we're unlikely to worry about the eventual results. Similarly, if we edit the human gene-line to make people smarter, or healthier, the spread of this gene would (barring those cases where we may worry about homogeneity of genetics, or the role social norms have in determining what is 'good' for a human genetically) likely not be cause for alarm. What worries we do think hold in these cases are often simply expanded versions of the arguments from ignorance or species jump - where the larger population might make the possibility of off-target effects or jumping more likely. Similarly, we might have some expanded worry from the argument from natural states where we think there is some loss that comes from modifying *an entire* species, rather than a subset of it, and thus preserving the original.

The use of population suppression-drives, however, changes the stakes. Here, the biggest problem is not the possibility of altering an entire species, but of destroying it completely - a worry that is embedded in the optimisation of the technology itself rather than unpredictable effects from a faulty or off-target drive. Given these stakes, serious consideration needs to be given to the question of global species wipe, or in the less extreme case, a global crash in the target species that significantly reduces its numbers in its native range.

Let's look at this option in more care, then. What are the actual risks involved, how likely is it that we can mitigate them, and how should we weigh them up against the moral obligations we have to solve the invasive species crisis? The answers to these questions should guide our attitudes towards the release of suppression-drives going forwards.

2.4.1 The Chance of an Escaped Population Suppression-Drive

Suppression-drives have two fundamentally competing requirements. Firstly, they need to be able to drive effectively and efficiently through the target population. Conversely, we also need them to be containable, such that they don't spread to non-target regions. Isolation is often the easiest way to resolve this problem - one reason that most currently proposed applications of gene-drives are on islands.

There are, however, good reasons to think that population-suppressing gene drives will escape even the best quarantine. Primarily amongst these is simply the fact that the species reached the area in the first place. This, in itself, suggests an at least historic transferal of individuals between native and non-native regions. While there are some cases where species were deliberately moved from one area to another (take the eucalyptus in California, for example, or the cane toad in Australia), many transferals were accidental - lanternflies in plane cargo, zebra mussels on the bottom of ships, cats escaping captivity for a life in the wild etc. These mechanisms are, in the majority of cases, still in place, and are often not one-way. Ships still routinely move rats from port to port, people emigrate with pets, warmer weather extends the continuous range of insects and plants. It is difficult, if not impossible, to prevent this movement over long time-spans, as people battling the initial invasions of invasive species have found [82].

This is particularly true of those species which pose the greatest global threats to biodiversity. Ship rats are so named precisely because of their predilection for stowing away on transports, and are now found on over 80% of the world's islands [138]. Restricting their movement is notoriously difficult [124, 82], and while quarantine can be put in place the chance of escape, even for highly immobile species, is non-zero, escalating quickly when we consider those species for which suppression-drives seem like they can do the most good - namely those which are widespread, where traditional methods are ineffective, and where they have significant reintroduction rates. Sparrows, mice, carp, rabbits, mosquitoes, and other major invasives are the target of reduction efforts often precisely because they are difficult to control and are highly mobile, while threatening a large number of native species. What this means is that, despite our best efforts, it is likely that the release of a drive in one region will mean the spread of it to other adjacent areas. This spread, in turn, increases the further chance of escape, creating a domino effect where any initial quarantine failure significantly increases the chance the drive will eventually effect the target species in its native range. Consider, for example, the use of a suppression-drive on rats in New Zealand. The chance of the drive jumping across to mainland Australia is not inconsequential, for all the reasons we've just discussed. Once in Australia, however, this significantly increases the chance of it migrating upwards into Asia, or jumping to another continent, given the much more extensive trade ties Australia has with other nations. Once it's on more than one continent, it will be almost impossible to prevent the spread of the drive towards Europe, where the rat is native.

There are, of course, greater and lesser risk species. Carp in Australia are unlikely to reach Europe accidentally via human trade routes, nor are cane toads or foxes. Conversely, we don't always need human movement for a suppression-drive to escape. Lampreys in the Lauritian Great Lakes are a good example of an invasive species that has the potential to naturally migrate back to a native range - while lampreys from the Atlantic are unable to move upstream to reach the lakes, once the fish are upstream there's a non-zero chance that successive generations might migrate downstream and reach the ocean, spreading a suppression-drive with them [137].

On top of this, we have the natural tendency of humans to do *stupid things*™, which includes multiple instances where people have deliberately spread species and bio-controls to places they will do damage, either through ignorance or malice. While it's unlikely that a suppression-drived carp in Australia will accidentally hitch a ride on a plane to Europe, this doesn't prevent someone from simply capturing a specimen and deliberately moving it across continents. The history of human expansion is marked by instances just like this [59]. Consider, for example, the release of non-native species on colonised landmasses by the British for hunting, agriculture or, in one memorable case, as a tribute to Shakespeare [155]. In the modern day, the introduction of invasive species is often not malicious or with any goal in mind beyond planting a particularly pretty type of exotic flower in one's garden, or keeping an unusual pet, with the mistaken belief that these actions won't cause harm in this one specific case. On the subject of bio-control efforts, we also have cases like French researcher Paul-Félix Armand-Delille catching rabbits deliberately infected with myxomatosis in order to infect rabbits on his estate in Eure-et-Loir to lessen the damage they were causing to his grounds. Less than a year later the rabbit population in France and Iberia had dropped by 45%, with rabbit hunters reporting a 98% drop in yields in the subsequent season [33].

Similarly, farmers deliberately and illegally released the haemorrhagic disease virus to New Zealand in 1997 [103]. Humans, as history reminds us, are often the biggest danger to containment efforts [59, 42].

With the right quarantine in place, and by selecting the right species, region, and suppression-drive type however, the chance of escape can be extremely low [19, 69, 82], but critically, is never non-zero. In general, then, unlike with the horizontal gene transfer or mutation cases, we cannot dismiss this as an extremely remote possibility, or one that can be solved before the release of the drive. By one means or another for many of the invasive species it is most important we address, the release of a drive will likely mean the suppression of the species in its native range, and at the most extreme, complete extinction.

2.4.2 The Chance of Global Extinction given Escape

Given that we can never guarantee the isolation of a suppression-drive, then, what are the chances that an escaped drive actually wipes out the target species, rather than simply temporarily suppressing it?

The actual numbers on this, as with the actual risk of escape, are hard to calculate, and seem to strongly depend on context. Several models have, however, been put forward [37, 9]. North et al. (2013) [101], for example, model the propagation of Homing endonuclease genes (HEG) through mosquito populations that are in stable or semi-stable equilibrium. These models are also applicable to Crispr/Cas9 drives, and show that extinction occurs only when population growth is low with sparse density. When breeding and feeding sites are common, the release of a drive results in suppression instead. They also note that the success of these drives depends strongly on the drive being slow enough acting that it has time to spread to the entire target population before wiping out the initially impacted groups. Too fast, and it eliminates itself before it can propagate fully.

There is also a reasonable amount of evidence to show that even stronger drives might not be as effective as we first thought. There is strong evolutionary pressure acting against members of a species that are less viable. A gene-driven individual might, for example, have half their offspring be unable to reproduce. This makes them less fit than a competing non-suppressed member of the same species. Current models suggest that even a fitness viability loss of 25% could be enough to counteract the 'drive' aspect of the suppression-drive, preventing its spread through the population [34]. There is also significant evidence that this viability mismatch can cause target species to evolve resistances to the drive [54, 18]. Indeed, in the models given by Unkless et al. (2017) [139] such resistances are almost inevitable given enough time, although it's possible to reduce the chances of them cropping up given the right techniques.

Note, however, that those factors which make extinction unlikely are also the aspects of suppression-drives we are actively trying to 'solve'. Suppression-drives rely on their ability to quickly and effectively spread through a population - there is little point in releasing one to which a species will develop immunity, or one which cannot effectively move between regions containing the target invasives. The Kyrou et al. experiments [73] which eliminated a lab mosquito population in eleven generations was celebrated precisely because it managed to solve the problem of evolved immunity. As such, any version of a suppression-drive which is eventually released will likely have done almost all it can to maximise it's chances of becoming an extinction-drive.

Scientists are, however, not blind to the risks involved in a suppression-drive [42, 77, 121, 10]. Much effort has been put into mitigating escape worries [2], primarily through the use of artificially limited lifespans for the genes. Daisy-chain drive technology is a paradigmatic case of this, although recent work has suggested it may be less viable than previously thought [88, 100]. These limitations often involve a time-delay included with the drive that kills it after a certain number of generations, preventing the kind of global spread that we're worried about here. Tied into this are also calls for a drive to not be released until a counter-drive is also ready to go [42]. The idea here being that in the case where the drive turns out to have deleterious effects, or escapes the bounds of the target region, a second drive to remove the gene can be released. This drive, in turn, is obviously subject to all the same worries as the original drive in terms of mutation and transfer, but at least theoretically acts as a safeguard against accidentally wiping out a whole species.

So then despite what look like good fears about our ability to contain a drive in the wild, the actual current chance of global extinction seem low. Significant suppression, however, is a very likely effect of the release of a drive, particularly for those wide-spread species which do the most damage and are the hardest to contain. Suppression, however, can be almost as bad as extinction - just look at all the species on the endangered list whose reduction in numbers are seriously impacting their native habitats. As such, while we should likely not worry too much about the possibility of a suppression-drive causing global wipe-out, this doesn't mean that the risks involved don't have serious stakes.

[2] It is worth noting here that in some cases the global elimination of the species is actually the goal of the project. One approach that has been floated for dealing with the malaria crisis has been to wipe out malaria-carrying mosquito species using a suppression-drive, in which case scientists are obviously not trying to mitigate the chance of escape [53].

2.5 Weighing up the Risks

Unlike other methods of invasives control, gene drives have potentially global implications for a species. We cannot accidentally poison all rabbits on earth, nor would the release of a virus likely prove fatal in the long-run, even if every country got infected simultaneously. Suppression-drives, however, no matter how careful we are will always have a non-zero chance of escape and total species wipe - so how should we think about such scenarios? This is where the precautionary principle is often evoked - given the uncertainties involved, and the potentially serious repercussions, the consensus seems to be that we should hold off on using suppression-drives in any wild-release context.

Still, we can attempt to make some primitive calculations, treating it as a balancing act: If the benefit of act A is n and the act will do damage worth m, and there is a $x\%$ chance of the benefit, and a $y\%$ chance of the detriment, then we might simply assert that we should do A iff $xn - ym > 0$. That is, if the benefits outweigh the detriments.

Such calculations are difficult in real-life, however. Here's one attempt at one - ship rats in New Zealand are responsible for the extinction of 23 local species, and they threaten about 40 more [138]. If we were to release a suppression-drive on invasive rats in New Zealand, then we are potentially saving many of these species, and potentially eliminating one (the ship rat). This, in the grand scheme of things, seems like an overall win. Even in the case where there is a 100% chance of the drive escaping and killing all ship rats, we have still made a positive contribution to saving species diversity because we've killed one to save many.

This is, obviously, an incredibly naïve position to take. There are a number of salient factors in any such calculation beyond simply brute species diversity. Rats, for example, might be far more numerous globally than the species they are replacing, even before they were invasive. As such, we might think that the sheer number of ship rats gives some weight to the idea that they are more worthy of saving than the native species, as the collective right to continued existence of the first outweighs the additive collective rights of the others. Alternatively, rats may be a keystone species in their native range, as with bees in North America, such that eliminating them would disrupt the entire ecosystem, leading to further extinctions and environmental degradation. On the other hand, it might be the case that rats are one of a number of species that fill a similar ecological niche in their native environment, and thus their elimination wouldn't seriously disrupt the ecosystem, as with the northern and southern white rhinos which are technically genetically distinct but in every salient way are identical from an environmental perspective.

We also know that the chance of wiping out rats globally is far from 100% - partially because there will be at least some attempt at quarantine, and partially because unlike poisoning or viruses, the speed of suppression-drive spread is limited by generational considerations. We will have numerous generations of a species to deal with an escaped drive, either by quarantining affected areas, or by introducing reverse drives into the native range being affected. All these factors weigh in favour of suppression-drive release, even in the face of global species wipe risk.

All this is to state the relatively uninteresting conclusion that the release of suppression-drives is incredibly situation dependant, and that many factors will need to be accounted for in each case where it is considered for use. While there will be times when suppression-drives prove to be morally problematic, as with cases where simple trapping or hunting will solve the problem and thus we need not run the risks that CRISPR/Cas9 technology opens up, in other scenarios we might seem more justified in using the suppression-drive approach. These calculations, and the debate around them, form a large part of why suppression-drives are so controversial - no one is entirely sure what constitutes a right to release such a technology, nor what types of risk analysis we should be doing to answer the question.

2.5.1 The Worst Case Scenario

It's very easy to get caught up in the details of this complex issue. When do we have the right to release a technology that could wipe out an entire species? What factors should we consider, and what relative weights should we give them? How do we ever analyse the risk of accidental extinction by our own actions? The uncertainty and unknowns make the problem all the more difficult - and is one reason we often take the path of precaution, determined to not take any actions until we have a clearer picture of what the possible problems could be given the severity of getting it wrong.

While all of this debate is important, and will indeed be critical in our eventual decisions over whether to release suppression-drives into the wild, I don't think that this uncertainty completely precludes the release of suppression-drives in certain specific cases. In fact, I suspect that even without knowing the full impact of what a suppression-drive might do, we can still motivate a case for release on some species that meet certain criteria.

To justify this I will be appealing to what I will refer to as The Worst-Case Clause (WCC),

> **Worst-Case Clause** - If we are warranted in taking action x even after assuming that the worst possible outcome will follow, then we are warranted in taking action x in general.

That is, if our reason for precaution is that we don't know how risky our action could be, and lack enough information to make an informed decision, then one way to bypass this is to simply assume the worst from the beginning. If, in this context, we still seem warranted in taking the action, then we no longer seem to need to worry about the precautionary principle.

In the context of suppression-drives that could, with some unknown probability, eliminate a species, this means assuming that any suppression-drive released is guaranteed to escape quarantine and wipe out the target species globally. For the sake of clarity, we'll refer to suppression-drives with this added worst-case scenario caveat as 'extinction-drives'.

The Worst-Case Clause is a type of dominance argument. In this we're utilising a methodology which is robust to changes in the underlying probabilities or uncertainties. Instead of entering into these types of calculations we're instead claiming that for certain species the worst-case scenario of extinction is still such that it is strictly the best outcome for the system as a whole.

Let's look at one example that I think meets the Worst-Case Clause.

European rabbits are one of the most pervasive invasive species in the world[3]. In New Zealand, they have on occasion reached plague-like numbers since their introduction in the early 19th century, ravaging crops, overgrazing native flora, and out-competing native species for resources thereby reducing native biodiversity [76]. For this reason they are included in the New Zealand 2050 'mammal free' goal. Current control methods include hunting, trapping, poisons (via bait, aeroplane drop, and being pumped directly into bur-

[3] Its been suggested that European rabbits are the first recorded invasive species, with Pliny the Elder noting in the *Naturalis Historia* (VIII.80) that they were invasive to the Balearic Islands around 75AD where they caused famines, brought down trees, and collapsed houses, leading to Divine Augustus sending a troop of soldiers armed with ferrets to help the region.

rows), the building of rabbit-proof fences, the introduction of competitive and predatory species (including the now invasive ferrets, stoats, and weasels), and targeted disease (Myxomatosis and Rabbit Haemorrhagic Disease (RHD), the former failing to become established, and the latter being introduced by private individuals against the government's wishes. Recent studies have shown that New Zealand rabbits have begun developing a resistance to RHD) [28]. None of these methods have had any long-term impact on rabbit populations in New Zealand.

Similar stories may be told throughout the world. In Australia, rabbits have had massively detrimental impacts on agriculture, native flora and native fauna [24, 27]. Africa and North America have similar problems [113, 64] - indeed, European Rabbits are now invasive on ever continent except Antarctica, including over 800 islands worldwide [76].

Given the failures of current methodology, very few options are left on the table for rabbit control. The most promising of these is, unsurprisingly given the contents of this paper, CRISPR/Cas9 gene-editing technology. This method obviously comes with all the risks we've already discussed. There is, in as much as there ever is, the remote chance of horizontal gene transfer via bacteria, although this has a very remote chance of serious damages via random insertion [154, 142]. However, there are no recorded instances where European Rabbits have hybridised successfully with native species outside of their own genus in a way that creates viable offspring [20, 112] making them a good candidate species for a suppression-drive along this dimension. Spread does also means all the worries about 'loose' genes crop up though, where the perpetual existence of the drive may, in remote cases, lead to unexpected off-target effects. Finally, European rabbits are a keystone species in their native range in the south of France and Iberian Peninsula, and should the drive escape and spread world-wide their destruction has potentially devastating implications for these environments.

Imagine, now, that in attempting to remove European Rabbits from New Zealand using CRISPR/Cas9 the worst happens and it is wiped out globally. What then? This is obviously a difficult question to answer - European rabbits have been invasive around the world for centuries and it's hard to predict what removing them might do even in ranges where they're non-native but might now be providing some stop-gap role. As above, there are primitive calculations we can make. Take, for example, the fact that although rabbits are not native to the UK (having been introduced in the 12th and 13th centuries through trade), many studies have proven that they now provide critical services maintaining the natural landscape. So much so that an outbreak of myxomatosis in Southern England led to a radical restructuring of local heathlands, and the subsequent decline of the native Maculinea arion butterfly, Myrmica sabuleti ant, stone curlew, red-billed chough and woodlark [76]. In cases like the UK it is likely that the European rabbit has taken up the role of a now extinct mega-fauna, maintaining environmental stability in a highly-farmed and artificial part of the world. As such, its removal would have serious consequences for the region, despite the rabbit being a non-native transplant.

The rabbit itself is also endangered in parts of its own native range - the Iberian peninsula - giving us some further insight into what might happen if they were removed completely. In the Iberian peninsula the species has seen significant decline, threatening the further survival of other endangered species, including the Iberian Sphinx and the Spanish Imperial Eagle [76]. Should the European Rabbit become globally extinct, then, there would be significant knock-on effects for the ecosystem of its home region, likely leading to extinctions.

The actual impact is difficult to calculate. In accordance with the worst-case clause, however, let's assume the worst. If the rabbit goes extinct let's assume it will take down numerous species with it - contributing to the collapse of diversity on the Iberian peninsula and places like the heathlands in the United Kingdom. This is, of course, only to speak of biological disaster - European rabbits are also central to a number of economic systems in various regions, where hunting and farming the animal is a key source of income.

This worst-case scenario is pretty bad - we are effectively collapsing multiple ecosystems, robbing large groups of people of critical income, and driving multiple species to extinction. It could be argued, however, that this was a mild version of a worst-case scenario - if we really wanted to push things to the extreme it might be possible to argue that the elimination of the European rabbit could have unforeseen devastating consequences globally, perhaps even preceding the total collapse of the global biosphere. Or perhaps the European rabbit turns out to be the key to curing the next deadly virus to sweep mankind, or that it somehow has been preventing the complete economic collapse of multiple countries. How, I am unsure, but technically we cannot rule out the possibility if we really are aiming to consider the most devastating version of an extinction-drive. Should we take one of these, then, to be our worst-case scenario, rather than the less extreme one I've proposed here?

Given that I am attempting to distance myself from probabilistic arguments here, it is frustrating to have to turn to one now, but I think we are warranted here in pointing out that the chances that removing the rabbit will be more devastating than the first worst-case scenario are remote. In the absence of any reason beyond wild imagination to think that the European rabbit has become a keystone species in most of its non-native ranges, or that there is some reason to think it is propping up international economies, or that if we attempt to kill of rabbits they'll retaliate by stealing all the nuclear codes and causing thermonuclear devastation, we should act to construct a worst-case scenario based on what we could feasibly see happening at the extreme.

In contrast, consider the damages currently being caused by the rabbit in other environments. Whatever damages may occur by removing the rabbit in Iberia and the UK is already in effect in New Zealand, Australia, South America, and many other ecosystems where rabbits are causing the extinction of numerous native species. Played as a pure numbers game, global biodiversity and stability improves both if the drive works as it's supposed to and only removes rabbits from New Zealand, and if quarantine fails and we see a worst-case scenario of global wipe-out. Moreover, what if one of *these* species are the one that could save mankind from virus, or stabilise global economies? Indeed, the rabbit is driving to extinction many more species than we could reasonably see being threatened by it if it were to disappear, meaning that if we are aiming for a worst-case based on pure speculation we still need to give more weight to the side of rabbit elimination rather than preservation. Given this, we can conclude that the potential damages caused by removing the European rabbit globally are much less than the current damages already being caused where the species is invasive.

In many ways this argument is simply a numbers game - weighing up the benefits of removing an overwhelmingly damaging species against the problems removing it may cause. In the European Rabbit case, this seems to come out in favour of removal - the species seriously threatens agriculture, and native flora and fauna in almost all it's extensive invasive ranges, and its loss, however serious, would still in turn damage only a limited ecosystem and set of economies. Thus, we seem justified in thinking that even in a worst case scenario we are warranted in releasing a suppression-drive. The European Rabbit is, however, a very distinct case. It has very little evidence of hybridisation, it has a very limited natural range, and it is on the extreme end of invasive species in terms of spread and damage. If, on the other hand, rabbits were only invasive in New Zealand, and not anywhere else in the world, and the damage they did there was on any scale comparable to the damage that would be done by their extinction in Iberia, then even if it were the case that they were driving species to extinction, and that there was no way to remove them other than a suppression-drive, we seem much less warranted in releasing a drive.

2.5.2 The Right to Cause Extinctions

Here is one possible objection to the argument I have outlined above - throughout this paper I have assumed that the value of species derives from the value of biodiversity, ecological preservation and stability, human utility or some other similar principle. Thus, I have biased the calculations in my favour since invasive species almost by definition are those which threaten biodiversity such that any non-extreme action to remove them will count as a positive mark on the scales towards action.

Here is a different take on the argument. Instead of valuing biodiversity etc (either inherently or for its utility) we might think that either the species itself is inherently valuable, and thus has some primary right to exist, or that we have some duty of care towards species such that we do not have the right to eliminate them. In this way, we might wonder not whether we can balance the scales in favour of saving more species, but whether we ever have the right in the first place to act in a way that will knowingly eliminate it. Perhaps there is some more fundamental principle in play which forbids such actions on the grounds of our inherent duty of care towards these species, particularly given that it is our own interventions which have spread these species to the point where we are now considering taking an action which could wipe them out.

While I suspect that most arguments about our duty of care in this case don't go through, thus rendering this approach nonviable, I do think that even if we accept this premise we can still make the case for worst-case suppression-drives being deployed.

Let's break this objection down a little. Firstly, we have something that looks like the following principle,

> **Duty of Care Principle (DCP):** We have a duty of care towards any given species such that we should allow the species to survive.

This, obviously, is not enough to get the objection up and running. After all, if it is the preservation of species that is important then any scenario where an invasive species threatens to drive more than one species to extinction would license us to release the extinction-drive.

I take it, then, that we must add something like the following,

> **Active Participation Principle (APP):** If a species should be allowed to survive, then you should not act such that you knowingly cause the species to go extinct.

The wording of this is to distinguish a difference between actively killing a species and letting one die by omission. I take this to be a core aspect of the objection - that although we might be doing something wrong by not stepping in to save the birds of New Zealand from introduced predatory mammals, it is worse to actively try to eliminate the weasel in order to do so.

I cannot say I find this argument very convincing, but it will be useful to briefly discuss why.

Firstly, ‘letting die’ implies that we have not had a causal hand in the death of the species, which is patently untrue in most invasives case. While we may not have intended the effects, we as a species are still responsible for the introduction of the invasive pests, and thus the ones who set the extinction mechanisms in place. As such, the distinction between the ‘killing’ and ‘letting die’ sides are less straight-forward than in other paradigmatic cases of the action/omission distinction. To lean on an old trope, it is akin to being faced with a trolley problem where you were the person who accidentally set the trolley loose in the first place. In this case, it is unclear that refusing to pull the lever absolves you of the moral responsibility for the subsequent deaths. As such, the claim that the APP tells against extinction drives seems far weaker than the initial formulation suggests.

Direct action is, of course, not the only way we might bear responsibility. Perhaps it was not you who released the trolley (or transported invasive species), but instead you are receiving some benefit that comes from the existence of runaway trolleys as a whole (as we benefit from globalisation and the continued transport of goods which acts to move these species). Similarly, it seems difficult to claim that we should not pull the lever, as the distinction between action and omission is blurred.

There is, however, still a distinction to be made between accidentally killing someone by setting a process in action, and deliberately pulling the trigger. As such, we might think that even though we are still responsible *in some way* for the deaths of various endangered species, this is still morally preferable to setting out to deliberately wipe-out a species. Note, however, that even if we make the assumption above that releasing a suppression-drive will necessarily eliminate a species, this is not actually the aim of most invasive species controls. In the context of biodiversity conservation it is never the case that we are aiming at the complete global elimination of a species - only ever the removal of them from non-native ranges. Global wipe is, at worst, a foreseeable but unintended repercussion of the act.

Philosophy tends to draw a distinction between an act which aims at doing something, and an act which has the same unintended, but foreseen, consequences. Most formulations of the doctrine of the double effect have four conditions which must be met for an act to be a morally acceptable [85],

The Doctrine of the Double Effect (DDE):

1. The action in itself must be good or indifferent.
2. The good effect must be intended, the bad effect must not be intended.
3. The good effect must cause the bad effect. The bad effect cannot be used to achieve the good effect.
4. The good and bad effects must be proportional, such that the good sufficiently outweighs the bad.

The extinction-drive case meets all of these. In itself, the release of a gene-drive is at worst indifferent; we intend the good effect of saving the ecosystem and various species therein and by definition do not intend to eliminate the target species globally; the good of saving the ecosystem is the cause of the bad, rather than the other way around; and as we've just discussed in section 2.5.1 the good outweighs the bad in such a way that it is certainly a proportional action.

In the case of an in-principle reason to avoid releasing an extinction-drive then, we can see that

i. the choice between passively allowing a species to go extinct and actively eliminating it is difficult to maintain in the invasive species case due to our broader obligations towards the environment, as well as our either passive or direct hand in causing and benefiting from the problem,

ii. in the case where we actively release an extinction-drive, the global wipe-out is an unintended but foreseen consequence of our bid to save the environment, something that meets the requirements for the doctrine of the double effect and,

iii. this argument holds regardless of whether we take the value of the species, ecosystem, or individual animals to be inherent or instrumental.

This does not, of course, preclude us from being morally responsible for the ensuing results - just as we as a species have a moral obligation to deal with invasives, we will also have a moral duty to try and prevent the elimination of a species we have introduced an extinction-drive to, even if we have decided that the cost of the species loss is worth the payoff.

2.6 Conclusion

We often appeal to the precautionary principle when we are faced with a new technology that poses a poorly understood but potentially serious risk. Such is the case for suppression-drives which run the risk of accidentally wiping a species out globally, and is one reason that international moratorium have been extensively discussed by both scientists and policy makers. One way around these uncertainties is appeal to a dominance argument, assume the worst, and take it that any suppression-drive will become an extinction one, stepping away from the complex risk-analysis questions that underpin most debates on when to use suppression-drives. If it is the case that we can warrant the global extinction of the target species in the name of environmental stability, then it is the case that we can warrant the use of the (likely) less extreme suppression-drive, even in cases where this suppression-drive does serious damage to the population of the species world-wide.

The worst-case clause is a useful litmus test for what species we are warranted in focusing our attentions on. It is only one of a number of factors that I take to be relevant to such calculations, however, including but not limited to,

1. **Worst-Case Clause:** It must be the case that if the invasive species went extinct globally we would not cause disproportional damage to the environment, either in the species' native ranges or in ranges where it is invasive but now plays a key part of the ecosystem.

2. **No-Alternatives:** There must be no other viable/cost effective alternatives on the table to deal with the invasive species.

3. **Reasonable Timeline:** The time between subsequent generations of the species should be sufficiently short to allow the drive to work within a reasonable timeframe.

4. **Proportionality:** The damage caused by removing the species must be significantly lower than the damages prevented - if a keystone species with a large native range is invasive on a single island we aren't warranted in releasing the drive from a consequentialist position, nor do we meet the proportionality requirement for the doctrine of the double effect.

5. **Ark-viable:** If we assume, as we have here, that a suppression-drive will necessarily

be an extinction-drive then the creation of reserves for the species becomes a necessity before the release of a suppression-drive. As such it must be possible to create and maintain an effective ark for the species by which un-modified genetic diversity can be preserved. While this is possible with most terrestrial species, certain aquatic species will be difficult to maintain in isolation in great enough numbers to rebuild the species should the suppression-drive become an extinction one. I expect the onus for creating such an ark will fall primarily on the releasing party.

6. **Non-Value-ladened:** The target invasive species should not be one to which we attach other types of value that might outweigh environmental damage. This may include criteria like sentimental or cultural value (eg. cats and dogs), research value or uniqueness (eg. lab rats and coelacanths), or being possessed of higher-intelligence (eg. apes and dolphins).

7. **Non-hybridising:** It must not be the case that it is able to produce fertile hybrids with non-target species unless that species also meets all the other criteria on this list both alone, and in combination with the target species.

This is, obviously, not a complete list, nor should it be consider the case that species which meet all of these are thereby automatically an acceptable target for a suppression-drive. I do think, however, that meeting all these criteria, particularly Worst-Case, does warrant us boosting the species to the top of our potential lists, and sidestepping several of the worries that bring the precautionary principle into play.

What species meet these criteria? Of the International Union for the Conservation of Nature (ICUN)'s list of top 100 invasive species [81], if we set aside diseases, plants, and insects as being not well enough understood yet at an ecosystem-wide level, a cursory survey shows that only a few seem like they might currently be possible candidates for a suppression-drive including the Caribbean tree frog, ship rat, and European rabbit. This is, of course, a superficial glance at best, and many of the criteria above are highly circumstantial and subjective judgements, but it is worth noting how rare it will be for a species to meet the minimal criteria I have listed above.

In all of the above I've framed this argument in terms of extinction. As I pointed out earlier, however, the chances of a suppression-drive actually wiping out a species globally is vanishingly low in reality. Realistically a worst-case scenario looks not like extinction, but significant suppression. This, however, can be almost as bad - the practical difference between a species disappearing and simply being massively reduced are minimal. In both cases the species is failing to perform the role needed in the ecosystem, and while in the latter there is hope for recovery, this doesn't mitigate the biome collapse that will occur beforehand. Either way the arguments I've made above can be applied, and seem give the same answers.

In conclusion, then, while there are cases where we might think that the use of suppression-drives involves unwarranted risk, in a small number of scenarios they look to be either the best or the only viable option available to us to stop the massive loss of biodiversity on a global scale caused by invasive species. I have focused primarily on ship rats and European rabbits here, both of which I take to be species which pass the Worst-Case clause and thus allow us to bypass the precautionary principle for the extinction-risk aspect of suppression-drives. In both of these cases the species involved are too pervasive for conventional eradication techniques to be effective, and the current damage being done to the environment seems to tip the scales in favour of CRISPR/Cas9 suppression-drive use despite the remote possibility of species extinction.

This paper was published in Ethics, Policy & Environment on the 25th of November, 2020. DOI: 10.1080/21550085.2020.1848197

Chapter 3

The Tangled Principle of the Democratic Peace

Abstract

‘Democracies don’t go to war with other democracies’. This is the democratic peace (DP) principle, something which is widely agreed upon, but which remains the topic of heavy debate. There is, as of now, no single agreed-upon explanation for why it holds. Here I propose a new solution to the problem of why the democratic peace holds, drawing on work done in my forthcoming book the Tangle of Science to help conceptually engineer a new definition of the term. The DP is, I argue, an example of competitive and convergent constraints on theory, wherein the Ballung nature of the terms “democracy” and “war” force theories apart, while simultaneously the existence of standard touch-stones for the field, as well as inter-theory critique, prevent theories from moving too far away from one another. I then argue that we ought to read the DP as a principle which holds that “democracies

by anybody's book" will not engage with each other in ways that will count as "war by anybody's book". In proposing this, I suggest that the DP should be read as a principle defined only over core cases: the DP-principle-by-anybody's-book. This argument draws on the role of generics and ceteris paribus principles in science to create an overdetermination argument, where the DP holds not because there is one correct mechanism which prevents war, but because there are numerous overlapping ones acting at once. We can expect it to be rare, then, for all these mechanisms to fail in a given scenario, leading to the impressive reliability of the DP principle.

Here I include, in its entirety, a chapter from a forthcoming book written by Nancy Cartwright, Jeremy Hardie, Eleonora Montuschi, Matthew Soleiman, and myself. I present it as it will be published with two exceptions; first a foreword which gives some context of the contents of the rest of the book to help the reader understand the work the chapter is doing. Second, a number of comments and notes throughout the chapter where context or relevant information is needed to understand what is being said. These comments will be almost exclusively footnotes, and ***will be distinguished by being bold and italicised*** in contrast to footnotes which were included in the book.

3.1 The Tangle of Science

The forthcoming Oxford University Press book 'The Tangle of Science: Reliability Beyond Method, Rigour, and Objectivity', which I have co-authored with Cartwright, Hardie, Montuschi and Soleiman proposes a new answer to the question 'why should we trust science?'. This paper, which is a modified version of a chapter of the book, constitutes an attempt to use this new theory to help solve an ongoing debate in political science, namely why the theory of the democratic peace holds. Before going into detail with the principle of the democratic peace and how it can be secured, it will be helpful to briefly acquaint the reader of this dissertation what a 'tangle' is, and what features of it are relevant to the following discussion.

The Tangle is a concept designed to help explain the reliability of science. It is, we argue, the piece missing from the standard narrative that science works because it is objective, rigorous, and follows the scientific method. These are, certainly, a part of the story, but either separately or together they cannot be enough to secure reliable scientific results. Instead, underpinning all three of these 'usual suspects', and any particular scientific product more broadly, is a set of connections that weave together the supporting pieces of science in the right kind of way to make it credible that a given product will do what we expect of it in a given context.

We use the word 'reliable' very deliberately here. Indeed, it is a focus on reliability which distinguishes our account of science from others. This is in contrast to the standard narrative which holds that science works because it is true - or rather because scientific claims can be *confirmed*. In *The Tangle of Science* we step away from truth for a number of reasons including that by focusing on truth we loose sight of a good number of the scientific products which help make science so successful and yet are not candidates for having truth values. Thus, when we talk about the tangle we aim to bring back into view all these neglected parts of science including theories and models, concepts and measures, studies and experiments, data collection, curation and coding, methods of inference, narratives and devices, technologies, designs and science-informed policies, amongst many others. All of these products of science are expected to be reliable for the jobs we put them to an, in turn, are critical parts of the complex network of connections (a tangle) that, when connected in the right kinds of ways, support the reliability of other particular pieces of science. It is only by recognising the critical role of these pieces that we can come to understand exactly why science works.

The emphasis on reliability also helps bring into focus another important aspect of science - that context matters. Newton's theory of gravity is reliable under certain circumstances - building a bridge, for example, or calculating the trajectory of a ball thrown on earth. Conversely, one should not use it for calculating gravitational lensing effects around black holes or for navigating the Hubble space telescope. Why not? Because the tangles which make Newton's theory reliable in the first set of cases do not translate comfortably to the second set of cases. That is, the evidence, experiments, data, theories, models etc. which can be used to justify trusting Newton for bridge building, are absent or very weak when trying to justify it's use in gravitational lensing calculations. The concept of the tangle helps us understand why these contexts matter and under what circumstances we're warranted in trusting a particular product for a particular job.

Visually, we find it useful to think of a tangle as something akin to the floating nests of the Jacana birds. That is, for a given product of science (the eggs), there must be a supporting, carefully woven, tangle (the nest) comprised of a variety of different connecting pieces of science. These pieces must be varied and woven in the right kinds of ways, or the nest will fall apart and we cannot trust the safety of the eggs.

The better a supporting tangle is, the more likely it is that the product it supports will reliably do the job asked of it. Virtuous tangles, as we call them, have three particular properties;

> **Rich**: Not only does the tangle have a lot of closely connected pieces but these pieces are varied in type in the right kinds of ways.
>
> **Entangled**: The pieces relate to each other and to the product/aim pair in question in a variety of different ways.
>
> **Long-Tailed**: The pieces figure in support tangles for other scientific products in other domains that are succeeding at other kinds of difficult jobs, including successful interactions with the empirical world.

When a product of science has a tangle which exhibits these properties, it is more likely to be trustworthy to do the job it is aimed at. It is worth noting, however, that a virtuous tangle doesn't guarantee reliability, it only makes it more likely. Plenty of tangles throughout the history of science have looked (and been) virtuous for a given job, despite later being discarded in favour of new scientific products which work for a wider range of aims.

The Tangled Principle of the Democratic Peace

'Democracies don't go to war with other democracies'. This is the democratic peace (DP) principle, which we will discuss extensively in this chapter. This principle has been the subject of a huge body of research in political science over the last 60 years. It is widely assented to, but the reasons for it holding are heavily debated. As one distinguished political scientist reported to us, 'Everyone took it to be true, then jumped in trying to explain it'.[1]

As with, we hope, most of our chapters, this chapter has ideas to offer independent of the role it plays in developing and defending our overall claims about tangles and their role in supporting reliability in science. In particular, here we offer our own positive account of the democratic peace principle that we claim makes more reliable (although more restricted) predictions than any of the other accounts we have found and has behind it a better tangle to warrant that reliability.

With respect to our overall project to explain and defend the tangle, we use the extensive studies of the democratic peace to do four jobs.

First. Our own positive contribution to the democratic peace literature introduces new products into the tangle of support for the principle. As is often the case, this alters the product (in this case what the principle says) and refines the description of the job it can be expected to be reliable for. This illustrates the usefulness of the tangle, and of the focus we urge on a broad swathe of product types, as instruments for evaluating and contributing to ongoing scientific debate

Second. In our book section entitled *'In defence of the tangle: Constraints make it*

[1] Stephan Haggard in conversation February 25th, 2020.

hard to misstep' we argue that tangles make failings less likely by supplying a network of constraints that makes it hard for something to go wrong. There we explore briefly how a particular measure of democracy, V-Dem, is constrained by a surrounding network of empirical and theoretical research. In this chapter we use the work around the democratic peace to illustrate how constraints help build tangles in contexts where there are simultaneously in play multiple competing understandings of what looks to be 'the same' principle and multiple competing theories about it. Although we illustrate with a social science example, the points we make hold equally in the natural and the social sciences. These constraints can push in opposing directions: they can drive divergence between rival theories or they can encourage them to converge on similar positions. In Section 3.4 we illustrate how the Ballung nature of the terms "democracy" and "war" pushes theories apart, expanding on Hempel's concept of alienation that we discuss in the book section *'What is Objectivity'*[2] Conversely, we show how the existence of standard touch-stones for the field, as well as inter-theory critique, prevents theories from moving too far away from one another, as they are required to draw on similar resources to build their tangles or risk being discarded by the scientific community.

Third. Our chapter '*Illustrating the Tangle: Episodes from the History of Science'* provides examples of tangles in different domains of the natural sciences. This chapter provides concrete examples of tangles in the social sciences and of what our three criteria for virtue look like in a real social science case. We are keen to look in detail at an example in social science because tangles in the social sciences may differ considerably from those

[2] ***'An explication sentence does not simply exhibit the commonly accepted meaning of the concept under study but rather proposes a specified new and precise meaning for it'. [58, p.663]***

in the natural sciences for all the standard reasons that explain why success in the social sciences is generally harder to obtain than in the natural sciences. These include a number of problems facing the social sciences that are particularly salient in our discussion of the democratic peace:

1. As Max Weber (1904, ed. 1949) [145] argued, the social sciences are supposed to study concepts we care about. They are not at liberty to forsake these and shift focus to those that behave nicely. (Recall mention of this in Section 3.3.4.)

2. These concepts are very often difficult to define precisely and tend to manifest differently in different settings.

3. The things picked out by a great many social science concepts are socially constructed, like marriage, poverty, socioeconomic status, money.

4. So their behaviour is governed not by natural law but by convention, legislation, and habit, and although any of these may promote certain behaviours they far from guarantee them.

5. The social sciences are generally expected to deal with complex open systems and are not able to shield them off in the way that so many of our successful physical creations are, as in the casing on an ordinary battery, the capsule that commonly encloses the drugs we swallow, or the Faraday cages that shield MRIs. As the Norwegian economist Tyrgve Haavelmo, who won a Nobel prize for his work in founding econometrics, notes: Physics has it easy. No-one asks physics to predict the course of an avalanche. But economists are expected to predict the course of the economy.

6. Relatedly, as JS Mill (1836, ed. 1967) [87] argues, by contrast for instance with the planetary system, there are apt to be a large number of factors at work at once and in shifting arrangements. This makes prediction extremely hard.

These special problems facing social science help explain why the notion of a Ballung concept is particularly important for the social sciences, why social science principles must so often be in the form of generics, and why overdetermination of outcomes is so central for reliable prediction there: to secure an effect with relative certainty, throw all the causes possible at it. These three — the Ballung character of "democracy" and of "war", the generic form of many social science principles, and the importance of 'overdetermining' outcomes — all play a central role in our own account of the democratic peace, as we explain next.

Fourth. Although we claim that our account is normative, we do not see ourselves as recommending that the sciences do anything different in general from what they already do. The tangle is always there in good science, although that has not, we think, been fully realised and articulated.

In particular, filling gaps in a tangle to make it more virtuous is business-as-usual in science. There are two ways to do this that we have only implicitly noted before. One is to propose something new that you think may do the job required. That can be a path to scientific advancement and discovery—better science in the future. But it is generally not the soundest way to fill a hole in a tangle for this product to do this job here and now. If this tangle is to provide support here and now for the reliability of this product, you need a filler that itself is warranted, here and now, as reliable for the job that needs doing in the tangle. So generally for this task it is better to take a product 'off the shelf' — of course, as always, with caution and requisite adjustment. This is one of the principal reasons it is important for scientific products to be clearly labelled, even if only implicitly, ‘This product is attested as reliable for this job’. That is a central way in which science can claim to be cumulative and to build an ever-expanding store of new products efficiently on its store of already attested ones.

This chapter provides an illustration of how products developed elsewhere that weren’t at first seen as relevant to the task at hand can be taken off the shelf and, with appropriate adjustment, put to good use in new contexts. If successful, it is a poster child for wide-ranging interdisciplinarity. For in the case of our own development and defence of the reliability for prediction of a new version of the democratic peace principle, one of these shelves is labelled ‘philosophy products’. In particular we use the three products that, we note above, help address the problems that make life difficult for the social sciences, two of which are philosophical contributions.

The first is our defence of the practice of retaining and employing loose Ballung concepts in science. In this case the relevant concepts are “democracy” and “war” (though we focus primarily on democracy for illustration). We propose a special reading of the democratic peace principle itself and of the predictions it makes using Ballung-style concepts. We argue that the 60 or so years of intensive work on the democratic peace provide good reason to predict that what count as “democracies by anybody’s book” will not engage with each other in ways that will count as “war by anybody’s book”. In proposing this, we suggest that the principle should be read as a principle defined only over core cases: the DPP-by-anybody’s-book (DPP$_{BAB}$). We explain this in Section 3.5.

Second, in analogy with our defence of loose concepts we also, throughout the book, illustrate the central role played in science by generic principles and the practices that the social sciences have evolved for their use, despite their inexact content. This is how we read the standardly offered explanatory principles for the democratic peace. We take these to be generic principles that pick out what Jon Elster calls ‘mechanisms’ — which often refer to dispositions of individuals, institutions, or structures in society. These are like the causal principles in the process-tracing theories of change in the book section '*Tackling the job of prediction directly'* — the causes cited may need to be triggered, they generally need support factors to operate, they can be interfered with, and more than one can operate at once. This is not the usual reading of these principles in the democratic peace literature, but we believe that it is the most plausible one, and it provides a very different take on the democratic peace principle than is otherwise available.

The third product we call on is a familiar tactic that is used successfully over and over again in practical technology and social policy: when you can’t be certain that any one cause can be relied on to carry through start-to-finish to produce the effect you need, introduce a number of different causes, the more the better so long as they do not interfere with each other. We use this idea to build a kind of overdetermination argument for the reliability of the democratic peace principle.

In the case of the warring camps in high temperature superconductivity (described in the Preface), the great volume of work that supports the explanatory mechanisms of each camp undermines the claims of those endorsed in others since it is assumed on all sides that only one of these explanatory mechanisms can be at work. So too with the explanations of gravitational waves that we talk about in the Afterword.

The opposite is the case with the explanatory mechanisms offered for the democratic peace, as we read them. Each can operate under certain circumstances, and in many cases, the circumstances allow that many can operate at once. So, we shall argue, in 'bog-standard' cases peaceful outcomes will generally be overdetermined by a number of different mechanisms all working to the same outcome. You can expect it to be unusual for all to suffer sufficient interference at once in cases where democracies-by-anybody's-book come up against one another. We develop this argument in Section 3.6.

We begin in Section 3.2 by describing the bones of our argument and in Section 3.3 by briefly surveying the literature on the democratic peace to provide context for what is to follow.

3.2 The Democratic Peace

In 1898, French and British troops found themselves at the centre of an international crisis focused on Fashoda, Sudan. Several dozen members of the French military, seeking to evict the British from Egypt, took possession of a fort in a region upstream of the Nile. Some months later they were confronted by a several-hundred strong force of combined British and Sudanese troops. While interactions on the ground were largely amicable, back in Europe the incident became the locus for recent tensions between the two imperialist powers over control of Africa. The British public in particular agitated for military intervention, while the French were more reticent, mindful of a weaker armed presence at the fort and unwilling to risk alienating Britain with an aggressive Germany on their borders. After months of negotiation and with rising fears of wider conflict in Europe over the incident, the French finally ceded the fort to the British, ending the standoff and paving the way for British dominance in the region.

This result might be thought of as surprising. Prior to this, Britain and France had had a rough relationship characterised by both short and extended conflicts, and the two were currently locked in a land-race for territories in Northern Africa lending pressure to defend their position aggressively. Given this, and despite worries about Germany, one might expect that such a confrontation would, as it often had before, lead to open conflict between the two nations.

What made the difference? Political scholars argue that it was the newly extended democratic nature of both. That is, the Fashoda Incident, as it has come to be known, is seen as part of a broader pattern in international politics where democracies do not go to war with one another. This is the central principle of the Democratic Peace. It is as close as political science gets to a foundational rule. Empirically, most academics agree that the principle holds and continues to hold across almost all modern and historic international conflicts. There are as of today no generally agreed upon instances of generally-agreed-upon democracies going to generally-agreed-upon war with each other. Democracies don't go to war, and so one could have reliably predicted that the Fashoda Incident would not end any other way. The DP has been the backbone of many areas of political and international policy research for decades. It helps set agendas, drive treaties, and has been cited by numerous American Presidents as central to their international politics (for example, Wilson (1917) [150] and Bush (2004) [14]). Despite this broad-scale acceptance and use, however, the means by which the DP instantiates remains controversial. Numerous explanations have been proposed, but no one has emerged as a clear winner.

Proponents of these explanations have long argued that their particular theory is the ‘correct’ explanation for why the principle holds. We think this is a mistake.

Our take on the DP principle involves three theses.

1. Many of the different theories on offer about the DP principle have what look to be reasonably virtuous tangles to back them up. So there is good reason to expect that they reliably identify explanatory mechanisms responsible for the DP principle.

2. The different credible theories converge on what we call ‘the democratic peace principle by-anybody’s-book’ (DPP_{BAB}). The tangles behind these theories are all entangled with one another in multiple ways. This tangle of tangles supports the reliability of the predictions about peace and war that follow from DPP_{BAB}.

   As part of this, one of the primary goals of this chapter is to expand on the constraint work done in the chapter *‘In defence of the tangle: Constraints make it hard to misstep’*. We single out three kinds of constraints that do much of the heavy work in support of reliability for the individual theories and the DPP_{BAB}.

   (a) The kinds of constraints demanded of good science, like consistency, looking for and responding to empirical evidence, and ensuring concept validity, to name a few. These force theories of the DP to adopt precise definitions of the concepts they employ, beginning with democracy and war and building out from there. We illustrate how these constraints reinforce divergences among the different theories. We call constraints divergence constraints when they function this way.

   (b) In contrast to this, we claim there are two types of convergence constraints at play which entangle the theories together. These take the form of,

   i. Responsiveness constraints, generated by the demand that each theory be responsive to the standards of the field.

   ii. Competitive constraints. As we review in discussing HD-with-bells-on in our chapter '*In defence of the tangle: from truth and confirmation to the tangle'*, it is not enough for a theory to provide evidence for its own hypothesis. It should also show what's wrong with alternatives. Each theoretical camp is watching for problems in the alternatives and working to guard against any that competitors might identify in their own explanations.

3. Finally, we conclude that the different theories are not in competition after all. The explanations they identify can all be 'correct'. This is because the explanations use not universal but generic principles that pick out Elster-style mechanisms, and these can all operate at once.

The existence of these divergent and convergent constraints helps us see why the DP principle is reliable for making predictions about conflicts. While each theory predicts its own particular version of the principle all agree on what we term the DPP-by-anybody's-book. Contrary to standard robustness arguments, however, we do not take this convergence to suggest that there must exist some single 'correct' explanation for the DP principle. In contrast to what is commonly supposed in the literature, no individual theory is, in itself, enough to ensure that democracies don't go to war with each other. Instead, each of the wealth of theories, constrained by and interwoven with their respective tangles, provides one possible Elster mechanism for the DP principle, described in a defeasible generic or *ceteris paribus* law. Taken together, they ensure that democracies-by-anybody's-book don't go to war-by-anybody's-book. In many specific conflicts various of the proposed mechanisms may fail to trigger, be shielded off, or otherwise be absent, but the chance that all of them fail is small. So the DP principle is not a case where political scientists should be looking for the 'one true explanation'. It is a case where the combined weight of the theories is, itself, the reason the DP principle can be expected to hold—at least between bog-standard democracies with respect to bog-standard wars.

Our argument is, as we note in the chapter *'What's so good about a virtuous tangle?'*, reminiscent of Woody's 2015 [152] functional analysis of science, arguing as she does that science is 'a coordinated activity of communities'. Convergence of the kind discussed here may be seen as one part of her broader coherence work on how science is reliant on scientists working as a group to constrain divergence.

3.3 Theories of the Democratic Peace

There are a great number of highly developed theories of the democratic peace principle on offer, some compatible, others conflicting – each with its own tangle of support. It will help going forward to briefly review a number of these proposed theories. Following Harald Müller and Jonas Wolff (2006) [93], we roughly categorise them along two dimensions—the monadic/dyadic divide and the normative/structural divide.

The monadic/dyadic dimension attempts to capture something about how democracies might interact with one another. A monadic theory is one which holds that there is something internal to democracies which makes them less prone to enter into wars, an effect which is amplified when their opponent also possesses these particular traits. Conversely, a theory may take there to be something critical to dyads where both states are democracies. Under this type of theory, democracies are just as warlike as their non-democratic counterparts, but some new element comes into play when they face off against another democracy.

The normative/structural divide attempts to capture a difference in what theories think is the driving force behind the peace. Normative explanations take the role of democratic institutions to be transmissive in nature — they allow normative preferences for peace and conflict-avoidance to guide national policy, but the structures themselves have no role beyond this transmission. Structural explanations, conversely, take the institutional features of democracies to be central to the success of the DP principle. Here, these structures play a larger role, often in a capacity that forces one or both parties involved in a conflict to back down or delay allowing time for conflict to be resolved, defused, forgotten about or otherwise avoided.

The following table highlights four current DP principle theories which demonstrate the monadic/dyadic and normative/structural divide. This is not, of course, anywhere near a complete listing.

Table 3.1: Types of democratic peace principle explanation arranged by mechanism type.

	Normative	**Structural**
Monadic	Kant's Perpetual Peace	Increased Signalling
Dyadic	Social Constructivism	Democratic Commitment

3.3.1 Kant's Perpetual Peace [Monadic-Normative]

Immanuel Kant is a monadic normative theorist. In his book Perpetual Peace: A Philosophical Text (1795, ed. 1903) he suggests,

> [T]he Republican Constitution [...] includes also the prospect of realizing the desired object: Perpetual Peace among the nations. And the reason of this may be stated as follows. According to the Republican Constitution, the consent of the citizens [...] is required to determine at any time the question, 'Whether there shall be war or not?' Hence, nothing is more natural than that they should be very loath to enter upon so undesirable an undertaking; for in decreeing it, they would necessarily be resolving to bring upon themselves all the horrors of War. [...] On the other hand, in a Constitution where the Subject is not a voting member of the State [...] the resolution to go to war is a matter of the smallest concern in the world. [68, (Section II, Part I)]

Ernst-Otto Czempiel (1996) recasts this in the following way,

> The rational citizen in liberal capitalist societies is generally peace prone because war endangers not only his life (as combatant or civilian victim), but is economically expensive as well. If the political system allows for the translation of this preference into foreign policy (like democracy does), the respective state will refrain from violent behaviour. (In Müller and Wolff (2006), [93, p.80])

This can be seen as a normative theory because the relevant factor is the preferences of the citizens, the democratic institutions merely allow for preference transmission. It is monadic because democracies are supposed to have this tendency towards any war, regardless of their opponent. When both sides are reticent, war is avoided. In mixed dyads, however, the non-democracy acts as the instigator and can force the democracy into war against the citizenry's preferences.

A related theory, with similar structure but slightly different content, can be found in the 'cultural argument'. Müller and Wolff characterise this particular normative/monadic theory as the idea that,

> Democracies (democratically socialised citizens and leaders) are used to and prefer to solve their conflicts in peaceful and consensus-oriented ways. In democratic societies prevails a "democratic norm of bounded competition". [93, p.9]

That is, whereas Kant takes it that all citizens wish for peace but only democracies make this relevant on the inter-state level, one might alternatively think that citizens of democracies are socialised to be less conflict-prone.

3.3.2 Social Constructivism [Dyadic-Normative]

Thomas Risse-Kappen (1995) is the progenitor of this theory, arguing that

> [Democracies] to a large degree create their enemies and friends—'them' and 'us'—by inferring either aggressive or defensive motives from the domestic structure of their counterparts. On the one hand, they follow behavioural norms externalizing their internal compromise-oriented and non-violent decision rules in their interactions with other democracies. On the other hand, the presumption of potential enmity creates a realist world of anarchy when democratic states interact with authoritarian regimes. [119, p.1]

Because democracies follow internal norms of peaceful conflict resolution using compromise and negotiation, Risse-Kappen suggests that the DP is the result of democratic states assuming similar norms for other democracies. Thus they are more likely to trust overtures and negotiate in good-faith, assuming the other party is similarly peaceful. In mixed dyads, however, the democracy has no such belief, instead often seeing autocracies and the like as overly aggressive and untrustworthy, exacerbating tensions.

This, in turn, tends to lead to clusters of democracies working together along several dimensions including trade and military agreements. This further reduces the chances that these states engage in open war with one another. Conversely, such groupings tend to exclude non-democracies leading to a prevalence of tension between democratic and non-democratic states.

3.3.3 Increased Signalling Theory [Monadic-Structural]

This monadic-structural explanation for the DP principle draws on the idea that states are rational but lack perfect information. In the absence of complete information about their opponent's desires, commitments, and capabilities they are unable to accurately call bluffs, judge stakes, or trust overtures. This, in turn, leads them to engage in war with other states when peaceful moves are disbelieved or bluffs are misread as more aggressive than intended.

Under this assumption, states which are better at signalling their motivations and capabilities are less likely to be drawn into wars with other states. The Increased Signalling theorist, then, suggests that because of how democracies are internally structured they signal more openly and accurately than other types of government, leading them to engage in war less often. This property, in turn, is boosted in democratic dyads where both states are clearly signalling to one another.

The types of structures that are central to this theory vary from proponent to proponent, but the generally offered ones include the existence of elections which are determined by the public. Elected officials, wishing to stay in office, are under pressure to conform with general public opinion—a widely available piece of information for other states due to the existence of free speech and public debate in democracies. Thus, if the leaders of a democracy signal in line with public opinion, they are more readily believed, while signals which go against the public are more often accurately called as a bluff. Similarly, the Increased Signalling theorist might point to the existence of oppositions as a means of increased information—governments which are backed by their opposition are less likely to be bluffing, governments which are opposed by them are likely to not be as strong as they wish their rivals to believe.

3.3.4 Democratic Commitment [Dyadic-Structural]

This particular theory rests on the idea that leaders in democracies face harsh penalties if they enter into, and then lose, a war in that they are likely to be replaced by the public in the next election. Conversely, leaders of autocracies need only keep a small subset of supporters satisfied, something that can be achieved via various, often easily accessible, means. Thus, democratic leaders are less likely to enter into a war unless they're certain to win, and once engaged they are likely to commit more fully than their autocratic rivals.

Democracies are, then, more difficult opponents to provoke and more deadly once war has been entered. In mixed dyads, this manifests as a slightly increased chance of war, as democratic leaders know they have a better chance of winning against autocracies which will put less effort into the fight. Conversely, when two democracies butt against one another the leaders of both sides are aware that their opponent would be more difficult to defeat, leading to a trend away from war as the states are wary to start something unless they are certain of victory.

3.4 Defining democracy and war — fuzzy and precise boundaries

As we argue in the chapter '*The Tangle*', in order to have a credible claim that they provide a correct explanation of the DP, each of these theories must be supported by its own rich, tangled, and long-tailed network of definitions, measures, principles, theories, models, observations (amongst other things). They must use sound principles in sound ways with sound, validated concepts.

Central to our work in this chapter, however, is that each theory has a different underpinning tangle, though the different tangles share much in common. They may draw on different measures of public opinion, emphasise different structures of state government, or require the existence of different means of information transmission between states and citizens, to name a few.

In this section, we focus on constraints that enforce this divergence. In particular, we look at the role that definitions of "democracy" and "war" have in constraining a DP theory. These terms are, as they come, Ballung concepts, fuzzy at the edges and comprising a set of overlapping but related defining features. Their fuzzy nature is one significant source of theory-tangle divergence, with each theory attempting to precisify the concepts along different lines. As we will show, however, divergence can only go so far, since each definition is beholden to the same standard candles of the field, which we term responsiveness constraints.

3.4.1 Divergent definitions and measures of democracy

Setting aside “war” for the moment, let us focus on “democracy”, which is so widespread a concept that one would be hard-put to find many academics who agreed entirely on how it is to be defined and measured. Attempts range from that found in Frederick Whelan’s 1983 book,[3]

> [...] [G]overnment with the consent of the governed. This formula is indeterminate with respect to institutional forms, or the procedures by which consent is to be expressed - questions on which consent theorists have historically differed. [147, p.15]

to Elmer Schattschneider (1960),

> [...][A] competitive political system in which competing leaders and organizations define the alternatives of public policy in such a way that the public can participate in the decision-making process. [125, p.141]

to Robert Dahl (1989), who provides a more formal set of criteria whereby for a state to count as a democracy, its citizens must have

1. Effective participation
2. Voting equality at the decisive stage
3. Enlightened understanding
4. Control of the agenda
5. Inclusiveness [31, p.109-115]

[3] Note that this isn’t a definition Whelan himself endorses, but one he recognises as a common, if underspecified, attempt at defining democracy. We have used his wording because it is a nice succinct definition.

We have also already talked about the variety of measures put forth by political science including V-Dem, which we discuss extensively in the Section *'In defence of the tangle: Constraints make it hard to misstep'*, including the various constraints and elements that go into the tangle that supports their reliability as accurate measures of democracy.[4]Alternatives to V-Dem include Freedom House, which takes into consideration two axes along which states are scored—political rights and civil liberties; Vanhanen's Index, which ranks nations on competition and participation and provides an aggregate score; and Polity IV, which uses weighted scores for a number of categories including 'Competitiveness of Executive Recruitment', 'Constraint on Chief Executive', and 'Regulation of Participation'.

These discrepancies in definition and measures are unsurprising when one considers that democracy, much like the terms "poverty", "objective", or "chair", fails to have a clean set of defining features, but instead is, to paraphrase US Supreme Court Justice Potter Stewart (1964) on pornography, the kind of thing where 'you know it when you see it' [132].

3.4.2 Divergence constraints – making "democracy" precise

The formulation of a proper scientific explanation for the DP, however, requires that theorists give a clear definition of what they mean by "democracy". It is not enough to gesture at the Ballung concept and claim a theory applies over its instances since an explanation is necessarily predicated on particular democratic properties of the state. This predication, in turn, generates a subset of the Ballung democracy concept for which the theory supposedly holds, thus doing violence to the original fuzzy concept by moving from the vague to the

[4] ***V-dem is a measure of democracy which measures seven principles; electoral, liberal, majoritarian, consensual, participatory, deliberative, and egalitarian. These principles are each represented by a separate index and are designed to reflect the multi-faceted nature of democracies.***

precise. Recall Hempel from Section 3.3.2, ‘An explication sentence does not simply exhibit the commonly accepted meaning of the concept under study but rather proposes a specified new and precise meaning for it.’ [58, p.663] What we see here is a form of alienation, with all the benefits and downsides we point out in our discussion of objectivity.

Take, for example, Kant, who emphasises the role of constitutional structures in democracy which enshrine equality:

> The only constitution which derives from the idea of the original compact, and on which all juridical legislation of a people must be based, is the republican. This constitution is established, firstly, by principles of the freedom of the members of a society (as men); secondly, by principles of dependence of all upon a single common legislation (as subjects); and, thirdly, by the law of their equality (as citizens). [68, Section II]

This definition fails to capture aspects other theorists take to be central to a democracy. In contrast to Kant, the Increased Signalling theorists’ definition of democracy must involve the existence of elections, a free press, or open debate because it is upon these structures that their proposed explanation hangs, while the Social Constructivist needs to consider democracies to be those states with internal norms of peaceful conflict resolution. Including states which lack this property means creating a theory which necessarily cannot hold over the objects it is defined across.

Indeed, each DP theory will require some different function from the concept “democracy” and so, while some may share a definition, many will diverge in the precisification they adopt. Note, however, that by moving to a precise definition of democracy, the concept necessarily loses information. What was once a fuzzy set of related concepts is now rigidly defined along one or more specific dimension. Though moving from a Ballung concept to a precise definition is a requirement of a functional theory, it necessarily involves excluding states which other people may wish to count as democratic, thus forcing divergence between theories.

3.4.3 Convergence constraints – how uncontroversial examples constrain the tangle

Each theory has a different meaning for the term "democracy", brought about by divergence constraints demanding precisification. But no DP theory can use a definition which differs too greatly from the core concept or it risks being discarded. This gives rise to convergence constraints. These come in two flavours:

A. Responsiveness constraints, which include,

   i. The existence of uncontroversial examples of democracies and non-democracies, which is the focus of this section.

   ii. The existence of extensive literatures about democracy, as Crasnow stresses (recall the Section *'What's in a tangle?'*)

   iii. The existence of diverse measures that give roughly correlated results over various ranges, as we illustrate in the Sections *'The 'virtuous' tangle: Three general features'*, *'What's in a tangle?'* and 3.4.1.

   iv. The existence of respected databases that record which states have which of a variety of different features associated with democracy, as for example V-dem, Polity IV, and Freedom House.

B. Competitive constraints: The existence of critique from other theorists, which we return to in some depth in Section 3.4.5. Note that competitive constraints can also be a force for divergence as theories attempt to distance themselves from problems they perceive in other theories.

Here we focus on the first responsiveness constraint, the existence of uncontroversial examples of democracies and non-democracies, which plays a critical part in our argument for why the DP principle is reliable for making predictions, drawing each theory inwards and forcing them to converge on a single, overarching, formulation of the DP principle.

Uncontroversial examples are what we term "democracies-by-anybody's-book". Democracies -by-anybody's-books are examples of democracy which must be covered by any definition of the term offered by any theory or that theory risks having an unforgivable hole in its tangle. Examples of democracies-by-anybody's-books include Australia, the United Kingdom, and the United States of America. Any definition of democracy which fails to include these, by any theorist's light, has made a mistake and so it, and the DP principle theory which draws on it, is generally discarded from the mainstream body of science. Consider, for example, a definition of democracy which claims that only states in the northern hemisphere can be democratic, excluding democracies-by-anybody's-books like South Africa and New Zealand. It would immediately be thrown out by the political-science community.

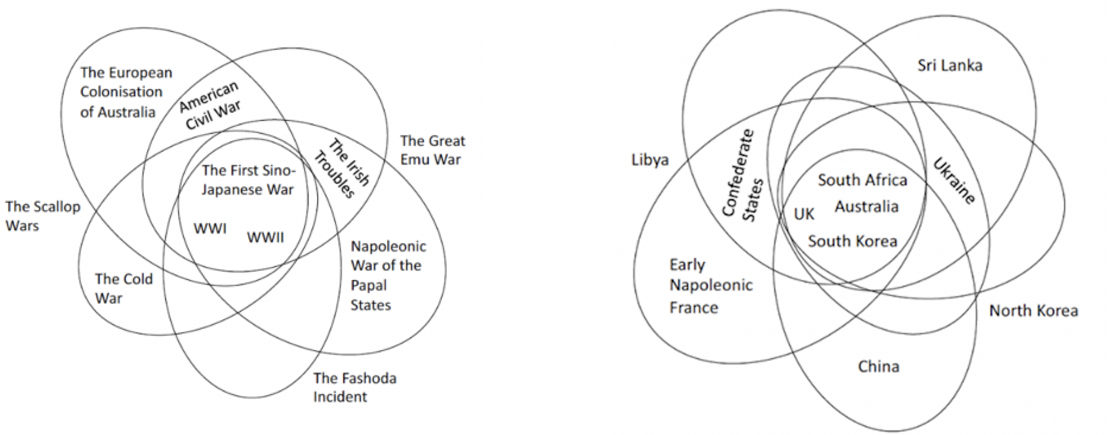
The existence of a set of pieces to which all tangles must be responsive is something found in every field. Scientists can and should discard any theory of quantum mechanics which doesn't fit with the results of the double-slit experiment, for example, or an explanation for fish evolution patterns that doesn't consider the salmon a fish. To ignore these central pieces is to have a bad tangle because one is missing critical, agreed upon, standards for the field. Again, we see the scientific community constraining divergence. [5]

A similar argument may be run on the definition of "war". Each theory of the DP principle will attempt to precisify the Ballung concept, losing fuzzy cases in the process and creating disagreement over how the concept is used. Just as with democracy, there will also be a set of wars that are "war-by-anybody's-book" (eg. World War I and II, the Vietnam War, and the first Sino-Japanese War), and sets of conflicts that become points of contention for their inclusion or exclusion under the term (eg, the Cold War, the Irish Troubles, or the existence of large-scale, persistent cyber-attacks between major world powers).

We should note that just as there are democracies-by-anybody's-books and wars-by-anybody's-books, so too will there be not-democracies-by-anybody's-books and not-wars-by-anybody's-books. It is just as bad for a definition or measure of democracy to decide that North Korea is a democratic state, or that the "Scallop Wars" of 2012 and 2018 between French and British fishermen counts as genuine inter-state war. Going forward we focus on the positive inclusion restrictions, but readers should bear in mind that this is only half the story.

You can visualise the varying definitions of democracy and war as Venn diagrams, with the core examples that all theories must include in their definitions sitting in the centre (Figure 3.1).

[5] This is not to say that reasonable theories which reject the stable examples cannot both exist and do well. To make such a move, however, requires a significant tangle of work behind it to justify why one should disregard the established norms of the field. It requires a richer, longer-tailed, and more entangled tangle to make up for, and justify, the significant gap that comes with disregarding a standard example. Such instances are unusual, but they do occur in science. Take, for example, the creation in biology of the fungi kingdom, which moved mushrooms, at the time an example of a plant-by-anybody's-book, out of the taxonomic grouping of plants and into the newly discovered category. Similarly, reclassification of the term "planet" moved Pluto — up until then a clear-cut case of a planet as far as most people were concerned — to the new "dwarf planet" grouping. In both cases overwhelming evidence was presented that the current systems were mistaken about the core examples which justified the rejection of "by-anybody's-book" elements.



(a) War

(b) Democracy

Figure 3.1: Overlapping concepts of war and democracy

3.4.4 Inter-theory criticism, fringe cases, and constraining the tangle

Once one has established the subset of elements which must be included or excluded by a measure or definition of democracy or war, inter-theory criticism (competitive constraints) drives how the other responsiveness constraints are applied. There are two ways this occurs.

1. The field as a whole will reject any theory which fails to adequately meet the responsiveness constraints. This acts to excise theories which wander too far from the rest along the given dimension.

2. Each theory, due to including and excluding a different subset of tangle pieces from other theories and other domains, will criticise other theories for failing to meet what it judges to be appropriate standards.

The second way rests on the existence of fringe cases. Fringe cases, in this context, are those states which may or not be democracies depending on how you look at them.

This directly leads to inter-theoretic critique: because Theory A includes 1849 France as a democracy while Theory B does not, the two will act to constrain each-other, pushing and pulling the other theory to justify its inclusion or exclusion.

Such conflict is inevitable. While precise definitions cannot wander far from the core set of democracies-by-anybody's-books, the very act of precisification will create edge disagreements as some theories include or exclude states that others take to be clear democracies or clear non-democracies by their own light.

Why think France was a democracy in the period leading up to the outbreak of the War of the Roman Republic between Napoleon and the Papal states? Recall Whelan's 'government with the consent of the governed' definition from Section 3.4.1. Napoleon, the leader at the time, had the support of the masses and was duly elected leader of France, and so it seems it must count as a democracy. Alternatively, Schattschneider required that 'the public can participate in the decision-making process'. By this light, France was not a democracy—Napoleon infamously declared himself Emperor shortly after this period, indicating a trend away from public participation and towards dictatorship, and France at the time lacked any other politically competitive groups besides Napoleon himself. [117]

Which is the correct decision? Was early Napoleonic France a democracy or not? Not only is it unclear which is right, it's unclear that such an absolute result even exists. Both Whelan and Schattschneider advocates, however, will act to force the other to justify their choice of definition.

Let us consider another example, the American Civil War as discussed by James Ray [117]. Some academics have suggested that this conflict constitutes a violation of the DP principle. A lot hinges both on whether each was a democracy, and whether the conflict between them counts as an international war between two states as opposed to an internal conflict.

On the first point, both had almost identical constitutions, suggesting that if the North (as is standardly thought) was a democracy, so too was the South. As has been pointed out by academics like Ray, however, while the South's leader, Jefferson Davis, was technically elected by representatives of each state, there had not yet been open elections such that it could be said he had the mandate of the broader citizenry. His position, moreover, was supposedly a temporary one, but war broke out before a general election could be held.

If one thinks, as the Democratic Commitment theorist does, that what defines a democracy is the ability to be removed from control, then the South looks like a democracy.

Alternatively, if one requires that instances of peaceful norms for conflict resolutions be demonstrated, the short-lived nature of the state might not be enough to make it a democracy under the Social Constructivist position, so the conflict is not a failure of the DP principle.

There are, of course, good reasons to think that both the first definition and the second are equally valid conceptions of democracy. The first prioritises legal structures, while the second attempts to exclude those states which pay lip-service to democracy while remaining in practice a non-democratic system.

If a theory wishes to use a definition of democracy that includes the Confederacy, but doesn't want to say that the American Civil War was a violation of the DP principle, then they will need to carefully choose their definition of war to reflect this, either by excluding the Confederate states from being a genuinely independent state, or by gerrymandering their definition of war so that the conflict counts as an internal civil war rather than an open war between two nations. This gerrymandering may mean an even larger discrepancy between theories, however. As one contorts a certain way to avoid violations to the DP principle, another will contort in a different direction, further limiting the types of wars that the theory is taken to be able to predict outcomes over and creating even stronger fringe-case divergence.[6]

These divergences, then, are a driving force for why DP tangles are conducive to the reliability of the theory to provide an accurate explanation of why the DP principle holds. Critique binds the tangles behind different theories together by forcing them to interact in dialogue with one another, creating a virtuous tangle of rich, entangled connections between the different theories. These connections, in turn, make each theory more reliable because the field as a whole is better at culling bad apples. Thus, if a theory's tangle is sufficiently responsive to the standard pieces, and it engages with other theories regularly without being completely discarded from consideration, one should think that it is more reliable than theories which fail to pass these bars.

Of course this goes far deeper than democracy and war. This argument applies to many of the other pieces that go into making up a "good" theory of the democratic peace. For each measure, definition, observation, study, dataset, mechanism, or other piece there will be similar constraints and norms that it must adhere to or risk a large, non-virtuous,

[6] As a brief aside, there are reasons one should be wary when one sees this type of contortion to fit evidence. It is reminiscent of the Deads Sea Scroll example from the section *'What's in a tangle?'*, which we argue is an example of a bad tangle. Recall that Wylie maintains that while the Qumran-Essene hypothesis ***(A theory detailing the origin of the dead-sea scrolls which argues that they were written by the Essenes, a sect that resided in nearby Qumran)*** is both rich and tangled, the tangle lacks long tails. Each piece turns inwards to support its own internal consistency and story rather than drawing on broader claims, observations, and theories from archaeology, anthropology, theology, geology, and more. Similarly, the twisting that needs to happen to include or exclude controversial examples of the DP will potentially reduce the long-tailed nature of the tangle as the theories start to pull further away from broader work in other fields on democracy and war, looking inwards and lacking strong external support.

hole in its tangle and thus also risk being excised from the field.

3.5 Inter-theoretic critique and the reliability of the democratic-peace-by-anybody's-book

We have spent a lot of time on the pieces that make up a DP theory and how they are constrained, but what does all this mean for the reliability of the DP principle as a whole? In this section we will see how a responsiveness to democracies-by-anybody's-book and war-by-anybody's-book, and inter-theoretic critique, leads to the existence of a core DP principle, termed the DPP-by-anybody's-book. This, we suggest, is the DP principle that has proved so useful in political science, not some supposed correct, precise formulation which has so long been the goal of DP principle theorists. As we've just seen, each theory of the DP will have its own particular formulation of democracy and war which it is predicated over. This, in turn, means that each proposed theory of the DP will have its own unique version of the DP principle. Thus one no longer has 'The Democratic Peace Principle', but instead there is DPP$_{Kant}$, DPP$_{SocialConstructivism}$, DPP$_{IncreasedSignalling}$, and so forth, each one supported by a distinct tangle and providing a distinct account of why the DP principle holds. Moreover, these different formulations make different predictions, some of which are incompatible with each other. Each posits rules that cover a different set of conflicts over different dyads of states. They can make predictions—whether correctly or not—only over these instances and not over all the instances covered by other formulations.

As with our myriad of overlapping definitions for democracy and war, however, while

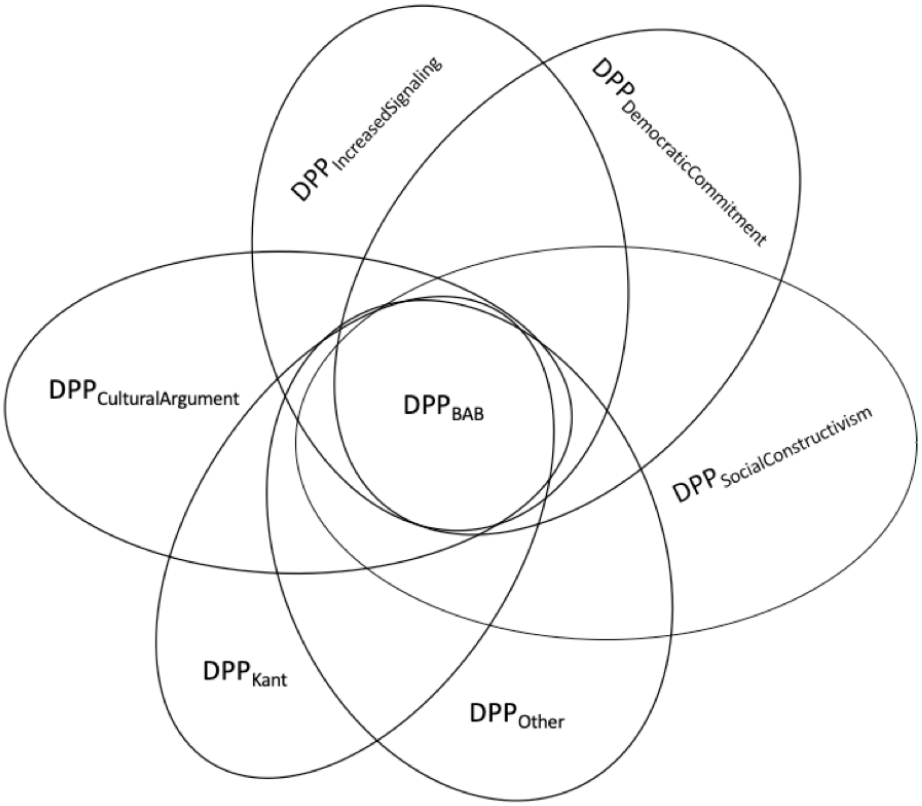


Figure 3.2: Overlapping DP principle explanations.

each theory is distinct there will also be significant overlap caused by the existence of democracies-by-anybody's-book and war-by-anybody's-book. Every theory will, necessarily, admit to the following principle:

> **Democratic-Peace-Principle-By-Anybody's-Book (DPP$_{BAB}$):**
> Democracies-by-anybody's-book don't go to war-by-anybody's-book with other democracies-by-anybody's-book.

This picks out the centre of the Venn diagram of DP theories (Figure 3.2).

Now one can begin to see more clearly why, even with the proliferation of possible DP principle formulations and fundamental disagreements between them, the DP principle remains reliable about a great many predictions. The DP principle is not one particular formulation which happens to be 'correct', but it is better thought of as a principle formulated across democracies-by-anybody's-book and war-by-anybody's-book, and thus amenable to explanation by a host of different theories.

Let us return to the Fashoda Incident. Why could you reliably predict that it wouldn't end in war between France and Britain? Because both countries at the time were democracies-by-anybody's-books. Very much in brief, according to the Perpetual Peace Principle, both Britain and France were countries where leaders were in some way dependent on the general public to hold their seats, Britain having passed the Representation of the People Act and Redistribution Act in 1884 and 1885, which increased the voting population to approximately 60% of all men and France being firmly in the middle of the Third Republic which had universal male suffrage to elect the chamber of deputies which, in turn, dictated the ministries. Each had a common legislation (dictated by Parliament in the UK and the Chamber of Deputies in France), freedom (provided one was a man with land), and applied the law equally to all citizens (at least officially). Thus, according to the Perpetual Peace theorist, neither country would go to war-by-anybody's-book with the other.

For the Social Constructivist, both France and Britain were systems which used internal norms of peaceful conflict resolution, including parliamentary debate and voting. Thus France, when presented with the possibility of war with Britain in tandem with threats from the non-democratic Germany on its border, was more likely to infer from their own norms that British overtures for peace could be trusted more than German ones and to seek alliances along those lines.

Increased Signalling theorists simply require that there be open and externally accessible debate within both countries, which is given by the existence of rival parties sharing a common legislative floor in both states.

And finally, the Democratic Commitment Theorists takes both France and Britain to be democracies for the purposes of their theory because, as with Kant's Perpetual Peace, both countries had leaders who were beholden to the general public for their jobs, placing them in precarious positions if they entered into wars they would lose. Indeed, the French were explicitly wary of entering into the conflict because they feared the superior might of the British military and recognised that further instability on a military front would pose a significant threat to the current government which was already embroiled in the Dreyfus affair. Leaders were hesitant to commit unless they were guaranteed victory and knew the other was in a similar position.

Thus for each of the four kinds of theories discussed in Section 3.3, the Fashoda Incident sits in the centre of the Venn diagram—the part that matters for DPP$_{BAB}$ – even though each emphasises different aspects of the states in maintaining their theory.

Before going further it will be useful to recognise how what we have just said differs from standard robustness arguments. Robustness, as generally held in the literature, is 'the idea is that if there are many ways of measuring, detecting, producing or deriving something, and if those ways are sufficiently independent, then it is very unlikely that all of them turn out to be mistaken or erroneous.' [41, Section IV]

There are two standard versions of this, one where the input pieces are compatible and one where they are incompatible.

Begin with the compatible. This kind of robustness supposes that there are a number of different but compatible arguments with different credible premises that yield the same conclusion. You may allow that no one of these arguments has premises that are entirely certain. Yet, by hypothesis, the premises are all supposed to be reasonably credible—they have a lot to back them up. Suppose for instance you have two relatively established compatible theories, and three different kinds of experiments calling on different techniques and assumptions that have been carefully conducted as well. All lead to the same conclusion. It seems you can generally assume that it is very unlikely that all of these, all of which point to the same result, are flawed at once, a rare coincidence you are usually justified in discounting.

Where the input arguments are incompatible, the defence of the conclusion rests on the assumption that at least one of the many ways of getting to it is right. Again, it is supposed that the set of arguments is comprehensive enough to make it unlikely that all are flawed.

Our case involves a set of theories which almost all converge on DPP_{BAB}. Note that if you suppose that all of the arguments for a given result are flawed, neither form of robustness gives reason to accept that result. So if you are to accept DPP_{BAB} on the grounds of robustness, no matter which version you choose, you are left having to assume that at least one precise version of the explanation for the DP principle is correct, even though the political science community has not yet settled on which it is.

This is why we do not employ a standard robustness argument. We think that there is good reason to take the predictions of DPP_{BAB} to be reliable. But we also think there is good reason to suppose that none of the proffered explanations are 'the correct one'. So our defence of the reliability of DPP_{BAB} must rest on different grounds.

3.6 Overdetermination and the *ceteris paribus* nature of democratic peace principle theories

Instead of robustness, we want to suggest that because claims about why the DP principle holds only have meaning relative to a tangle of ingredients that give the DP principle and the claims meaning, including but not limited to the definitions of democracy and war, you can't simply appeal to the idea that one explanation is the 'true' one.

There are two reasons for this. First, because no matter how good each precise theory is, it is predicated on the precisification of a Ballung concept and thus will necessarily fail to capture the broad principle in full. This, as we've already discussed in depth, is one reason each theory is exposed to criticism by other competing theories.

When one precise theory critiques another, it is often by observing legitimate failures and weaknesses in the other's inability to fully capture everything the field as a whole wants to count as an instance of the DP. No single tangle is reliable for supporting a version of the DP principle that consistently predicts the outcome of conflict between democratic dyads that are not in the limited subset it is predicated over. For instance, DPP$_{Kant}$ is only relevant for predictions about war$_{Kant}$ between democracies$_{Kant}$, and no others, and for every theory of the DP principle that isn't DPP$_{BAB}$ there will exist examples that others cover successfully that they can't.

Second, because you should not read the principles in these theories as if they had universal quantifiers in front. None of the features they select out as characteristic of democracy are strong enough to compel peaceful resolutions. Recall our discussion of process theories of change in our Section '*Tackling the job of prediction directly* '. There we argue that a causal process can only carry through if all of the requisite support factors are in place at each step and no interferences strong enough to halt the process occur. For many DP principle theories, the tangles behind them look strong enough to support the claim that they are conducive to peace. But none has enough to expect that the support factors will always be there when needed nor to expect shields strong enough to prevent interferences from ever stopping that outcome.

Because of this, we urge that the different causal principles that the theories endorse are best seen as generics or *ceteris paribus* principles, which pick out what we call 'Elster mechanisms', after Jon Elster [39]. In the case of the DP, these mechanisms will generally be social, political, cultural, and psychological dispositions. They may need triggering. And even when properly triggered, they do not compel the associated outcomes but are only conducive to them via what are usually long causal chains. If other mechanisms are simultaneously at work pushing in other directions, their effects can be diminished, distorted, or totally overwhelmed.

This is why the principles that describe them should generally be expressed as generics or have the label *ceteris paribus* affixed at their start. John Stuart Mill (1836, ed. 1967) [87] calls them 'tendency' principles because they tell not what effect a feature will produce but what effect it tends to produce.

Let us take a moment to look at one of these proposed theories — Kant's perpetual peace — as an illustration (Figure 3.3)

Here we have a (rough) process theory of change for Kant's perpetual peace. Let's focus on the connection between 'Citizens do not want to put themselves through the horrors of war' and 'Citizens are against war'. Under ideal circumstances for Kant this connection

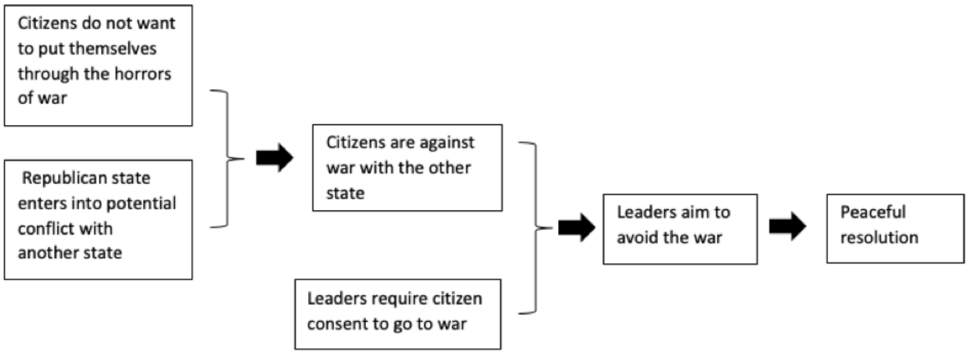

Citizens do not want to put themselves through the horrors of war
Republican state enters into potential conflict with another state
Citizens are against war with the other state
Leaders require citizen consent to go to war
Leaders aim to avoid the war
Peaceful resolution

Figure 3.3: A rudimentary theory of change for Kant's perpetual peace.

holds, but as a principle it relies on a specific connection: war means citizens go through horrors. One can imagine, however, numerous scenarios where the citizenry do not expect this connection to hold. It might be that the war would be fought overseas and with little impact on the average citizen (as, for example, with many of America's modern wars in the Middle East).

The connection holds *ceteris paribus*. This is also true of other steps in the perpetual peace process theory of change. Leaders may have overriding reasons to go to war without their citizens' consent (perhaps they have information the general public do not, or it may be the case that the leaders are not up for election for several years, in which case they believe they can repair their reputation in the intervening time), or citizens may give permission to go to war even knowing that there will be horrors visited upon them (perhaps because of overruling nationalistic tendencies in citizens who are willing to bear the burden in the name of pride or resources). The perpetual peace is not an absolute law, but a tendency in those states with the republican structures picked out by Kant's theory. Naturally, these tendencies may be overruled, or specific necessary aspects may fail to trigger given certain circumstances.

All DP principle theories will have this property of employing *ceteris paribus* or generic principles. Each invokes a tendency of those democratic systems they are formulated over which can be overruled or otherwise derailed by other factors. There's no reason to think that any of them holds in all situations or is indefeasible.

Because of this, each explanatory mechanism can be countered or otherwise fail to trigger. But, if as we claim, the DP principle is formulated across democracies-by-anybody's-book and war-by-anybody's-book then all the proposed theories will include democracies-by-anybody's-book and war-by-anybody's-book in their predictions. So in each case the mechanisms cited by every theory will simultaneously be viable means by which the DPP$_{BAB}$ can hold . DPP$_{BAB}$ would fail if all the theories turn out to be completely wrong despite the virtuous-seeming tangles that back them up, and/or all the mechanisms cited by them that are genuinely conducive to peace are disrupted or fail to trigger simultaneously. This is the coincidence that we think is unlikely. There's been a vast amount of diligent work on the DP over the last 20 years, creating a great tangle of tangles with long tails into other successful domains. This doesn't guarantee reliability, as we have noted (and will see in some depth in the Afterword). But it is good grounds for expecting it (at least from science's point of view). Thus, we argue, you can rely on DPP$_{BAB}$ to make accurate predictions. Not because any particular one theory is the correct one, but because it is overdetermined that at least one mechanism will be in place and work as predicted to prevent democracies-by-anybody's-book dyads from going to war-by-anybody's-book.

3.7 Conclusion

If what we have argued in this chapter is true, then the broad projects pursued by most DP theorists of searching for a single, true explanation is misguided. Instead of the theories being at odds with each other, the principle is reliable for bog-standard democracies and bog-standard wars precisely because there are so many plausible theories at play.

This is in contrast to other examples where multiple theories about the same phenomenon are genuinely in competition, such as the alternative mechanisms for high-temperature superconductivity discussed in the Preface and the gravity wave case we discuss in the Afterword, where at most one theory is expected to ultimately win out. If only one or a couple of the mechanisms cited by the different theories were genuinely conducive to peaceful resolution, you should expect exceptions to be more common even for bog-standard cases since so much else needs to be in place for the mechanism to produce the outcome it pushes towards. The DP principle would be far less reliable.

This result is due to the types of constraints that are at play in the tangles of the DP theories, in particular, to the tension between divergence constraints, which arise due to the fuzzy, Ballung nature of the concepts the principle is predicated over, and convergence constraints in the form of responsiveness to field standards and inter-theory critique. Reliability is secured by the fact that the differing theories are forced to agree on a single, overarching principle - the democratic peace principle by anybody's book.

This chapter is forthcoming in the book *The Tangle of Science: Reliability Beyond*

Method, Rigour, and Objectivity by Cartwright, Hardie, Montuschi, Soleiman, and Thresher, to be published by Oxford University Press.

Chapter 4

How Research Harms

Abstract

Science can harm us. Science has harmed us. Science is going to harm us again. One of our jobs, as responsible academics, is to try and figure out when and where this harm will come from, and try to limit the damages. In many cases, this involves placing bans on how and when research is used as, for example, with the current moratorium on human germ-line editing. The ban on application, however, doesn't often extend to research. Programmers continue to build the theory behind military AI and killer robots, geo-engineers continue to model the effects of releasing particles into our atmosphere, and immunologists are actively working on the theory behind modifying the flu virus to be more infectious.

This often unspoken assumption that research ought not to be restricted is a problem. It leads scientists and policy-makers to overlook the harms of research when designing and implementing bans and moratoriums, meaning that certain harms aren't being addressed or accounted for in deciding whether to pursue certain areas of science. Here I aim to address

this problem, categorising the harms of research according to whether they are acquisition or context based, and providing clear examples of each of these harms occurring in real scientific projects. I conclude by analysing recent research into ‘gay genes’ arguing that this project caused significant research harm and using my categorisation system to determine the best way to mitigate future harms via an appropriately tailored moratorium.

4.1 Introduction

> “The ignorance of humans involves at least two kinds of uncertainty. There is uncertainty about what will be found; and, there is uncertainty about how it will be used. Our background knowledge of science tells us that knowledge brings the potential good and bad in one piece.” -D. Johnson, *Reframing the Question of Forbidden Knowledge for Modern Science*, (1999), p.7) [67]

The question of when and where we ought to restrict scientific research has a long and rich history. Academic freedom has been the subject of debates, books, and policies, and is seen by many as one of the central tenets of scientific inquiry (CITE). Despite the importance of allowing scientists to pursue research without censorship or strict government control, it is also widely acknowledged that certain areas of research should be regulated for the good of both individuals and society as a whole. These regulations, and the justifications for them, tend to stem from the idea of harm - that science can, has, and will harm us if left unchecked. Consider, for example, the terrifying results of nuclear fission research, or the ongoing debates over the development of genetic techniques to control invasive species. It is seen as one of the jobs of responsible academics and policy-makers to try and figure out when and where this harm will come from, and to limit the damages where possible.

In many cases, this involves placing bans on how and when scientific research is used, as, for example, with the current moratorium on human germ-line editing which bans the implantation of modified eggs, but allows research into the technique to continue. Indeed, there is a broad trend of focusing harm-mitigation in science on the application phase. Programmers continue to build the theory behind military AI and killer robots, geo-engineers continue to model the effects of releasing particles into our atmosphere, and immunologists are actively working on the theory behind modifying the flu virus to be more infectious. In all these cases it is implementation which is seen as harmful, and this, paired with a belief in the benefits of free research, means that most control mechanisms shy away from banning research itself.

Of course, there are exceptions to this - we ban experiments predicated on harming test subjects for example, and build ethics boards explicitly to guard against these kinds of harms. In general, however, there is a (largely) unspoken bias in scientific harm mitigation techniques that means research is left alone, while governments and scientists focus on controlling its application.

Here I argue that this emphasis on controlling application, while allowing research, has lead to a potentially dangerous neglect of the harms that research itself can cause. Indeed, there are some circumstances where the act of gathering more knowledge should be banned, or at the very least carefully monitored, because attempts to prevent these harms at the application level will prove to be too late. In this I join researchers like Kitcher and Pamuk, who have similarly argued that there are cases where research can be the cause of harm, independent of its application. My work encompasses and goes beyond theirs, however, aiming to categorise and demonstrate the numerous possible harms of scientific research, and giving clear examples *in situ* of these harms occurring.

By categorising these harms we are, in turn, better placed to identify effective controls. Permanent bans are, for example, appropriate when the harms themselves are inherent to the research or act of doing research. In contrast, some harms are the result of social context, and can be lifted or softened once these contexts change.

Section 2 will define what I mean by harm in this paper and section 3 will demonstrate the bias towards controlling application, rather than research, in current scientific policy. In section 4 I will provide a categorisation system for what I take to be the most common harms of scientific research, and finally in section 5 I will put these categories to use in a brief case study - research into the 'gay gene' - which I take to be an example of research which does enough harm to warrant a temporary moratorium.

This paper is aimed at scientists and policy makers. As such, it is designed to be fairly pragmatic, providing clear language and advice on what types of harms ought to be factored in when considering pursuing or banning research. It is also the case that while sometimes the harms discussed here will be of the type that warrant bans and moratoriums, others times the harms will be outweighed by the benefits. In both cases, however, scientists and policy-makers ought to be aware that these harms exist, and where possible should work to minimise them.

4.2 A Focus on Harm

The existence of harms, either actual or potential, underpins almost all calls for controlling science. We care about stopping science from initiating pandemics, irrevocably altering the environment, changing the genetics of the whole human race, accidentally wiping out species, increasing the deadliness of war, and doing experiments on non-consenting subjects. Here, for example, is Pamuk (2021),

> “Since preventing harm is one of the main purposes of regulation, restricting certain kinds of actions can be justified on the grounds that doing so would be the most effective way to prevent foreseeable harms to others.” [108, p.7]

Here, too, is Sillgoe, Owen, and Macnaghten (2020),

> “Responsibility in governance has historically been concerned with the ‘products’ of science and innovation, particularly impacts that are later found to be unacceptable or harmful to society or the environment.” [133, p.1]

and the US Government,

> “The goal of oversight is to preserve the benefits of life sciences research while minimizing the risk that knowledge, information, products, or technologies generated by such research could be used in a manner that results in harm.” [105]

It is clear, then, that in order to justify a broader acceptance of restricting research we ought to get clear about what exactly the harms of research might be. Indeed, articulating these harms and what they look like *in situ* is going to be critical to any conversation about where and when we ought to restrict science.

Given this, we will need a working definition of harm. Here, I’m going to use the term rather loosely. Harm, for this paper, will include any negative impact the research has on individuals or society. In this, I have in mind a broadly comparative account of harm; “that to suffer harm is to be put into a certain sort of comparatively bad state - a state that is worse for one than some relevant alternative state.” [55, p.1] That said, I don’t think any of the categories of research harm we discuss below will hinge too strongly on taking this particular approach. Here, harms will include those states which are physically instantiated as damages as well as those which simply increase the risk of damages. Thus, one might be harmed by physically being in a car accident, but also by having someone tamper with your breaks such that your chance of being in an accident increases. You might also be harmed if social or legal structures are altered in ways that will give you worse outcomes if you need to rely on them (the law now require the death penalty for anyone found at fault for a car accident, or anyone who deliberately hits you with their car now has no legal or social obligation to check if you're okay or help pay for injuries you sustain). Obviously the inclusion of some of these under the category of 'harm' is controversial [114], but for the sake of this paper we need a practical and pragmatic understanding of what types of negative impacts research might have on individuals and society. Thus, while there is a rich debate to be had about the status of things like 'potential harms' it seems obvious to me that if science creates a more racist society, even if that racism hasn't yet had a chance to directly harm an individual, a harm has occurred.

4.3 The Separation of Knowledge and Use

Most calls for restrictions on scientific work place themselves between research and application. That is, we allow research into a topic, but aim to prevent the subsequent information from being misused when it is taken out of a purely academic context and brought into contact with wider society. There are a number of reasons given for placing restrictions on use rather than research, ranging from the importance of academic freedom [149, 52], to the uncertainty of where future discoveries will come from [49, 108], to the idea that more research helps inform the safe use of new technologies and techniques, to a belief that knowledge and research are inherently valuable [84, 78]. This distinction between research and technology, knowledge and use, can be found throughout philosophy, science and policy both implicitly and explicitly.

Let us begin with the explicit.

Here is Gärdenfors in his 1989 paper *Is There Anything We Should Not Want to Know?*;

> ”The upshot is that controlling fundamental research is not only unwanted but well nigh impossible because it presumes that we can foresee major breakthroughs. In the cases where we suspect that research will lead to dangerous technologies, we should try instead to control the technology. This may be extremely difficult, but it is not impossible” [49, p.8]

And Pera in *Should science be supervised, and if so by whom?* (1989),

> “The justification for the differentiated treatment view is usually said to lie in the fact that science is good in itself, since it answers purely intellectual needs, while technology is merely an instrumental good that fulfills practical needs and aims. Only when such needs and aims are legitimate and morally right should technology be promoted.” [111, p.58]

Thomas (1977) in *Notes of a Biology-Watcher, the Hazards of Science*,

> “We should be very careful with that word hubris, and make sure it is not used when not warranted. There is a great danger in applying it to the search for knowledge. The application of knowledge is another matter, and there is hubris in plenty in our technology, but I do not believe that looking for new information about nature, at whatever level, can possibly be called unnatural.” [p.326][135]

Agazzi (1989), as cited in Johnson’s *Forbidden Knowledge and Science as Professional Activity* (1996),

> “To put it briefly: while it is in principle morally acceptable to know every thing, and there are no *morally prohibited truths*, not everything that can be *done* is acceptable, and there are *morally prohibited actions*” [4, p.206]

Johnson goes on to note that,

> “Implicit in Agazzi’s view is a distinction between knowing and doing. Science is equated with knowing and knowing is understood to be passive. Since knowing is passive, it need not and should not be controlled. On this view, only when a step is taken beyond science do we have something powerful, something that can affect the world and be dangerous. The step is from knowing to doing, from knowledge to using knowledge.” [66, p.207]

And Nelkin (1978) in *Threats and promises: Negotiating the control of research*,

> The scientific community has persistently resisted public control. Only when research has direct technological applications are scientists willing to concede the need for regulation. [97, p.191]

Here is a more solid example of this common line of reasoning being used in science. Take human germ-line editing, a technique surrounded by controversy even before He Jiankui proceeded to create two children with heritable genetic modifications in 2018.

CRISPR/Cas9 developments now allow us to modify human germline genetic traits with relative ease.[1] There are numerous obvious dangers here and there have been explicit calls to place moratoriums on the technology based on a number of serious worries [44]. Take, for example, concerns about scientist’s right to effectively alter the entire human race. There are also questions over who will have access to this technology, with legitimate concerns that it may end up creating an even greater divide between the wealthy and poor. Further, we might worry that it would lead to even greater genetic elitism and racism, and that it might be used to try and eliminate ’unwanted’ genetic strains, including historically disfavoured traits like homosexuality, physical disabilities like deafness, blindness or Down’s syndrome, certain appearances like darker skin tones or facial features, and other traits associated with oppressed groups. In all of these cases we don’t need to go far to find people who have

[1] Well, depending on who you ask and what you want to do exactly.

attempted to eliminate certain genes (and the carriers of them) altogether. Imagine if they had the ability to actually do so at the genetic level in the name of making humanity 'better'.

On the other side of the equation are significant benefits to be had from the technology - it has been touted as a silver bullet for curing currently untreatable genetic diseases, and could be used to imbue genetic immunity to other diseases into the population. All this means that while there is a lot of justified caution about actually putting germ-line editing into practice, scientists are still keen to keep research going in the hopes of generating better outcomes as the techniques mature and our knowledge-base grows. We can see this explicitly in the 2019 call for a moratorium on germ-line editing:

> "To be clear, our proposed moratorium does not apply to germline editing for research uses, provided that these studies do not involve the transfer of an embryo to a person's uterus." [74]

More information is seen as beneficial in helping us navigate not just the social and moral worries, but ensuring that we minimise the technical risks, as well as helping get other options on the table that may not be as extreme.

The pattern of this case is visible in numerous other areas of study, lending credence to the existence of an implicit distinction between research and use in science and policy.

Take, for example, Gain-of-function experiments, which increase the virility of diseases in the lab and are criticised because of the risk of escape or the potential use of the techniques for bio-weaponry. Continued research into them, however, is seen as vital for understanding how diseases spread, and how to handle new variants;

> "Researchers defending gain-of-function work challenged Andrew Hebbeler, assistant director for biological and chemical threats at the White House Office of Science and Technology Policy, on whether the government had considered the public-health impact of the moratorium and the research, surveillance and drug development that would be lost." [118]

Similarly, nuclear physics research is protected, but the active development of nuclear bombs is banned. From the text of the 2021 Treaty on the Prohibition of Nuclear Weapons;

> “[N]othing in this Treaty shall be interpreted as affecting the inalienable right of its States Parties to develop research, production and use of nuclear energy for peaceful purposes without discrimination” [5]

This emphasis on controlling application rather than research is further codified in the wording of most major moratoriums. Indeed, moratoriums are often used as a tool to allow research to catch up with the dangers of uninformed implementation. As such one can find numerous calls for moratoriums that specifically allow for research on a controversial topic to continue, but prohibit application or the development of new technologies with it. In addition to the ones for germ-line editing, gain-of-function disease research, and nuclear development above, here are examples from geoengineering,

> “An International Non-Use Agreement on Solar Geoengineering would not prohibit atmospheric or climate research as such, and it would not place broad limitations on academic freedom. The agreement would instead focus solely on a specific set of measures targeted purely at restricting the development of solar geoengineering technologies under the jurisdiction of the parties to the agreement.” [74]

biometric surveillance,

> “There is now a significant body of evidence that illuminates both the potential benefits and harms of biometric technologies in different contexts. [...] [T]his research provides an invaluable resource, allowing for proactive actions that anticipate and mitigate known harms.” [21]

and lethal autonomous weapons,

> “In fact, the ban would apply to development only of fully autonomous weapons, that is, machines that could select and fire on targets without meaningful human control. Research and development activities would be banned if they were directed at technology that could be used exclusively for fully autonomous weapons or that was explicitly intended for use in such weapons. A prohibition on the

> development of fully autonomous weapons would in no way impede development of non-weaponized fully autonomous robotics technology, which can have many positive, non-military applications." [32]

Conversely, moratoriums on research, rather than implementation, are exceedingly rare. This is not, however, to say that research moratoriums don't exist. Often these are considered or created for what is known as 'dual use' research. That is, in the words of the National Institute for Health "research which can be reasonably anticipated to provide knowledge, information, products, or technologies that could be directly misapplied to pose a significant threat with broad potential consequences" [96] . The term is most often used for biological research but is also applied to other areas of science. Dual use conversations often involve weighing up the benefits and potential dangers of research once, as the name suggests, it is applied in real world contexts. If the damages are too great, it is then seen as reasonable to limit, prevent, or censure the research in some way.

In 2005, for example, Wein sought to publish an article detailing what a bioterrorist attack on US milk supplies might look like [146]. The government, concerned that it could act as a blueprint for an attack, stepped in and asked for the paper to be delayed or withdrawn from publication. Again, we see a focus on use, however. Where the research is banned or controlled it is because of how it might be applied, and in doing so we seem to miss something critical about what sort of harms ought to factor into our equations. Nature published a comment on the Wein controversy, stating that,

> "It is important to develop clear guidelines about what research is considered sensitive, what is expected of researchers whose work produces dual-use outcomes, and how the government should in practice respond without losing the priceless virtues of open scientific scrutiny." [1]

Alternatively, we sometimes see bans like the 2014 one which put a halt to gain-of-function virus research,

> "Experiments that create the possibility of initiating a pandemic should be subject to a rigorous quantitative risk assessment and a search for safer alternatives before they are approved or performed. Yet a rigorous and transparent risk assessment process for this work has not yet been established. This is why we support the recently announced moratorium on funding new "gain-of-function" (GOF) experiments that enhance mammalian transmissibility or virulence in severe acute respiratory syndrome (SARS), Middle East respiratory syndrome (MERS), and influenza viruses." [80, p.1]

Again, however, the call above for a halt to research that aims to create 'novel potential pandemic pathogens' was explicitly proposed because there were concerns about the safety of such diseases existing, even in labs for research purposes. That is, the worry is not that the research is harmful in itself, but that it might accidentally come into contact with the outside world in negative ways.

It is clear, then, that there is a broad pattern in scientific policy and philosophy of taking the harms of science to stem primarily from application. And that most attempts to mitigate these harms focus on interventions that sit after research, but before implementation. As part of this, moratoriums and bans are almost always written to explicitly allow ongoing research, with the underlying assumption that, where feasible, research as an endeavour is beneficial.

This is a problem. Even a weak belief that the harms of science can or should be mitigated via implementation control means that scientists and policy makers are failing to recognise or prevent a number of serious harms - ones which occur at the research phase itself. Indeed, the harms of research is a surprisingly underexplored are of science within philosophy, and one which needs more recognition should we wish to have informed and successful scientific policy. Nowhere as far as I can tell, has anyone sat down and collated these types of harms into a clear list, articulating the most common of these harms and how they instantiate in science.

Such a list is important to have however. It is, after all, only by articulating exactly what harms are occurring, and why they instantiate, that we can begin to formulate appropriate responses and controls. Here, then, I lay out the beginning of such a categorisation. In doing so I hope to give scientists, policy-makers, and philosophers the tools to more effectively mitigate these harms as they happen.

The exploration of these harms, and how they come about, will form the remainder of this paper. Some of these harms will stem from the very act of doing research, while others occur because of the broader social context within which the research is performed. In both cases, however, they are of a kind where attempts to prevent the harm by controlling the uses of the research is doomed to failure. Application control, in all these cases, will have already missed the boat.

4.4 An Incomplete Categorisation of Research Harms

Here I identify the following ways that scientific research causes harms which cannot be prevented by controlling its use;

1. Harmful research practices

2. Inherently harmful knowledge

3. Research that diverts resources and attention away from other beneficial actions

4. Moral hazards

5. Psychologically harmful research
6. Socially harmful research
7. Significantly increased risk of harmful application

This list is incomplete. There will be types of harms not considered here, and, as with all science, there will be difficult fringe cases that don't seem to cleanly fit into any one category. What I hope to do here, however, is offer the largest ones which are relevant to the way we talk about, control, and evaluate the impacts of scientific research.

Thus, while there could be a scenario where someone threatens to kill an innocent bystander if a particular line of research is pursued, and this would indeed be a harm of the research, we'll consider that too narrow to include here.

We can broadly categorise the above list into two subgroups - harms that come from the scientific research itself (what I will call acquisition harms), and harms which come from its position within broader society (context harms). (3, 4, 5, 6, and 7) all fall into the latter category. (1, and 2) fall into the former.

The divide I'm drawing here isn't a hard one, harms can be both acquisition and context, and migrate between the two depending on a diverse number of factors. Even so, it is useful to have the language to talk about the difference between harms that are inherent to the act of research or acquisition of knowledge itself, and those which occur because of the social or political context within which it is being done.

It is also worth noting that context harms are, in many ways, related to the reasoning for standard moratoriums on application. That is, they arise because the research is coming into contact with the real world. Here, however, I take the harms to be of the kind that warrant controlling the research itself, because placing bans on application would be ineffective specifically because once the research is done the harms have already occurred, or are almost certain to occur, and as such if we simply try to intervene on its use once the research has been completed we've missed our opportunity.

Let us look, then, at each of these categories in turn, defining them more clearly and seeing how they instantiate in some real-world cases. This last bit is particularly important for two reasons. First, the kinds of harms we are discussing here are relatively loosely defined and, as such, it will be helpful to see examples. Definitions are, unfortunately, often not enough in this kind of scenario to give us a good idea of what these kinds of problems might actually look like *in situ*. Second, science is, as we all know, a relatively messy endeavour - no two cases of the same type of harm are going to look the same. As such, seeing clear examples can help us more easily identify similarities to other cases. As with many things, the application of these categories will involve careful consideration within a specific context. All I can give is a useful guide.

4.4.1 Acquisition Harms

Acquisition harms are inherent to the research or knowledge itself. That is, they cannot be separated out and will occur under any circumstances where the research is performed. As such, these harms are likely candidates for permanent bans, rather than temporary moratoriums.

Harmful Research Practices

> "Those for whom the advancement of knowledge is a supreme value might believe that, in basic research as distinct from applied science and technology, no subject

> should be declared off limits. Yet there are clear inhibitions on some kinds of research involving human beings, and indeed animals." [38]

It is, perhaps, unfair to say that restrictions are rare for research. In fact, they are quite common within a single category - where injury or damage is inescapably done to morally relevant agents in the act of performing research. This type of acquisition harm stems from the act of conducting research itself and is (broadly) why things like ethics boards exist, and why we ban experimenting on non-consenting humans. Indeed, it is probably the best regulated of the harms we'll discuss here, with numerous international restrictions and agreements in place including the The Belmont Report [95], and the Neuremberg Code [134]. It is worth noting that these are generally not temporary restrictions but permanent bans due to the inherently harmful nature of the research involved.

Research harms of this kind are those where the very act of collecting data does harm. Within this category we can recognise a number of sub-divisions the most salient of them being information derived from the torture or exploitation of human subjects. Knowledge of how long humans can survive in cold water is directly derived from unethical research performed by the Nazis on war prisoners, for example [134]. It also covers cases like the Stanford prison experiment, Milgram experiment and Tuskegee Syphilis study, and is one of the primary arguments used against STEM Cell work, where critics argue that the act of gathering cells causes harm to human agents (albeit their inclusion of STEM cells as part of this category is controversial). Other regulated research in this category includes experiments on human subjects who cannot or do not consent, experiments that require hurting certain types of animals that humans consider to be of sufficient moral standing, and experiments that involve doing significant damage to the environment (think, for example, of the lingering effects of dropping an experimental atomic bomb at Bikini Atoll).

Geoengineering research is a particularly interesting example of this latter category. We will discuss the version of this technique which aims at solar reflection in depth in section 4.4.2 but an alternative form - iron seeding - involves spreading iron in the ocean to encourage algae blooms which, in turn, capture and sequester carbon from the atmosphere. Research into the viability of this technique for reducing climate change is, however, almost impossible without actually releasing significant amounts of iron into the ocean to see what happens - there are no small-scale experiments which can be done in the lab which would accurately reflect what we would expect to see *in situ*. Thus, the act of doing research is almost indistinguishable from deploying it and, given one of the goals of the research is to figure out what sort of disruptions a large-scale algae bloom cause for deep-sea ecosystems, we cannot do the research without, in turn, causing some of the harms we were doing the research to try and avoid.

There are other, lesser, versions of this harm. Research that involves dissecting mice harms the mice. As does tagging seals, removing samples from the environment, and building laboratories. All of these involve some harm (punching holes in seal ears, taking resources from the land, removing wildlife habitats to construct buildings), but we generally judge these harms to be of sufficiently low stakes that the benefits are acceptable. Alternatively, we might think that research such as that done by NASA during the space race, which involved significant risk to the pilots of the experimental craft and cost the lives of a number of them [99], caused acceptable harm because the pilots themselves consented to the program. Still, however, harm was done.

One thing worth noting here is that we should be careful to distinguish research that *causes* harm, from research which *stems from* harm. Not all research on harmful things is, itself, harmful. Differences in socio-economic prosperity between East and West Berlin, for example, certainly involves harmed parties - one side was less well-off than the other - but it is not the research itself which caused these harms. Similarly neuroscientists looking at the brains of individuals with unique or interesting mental damage or impairments are only causing morally significant harm if they themselves have damaged the patients, as opposed to accidents which happened over the course of the subject's normal lives. Because of these grey areas debate over banning harmful research practices tends to centre on whether there is *morally significant harm* happening, rather than whether we should ban this particular type of harm in general.

Inherently Harmful Knowledge

The second acquisition harm I categorise here is research which harm us is by generating knowledge that is inherently harmful. This is a little more difficult to give clear examples for, partially because I suspect there's something morally wrong about sharing such information with other people. I do not, after all, want to harm the readers of this paper. While I am (unfortunately) aware of a number of examples then, I'm going to create a few (more or less) fictional examples that I hope convey the gist of this category.

Let's start with a realistic but fairly innocuous example. Let's say that a group of scientists do some research and come to the conclusion that people who are nervous will do worse in job interviews. It also turns out that knowing this fact will, in turn, make you more likely to be nervous during your interview. This research, then, has done harm to everyone who hears it who will go on to do a job interview.

Here's a slightly more serious example; stereotype threat occurs when a group is primed to conceptualise themselves in terms of a negative stereotype about the group (eg. women are bad at mathematics). This, in turn, leads to anxiety which makes the group perform worse at related tasks (ie. women taking a mathematics exam). This can often lead to vicious cycles, as members of the stereotyped group become more convinced of their inability to perform the task, and see others of the group doing worse because of its effects, further entrenching the stereotype. This effect occurs regardless of whether the member of the group believes in the stereotype.

Plausibly this goes a step further. It isn't hard to imagine that that learning about the existence of stereotype threat might be enough for a member of a group to either consciously or subconsciously prime themselves to do worse on the relevant tasks. That is, as a woman who knows the stereotype that women are bad at mathematics, even if I believe I am good at maths, and have past evidence to back it up, there is a non-zero chance that before performing any mathematics task I will remember the existence of stereotype threat. This, in turn, causes me to think about the negative mathematics stereotype, which will cause me to do worse on the test. I am, then, harmed by the knowledge that stereotype threat exists. Indeed, if what I have said is true, then we have very good reason not to disseminate information about stereotype threat to vulnerable groups, precisely because it increases the prevalence of the problem, and leads to the broader reinforcement of the stereotype.

Moreover, it is worth noting that these types of harms are almost impossible to avoid because the scientists themselves, in learning this information, are harmed. That is, this research harm is caused not just by the dissemination of information, but by the very act of finding it out in the first place.

4.4.2 Context Harms

Context harms occur because of the context within which the research is done. As such they are contingent on numerous broader factors and are likely candidates for temporary or conditional moratoriums. These controls are able to be removed once the context changes, and the research is able to be pursued with less or no harm done.

Harmful Diversion of Resources and Attention

New scientific knowledge might cause you to abandon, deprioritise, or divert resources and attention away from other actions which might mitigate the problem you're trying to address more efficiently or that might help in tandem. This harm is defined in part by limitations on funding or on other broader social issues to do with things like media coverage and the limited focus of the general public. It also hinges strongly on the position from within which the funding is being criticised - Democrats and Republicans can and do disagree strongly about research priorities and what constitutes a waste of resources. Still, researchers ought to be careful in determining what topics to propose and pursue precisely because of these issues, and, regardless of how context-dependent it is, it is clear that this is a harm that research can and does cause.

Let's look at an interesting historic case: AIDS funding. In 1988 the AIDS crisis was in full swing, and the federal government was preparing to pass the HOPE act aimed at funding HIV/AIDS research, prevention, and testing. At the time there was extensive debate about how much funding ought to be put towards the disease, and it was openly acknowledged that what funding was given would likely be drawn from other areas of federal disease research grants.

> "Senator Edward Kennedy has publicly expressed concern that the necessary expansion of the AIDS budged might be financed - directly or indirectly - by slower growth or reductions in other health research programs. Although many citizens might prefer that expanded funding of AIDS-related research come from a reduction of nuclear armament or an increase in taxes, the political reality is that trade-offs are more likely to be made with other health research programs." [57, p.1]

One of the serious questions facing policy-makers and scientists was how to weigh up the costs and benefits of particular funding models. Inevitably money taken from cancer research would cause deaths, but so too would failure to fund AIDS research adequately.

Along similar lines, there is ongoing debate about the best focus for AIDS research itself. It is entirely possible that building immunity to HIV/AIDS using the modification of human genomes will prove to be overwhelmingly successful, but it is also a long-shot [75]. Conversely, it may be that more traditional vaccines are more promising, and what funding exists ought to be channelled in that direction. There are also questions of what information is needed to even make these decisions, and how to go about getting them. Debate over HIV/AIDS research priorities are a prominent aspect of the field [107, 7] and there are difficult decisions to be made - do you channel your resources into finding a vaccine as fast as possible, or do you work to try and prevent as many infections as possible? What groups do you focus your research on? Which countries are results rolled out in first?

This is perhaps one of the most difficult harms to navigate, for good reason; it requires difficult decisions in the face of uncertainty, and in the case of medical research it is inevitable that some decisions will lead to worse health-outcomes or even death for numerous people.

Of course, this isn't limited to medicine. There are limited resources available to fund research, and governments, universities, and public grant givers are continually making decisions about how to allocate not just these funds, but also how to focus public attention on the most important of them. Do we fund better solar panels, or more efficient nuclear power plants? Should researchers prioritise building better climate models, or figuring out better ways to communicate currently existing ones to the public and policy-makers? Do we fund malaria vaccines, or buying mosquito nets? The former could save millions of lives, but insectiside-treated mosquito nets have shown to significantly reduce the transmission of the disease [102], such that a reduction in funding would directly kill people, and an increase would directly save lives.

Of course in an ideal world this wouldn't be a problem, but given the limitations within which scientists work it is clear that pursuing certain areas of research is a cause of harms purely because it reduces the focus and funding of other areas of study. As such, scientists and funding bodies routinely grappling with this particular harm and how to minimise it.

Harmful Moral Hazards

A 'moral hazard' is a term borrowed from the insurance industry. In broad strokes research creates moral hazards when it reduces the incentive to prevent or minimise a damage or harm. In science this often instantiates as things like techno-optimism, which creates a feeling amongst the general public that we will be able to invent, research, or science our way out of current problems, and thus have no need to take any other steps outside of science to solve them. Unfortunately, because research is just that - research, there is no guarantee that these solutions will ever instantiate or even turn out to be viable.

A number of critiques of climate adaption technologies take this route, hinging on the idea that they are a band-aid, rather than a genuine solution, to the underlying problem of human emissions. Solar Radiation Management (SRM) Geoengineering - changing our atmosphere to reflect more sunlight - is a good example of this debate in the literature.

The idea behind SRM is that by releasing tiny particles into the stratosphere we can increase the amount of sunlight reflected by the atmosphere, thereby cooling the temperature of the planet. Advocates of this technology point to the relative ease and cheap cost of such techniques, as well as the fact that it could buy us time to deal with greenhouse gases without a radical change in global temperatures. Critics point to a number of downsides, including the difficulty of garnering international agreement to use it, a lack of understanding of the long-term effects of the technology, and the fact that it would 'lock us in' to the SRM path as we would need to continue releasing particles over an extended period of time to avoid a sudden rapid and catastrophic change in temperatures if the releases were stopped [86]. The technology is, of course, far from ready, but there has been an uptick in research interest over recent years and it has become the focus of a number of philosophy papers debating it's pros and cons. [11, 50, 130, 151, 61]

One major worry for the technology, however, is the Moral Hazard one - that publicising this research, or even simply having it available, will slow down or even prevent action on climate change.

> "[G]eoengineering could be inaccurately perceived as a comprehensive insurance policy against climate change. This misperception could create various incentives that would exacerbate the problems that geoengineering is intended to ameliorate. Individuals might curb voluntary efforts to reduce carbon emissions. Fossil fuel consumption and other GHG-generating behaviors might even increase out of a misguided belief that climate change no longer poses a threat. Societies might divert resources away from mitigation toward geoengineering schemes that ultimately prove futile or unworkable. Finally, political and financial support for mitigation and adaptation policies might decline." [79, p.678]

Exacerbating this is the worry that not only might this technology inadvertently undermine attempts to reduce emissions, but it might be actively weaponised by large-scale polluters to shift focus away from their responsibilities.

> "In light of their history, capacity, and fundamental commercial interests, it should come as little surprise that fossil fuel companies have been among the most active and sustained players in the geoengineering space. [...] The fossil fuel industry controls huge swaths of the technologies necessary to pursue CDR and SRM at scale. These companies have been involved in geoengineering research and debates from their earliest days and are not separate from — but rather inextricably linked to - any real-world execution of geoengineering." [92, p.10]

Companies like Exxon, Shell, and Boeing have all expressed support for Geo-engineering, funding research and advocating for the technology [92]. There are concerns, then, that these companies could push a narrative where SRM, or other geoengineering techniques, are an easy and comfortable solution to climate change that wouldn't require major changes to the way we consume resources.

Geoengineering, then, is a clear case of research creating a moral hazard. Of course, the implications of this on the continuation of research into the technology is less clear-cut. Some scientists argue that research is beneficial, despite the moral hazard aspects,

> "Some critics also argue that pursing [Geoengineering] research at any substantial level will reduce society's resolve to reduce emissions of greenhouse gases. But evidence from other cases suggests that this is unlikely. For example, research that resulted in the development of seatbelts and airbags in cars has still provided an immense benefit even if it, in a small way, influenced people to drive faster and more recklessly. Moreover, if [Geoengineering] proves to be unworkable or to pose unacceptable environmental risks, the sooner scientists know this, the faster they can take these options off the table. Indeed, if [Geoengineering] approaches are not subjected to serious research and risk assessment, [Geoengineering] might incorrectly come to be regarded as a safety net. The stakes are simply too high for us to think that ignorance is a good policy." [15]

Much has been said, and much more will be written, on how and where to draw the line on these kinds of harms, and certainly it isn't restricted to geoengineering, but to almost any technique which proposes that we adapt to, rather than mitigate, climate change.

More generally, moral hazards turn up wherever we might worry about 'technology-fixes' or 'science-fixes' coming into play. Why recycle when there are groups working to use bacteria to break down plastics so they're biodegradable? Why conserve energy when green-power is becoming more prevalent? Why enact international policy to help food shortages in third-world countries when genetically modified super-crops are being developed to be used in these places? This kind of thinking is obviously harmful where it guides our actions on the broad-scale, and where we then continually put-off making more effective but perhaps less pleasant, or more personally costly, changes in favour of hypothetical, or even real, results from science.

As with all the categories here, whether a moral hazard counts as serious enough to warrant restricting research will vary with context - some research, like geoengineering, seems to pose enough of a harm to broader attempts to mitigate and take control of climate change, that we ought to halt any research in that direction. Others, like research into artificially breeding more effective food sources, seem to have benefits that outweigh any moral hazard harms they generate.

Psychologically Harmful Research

Psychologically harmful research is research which causes emotional damage to those who learn the contents of it, whether it's the researchers themselves or the broader public. One of the most salient examples of this stems from one of the most pressing problems of our lifetimes: climate change. In particular, there is a recent wave of articles and think-pieces on the phenomenon of 'climate despair' [140, 23, 13]. Defined, roughly, as "a sense that climate change is an unstoppable force that will render humanity extinct and renders life in the meantime futile.", it is also sometimes called 'eco-nihilism', 'human futilitarianism' or 'klimatångest' if you're in Sweden [110].

The existence of such feelings should be unsurprising;

> "Few Americans are confident that humans will reduce global warming. About half (49 percent) say humans could reduce global warming, but it's unclear at this point whether we will do what is necessary, and about one in five (22 percent) say we won't reduce global warming because people are unwilling to change their behavior. Only 6 percent say humans can and will successfully reduce global warming." [46]

This despair is a direct result of learning new information about the state of our planet, and the problems that human-caused climate change is causing. Indeed, it grows worse the more up-to-date one is about the situation, and the more information one takes in. Individuals have reported suicidal thoughts, a hesitancy to have children out of fear for their future, and a desire to cut themselves off from learning new information. [110]

Climate science research is, then, having a direct and terrible toll on people's psyches. It is causing harm to individuals, many of whom would arguably be psychologically better-off not knowing information about how the climate is changing. This is, unfortunately, an inescapable aspect of research into a world-threatening phenomenon.

More personally, in doing my own due diligence in preparing to enter the academic job market, I have been researching negotiation tactics. As part of this, I have been presented with an overwhelming amount of research that is personally psychologically harmful. Take, for example, the fact that women who negotiate their salaries are generally seen negatively by the companies who hire them, regardless of whether they get the pay increase. This is in contrast to men, who are praised and admired for their confidence and understanding of their self-worth for the same action. Similarly, women are generally paid less than men, and are more likely to be successful in negotiating if they ’smile’ more. All of these difficulties get worse because I’m also gay. [141]

All of this is incredibly depressing. In many ways, I wish I didn’t know it. In fact, despite being aware that women are discriminated against in the workplace, having clear numbers and multiple studies presented makes it much more salient, and can lead women (and myself) to feel like they’re facing an insurmountable problem. [141] This research has harmed me.

Now, obviously it has also helped me. Being aware of these numbers, knowing what one is up against, and seeing real-world studies that clearly lay out the lay of the land all help one modify tactics and try to work around the problem. It is hard, after all, to solve an ill-defined problem. Similarly, the solution to the problem of climate despair is not to stop doing climate research, but to figure out better ways to help people work through the weight of such knowledge. In these situations the benefits outweigh the harms. This doesn’t, however change the fact that the harms exist.

Socially Harmful Research

Socially harmful research is research which creates further social barriers or difficulties for a given social group or individual. That is, socially harmful research is research which does social harm, rather than just being harmful because of social factors, unlike the other types of harm in this category.

This is the argument that Kitcher makes when he argues that research into the intellectual differences between races, regardless of outcome, harms the socially disadvantaged group [71]. He grounds this claim in a number of factors, firstly that where there is a political asymmetry, studies which tell in favour of the minority group are largely ignored, while studies which find against them are taken up and used to motivate further harms to the group. Second, that because of the epistemic asymmetries involved in being a member of a society with biases, members of the social group are more likely to believe research which confirms their bias, and disregard research which contradicts it. This is, of course, to not even speak of the fact that researchers themselves are members of society and so, while all due caution may be taken, it is certainly the case that these biases have influenced the outcome of studies and will do so again.

So, for example, if one were to try and study the inherent differences in mathematical abilities between Asian-Americans and African-Americans, any study which showed no difference is likely to have less uptake than studies which confirm public bias and show that Asian-Americans are inherently better at mathematics than their African-American counterparts. As such, there is no benefit to either group. Either the status-quo stays the same, or it gets worse.

Here, then, is a context harm caused by research. Unfortunately, research cannot be disambiguated from the society within which it is done. As such, while abstract knowledge may be value-free, it doesn't exist in a vacuum and can be positive or negative relative to the researcher or subject.

Significantly Increased Risk of Harmful Application

This category is aimed at research which, if done, significantly increases or even guarantees that it will be applied in real-world situations. It comes under a number of headings, but is sometimes referred to as ‘slippery slope’, ‘scientific momentum’ or ‘path dependency’ problems, or alternatively being ‘locked in’ [16]. While there has been some discussion of this problem for geoengineering, here we will look at lethal autonomous weapons systems (LAWS) research, which suggests that placing a moratorium on application of the techniques is useless if militaries will simply ignore moratoriums to gain combat advantage. While the research itself doesn’t seem inherently harmful, once it exists there is the worry that it is almost guaranteed that groups will use it unethically.

LAWS (also known as Killer Robots) are defined by the Department of Defence as “A weapon system that, once activated, can select and engage targets without further intervention by a human operator. This includes human-supervised autonomous weapon systems that are designed to allow human operators to override operation of the weapon system, but can select and engage targets without further human input after activation.” [104] There is extensive discussion in political, philosophical, and social circles about the use of these weapons due to the potentially devastating ethical implications of their deployment. In general, there is broad consensus that they will violate the principle of *jus in bellum* by removing the direct responsibility of killing from the hands of agents with moral standing. It has also been suggested that extensive use of LAWS will make it less daunting for countries to enter into conflict, as it reduces the risk to their own soldiers, as well as that it will lead to an increase in preventable war-crimes as militaries will be disincentivised to include the ability to refuse unethical orders as part of the LAWS programming. For these reasons there have been numerous calls in recent years for LAWS to be banned, both from development and deployment. Here, for example, is the Human Rights Watch,

> "Human Rights Watch and [Harvard Law School's International Human Rights Clinic] are calling on governments to:
>
> 1. Work toward an international instrument prohibiting the development, production, and use of the fully autonomous weapons.
> 2. Develop national policies on the issue, which encompass national moratoria on the development, production, and use of the fully autonomous weapons." [143]

There is a clear connection, in the LAWS case, between research and application. Militaries, after all, have strong pressure to have the best weapons at their disposal. Many militaries are also very well-funded, and are already using semi-autonomous weapons. As such, when a tactical advantage like LAWS are put on the table, the chance of deployment significantly decreases. As Ariel Conn says, "Once this Pandora's box is opened, it will be hard to close." [106]

This risk is well known in the literature.

> "Two paradigmatic examples of this category are stratospheric aerosol injection (SAI), which is a radical geoengineering technology that involves reflecting a fraction of incoming sunlight back into space, and killer robots, which are lethal weapons that select and kill targets without human supervision. In both cases, research beyond a certain stage creates the risk of serious and potentially irreversible harm—because of the complexity of climatic processes in the former and the autonomous intelligence of robots in the latter. Both technologies also involve the risk of deployment by rogue private or state actors to intentionally inflict harm, and researchers in both areas have publicly stated that deployment would be a bad idea." [108, p.3]

This is a relatively complex harm of research however. Indeed, arguably it isn't a harm of research at all, but a harm of application - something we're avoiding considering in this paper. That said, it is clear that research cannot be disentangled from application in this case - the very act of doing research significantly increases the risk of future harms to human beings. Moreover, this case is distinguished out from other simple application harms by the fact that any ban on application will be almost impossible to police. Even if major militaries could be convinced to give up the tactical advantage LAWS represent, the research acts as a blueprint for rogue states and actors to create such weapons for their own use. Indeed, it seems difficult to contemplate a scenario where, if the research is done, it is not eventually used, no matter how stringent international controls are.

Regardless of whether application actually occurs, however, the research increases the chance of harm and thus, under our understanding of harms in this paper, is itself doing harm. Indeed, this is a critical harm for researchers to consider as they advance projects - what would happen if someone applied this research. It is for this reason that we ought to be careful when publishing papers which describe how to increase the virility of diseases, build smaller and more powerful nuclear weapons, or modify human genetics for non-health purposes. All of these fall into the category of research which, once available, is unable to be put back in the box and which, as a direct consequence of availability, has significantly increased the chance of serious future harms.

4.5 Case Study: The Genetic Basis for Homosexuality

Having detailed our list, let us look at how these categories might help us determine the right kinds of research controls via a case study.

In 2019 a new study came out that looked at the genetics of almost half a million people [48]. This study, which aimed to identify the genetic markers for homosexuality, concluded that while there was no single ‘gay gene’ it was still the case that certain genetic traits increased the chance that their carrier would be homosexual.

The study was both lauded and criticised for a number of reasons. Advocates argued that it was a significant step forwards to understanding how homosexuality arose in nature and emphasised the fact that it reinforced the fact that queerness was a mix of both nature and nurture,

> “[The study] provides even more evidence that being gay or lesbian is a natural part of human life, a conclusion that has been drawn by researchers and scientists time and again. The identities of LGBTQ people are not up for debate. This new research also reconfirms the long established understanding that there is no conclusive degree to which nature or nurture influence how a gay or lesbian person behaves.” [40]

In the author’s own words,

> “Our findings should not in any way be interpreted so as to imply that the experiences of LGBTQ individuals are “wrong” or “disordered.” In fact, this study provides further evidence that diverse sexual behavior is a natural part of overall human variation. Our research is intended to improve our understanding of the genetic basis of same-sex sexual behavior. It should not be misconstrued to disparage LGBTQ people.” [47]

Critics, however, pointed out how this research could be used in ways that would seriously impact queer communities and individuals, including possible embryo-screening and genetic-tests to determine one’s sexual orientation. Steven Reilly, a queer member of the institute which performed the study (although he was unaware of the work until it was completed) said “We just went into harm reduction mode, but it was too late to really address some of our bigger concerns.” [90]

Indeed, it is unarguable that this research caused significant harms.

Let us look to our list of research harms.

Acquisition harms are largely absent in this case. Provided that the genetic samples were gathered with the consent of all parties, and followed best practice in line with standard bioethics, we can assume that any other acquisition harms were negligible.

The context harms of this research are, however, significant. Let us begin with **socially harmful research**. Recall that this occurs when research does harm to the social context within which a particular group or individual resides. The existence of a 'gay gene' and published work on it does considerable damage along this dimension for a number of reasons,

1. because should it find clear genetic markers this adds fuel to already existing arguments that homosexuality is a harmful genetic mutation ("Some go so far as to say that if God has allowed some people to be born gay, why should we not accept the person's sexual orientation? More, probably, see homosexuality as an unfortunate birth defect, like a hare-lip, crossed eyes, or Down syndrome, to be corrected if possible." [72]). It also plays into narratives of homosexuality as a disease or something to be treated medically.

2. Alternatively, and commonly, any evidence that homosexuality is genetic in origin is simply dismissed as irrelevant to the question, and so does little to help advance queer rights. ("Now the sober truth of the matter is that we don't fully understand the roles of heredity and environment in producing a homosexual orientation. But that doesn't really matter. Even if homosexuality were completely genetic, that fact alone wouldn't imply that such a lifestyle is morally acceptable and should be indulged." [30])

3. should it fail to find clear markers, this can be used to justify the idea that homosexuality is a choice. This increases the stigma and pressure on members of the queer community and can be used to justify gay conversion therapy or other techniques designed to ‘deprogram’ queer individuals. (“There is no ”gay gene” and studies of twins prove this, supporting the idea that it is a lifestyle choice” [26])

4. should it find a mix of both weak indicators for a genetic basis, and that it is partially due to social upbringing, the worst of both of the above will occur. This is, of course, what the study found.

5. because it confirms that locating the source of queer identity is a legitimate study for geneticists, potentially implying that it is seen as a ‘deviation’ from the norm, rather than part of the standard human experience. That is, in many cases scientists are looking to explain why queer people are different, rather than study human sexuality as a broader topic.

6. and, in the context of the actual study, Ganna et al. included in the publication a correlation between the genetic markers for homosexuality and the genetic markers for mental illness (Figure 4.1), adding to the already existing narrative that queerness is both an illness, and a harmful trait to possess.

In a society which already stigmatises queerness, and where regardless of outcome the study will be seen as adding legitimacy to homophobic arguments, the pursuit of the study causes clear and present harms to individuals within the queer community.

The work also poses **Significantly increased risk of harmful application**. The

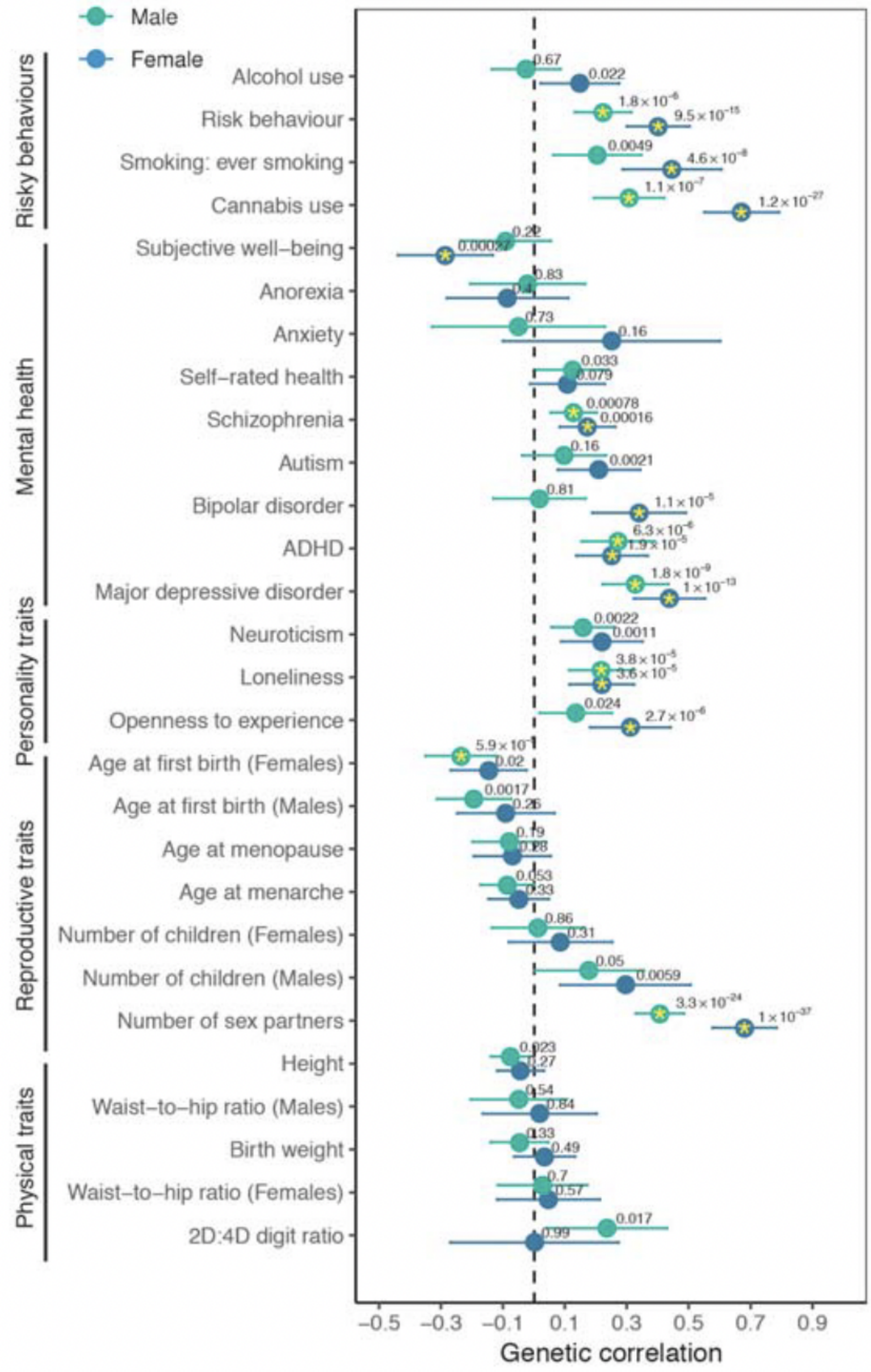


Figure 4.1: Genetic correlations of same-sex sexual behavior with various preselected traits and disorders, separately for males and females. Taken from Ganna et al (2019) [48]

release of the data was widely criticised because, in a world where genetic testing is increasingly easy to access, it was incredibly likely that someone would attempt to start determining who was ‘gay’ based on their genetic markers.

The scientists themselves acknowledged the existence of this harm,

> “We felt an obligation to try to represent the work in as honest and accurate a way as possible, to try to head off at the pass potential misuse to the best of our ability, and, you know, I’m not sure there’s more we could have done.” [90]

Perhaps predictably, within the month a phone app was available online called ‘How Gay Are You?’ The app promised to deliver answers based on the results of the 2019 study. Even more extremely the implication that such a tool exists puts huge numbers of people at risk world-wide, regardless of whether they actually work. Homosexuality is illegal in seventy-one countries. In ten of these it is punishable by death. Regardless of whether there is a clear genetic basis for homosexuality, it is extremely likely that groups looking to persecute queer individuals will use the results of the research to attempt to identify targets.

Beyond this, there is the worry that by providing genetic markers, even weak ones, it opens up the idea that homosexuality is something which can be cured via treatment, or even prevented via eugenics. Indeed, this is already being anticipated by a number of hate groups,

> “If a biological basis is found, and if a prenatal test is then developed, and if a successful treatment to reverse the sexual orientation to heterosexual is ever developed, we would support its use as we should unapologetically support the use of any appropriate means to avoid sexual temptation and the inevitable effects of sin.” [89]

All of these are serious implications of the research done by Ganna et al. and ones which cannot be easily or effectively removed by banning the use of the data.

This work is also **psychologically harmful**. Members of the queer community already face serious issues, many of which stem from insecurity over their place in society, and fear of reprisal from those around them. Research which proposes a genetic correlation between homosexuality and mental health problems, then, simply adds to this emotional load, and can lead individuals who are already disposed to think something is wrong with them to believe that their orientation is going to be the cause of even further hardships.

Finally, a less serious harm, but still a clear one of this research, derives from its **harmful diversion of resources and attention**. Ganna et al. are top scientists working at the Broad Institute, a world renowned research institute. In doing this work they actively chose to pursue this over other areas which could have proven to be more clearly beneficial, and spent significant research funding doing so. Moreover, the controversy surrounding their work forced other researchers, especially those involved in queer advocacy in science (an already difficult and exhausting pastime), to spend time and energy consulting with the group to try and minimise harms, and then dealing with the issues that subsequently arose from the fallout. This was attention and energy which could have been spent on projects to advance queer rights, rather than trying to prevent backslide.

Having identified the harms, we can now see what types of control methods would be most effective.

To start, it is clear that a focus on controlling the use of this research is insufficient to mitigate or prevent the damages. We might, for example, attempt to ban apps which tested for the 'gay gene' or the development of gene-therapy to remove it from embryos. While it is certainly important, perhaps even critical, that such controls now be put in place to attempt to minimise the future harms of the research, such a focus would fail to capture or even recognise the harms that the research has already caused.

Instead, if we want to prevent future harms of further ‘gay gene’ research, or had wanted to prevent the harms from occurring in the first place, we would need a research moratorium tailored to the categories we have identified.

It is apparent, for example, that the worst of the current harms are due to the social damages of the research, its guaranteed misuse by individuals who are looking to persecute members of the queer community, and the suggested correlation between homosexuality and mental health problems. All three of these are grounded in the existence of particular social prejudices, including the stigmatisation of mental health, and thus we can recognise that there are circumstances where these harms will disappear under the right conditions.

Conversely, the diversion harms will stop occurring either when enough resources are in play that the research doesn’t significantly detract from other more important scientific or activist projects, or when the more serious battles for sexual-orientation equality in the sciences have already been dealt with by activists.

So what does this mean?

First, the absence of acquisition harms should lead us to think that we need a temporary moratorium rather than a permanent ban.

Second, the most serious harms are derived from the social context within which the research is being done. As such, the prevention of these harms requires a moratorium in place until these particular homophobic social structures either disappear or are relegated to enough of a minority that the damages done are clearly outweighed by the benefits of knowing more about the genetic grounding of sexual orientation. Secondary conditions for the less serious harms can also be posited, namely to wait until mental health resources and support has changed, or more resources are available.

Of course, the cat is already out of the bag for 'gay gene' research, at least as far as Ganna et al.'s work. The harms are done and now our job is to minimise those that arise from use. A moratorium like the described here on future research into the topic, however, would prevent the further entrenchment and amplification of the research harms, especially given the minimal benefits of the work.

4.6 Conclusion

We, as a society, want to minimise the harms brought about by science. Often the mitigation of these harms looks like controlling the use or application of scientific research - banning nuclear weaponry, autonomous military drones, human germ-line editing, or geo-engineering. In some cases, however, controlling the use of scientific research isn't enough to prevent harm from happening. Instead, the research itself is the source of the damages, and should therefore be monitored, restricted, or factored into our weighing up of the pros and cons of pursuing a particular line of work. Thus far, however, it has been worryingly absent from the conversation surrounding topics like scientific freedom, dual use research, and moratoriums.

Here, I have begun to categorise the most significant of these harms, harms which cannot be prevented by controlling application alone. An awareness of these categories can, in turn, help us understand how and when to apply moratoriums, mitigate damages, and gives scientists and policymakers the language to discuss these harms in more sophisticated ways. It is my hope that by being clearer on exactly what types of harms research can do, we can also get clearer on how to effectively minimise them.

Above I have given two categories of harm that come directly from doing research; acquisition and context. Acquisition harms, I have argued, are inherent to the research itself, and thus if we decide these harms are significant enough to warrant control, these controls ought to be permanent. Conversely, context harms are those which occur because of their interaction with external factors, and thus are potential subjects for temporary moratoriums which are removed once certain conditions are met.

Within these categories we can find a number of different types of harm including those which, when serious enough, may warrant bans or moratoriums on research which causes them. Even in cases where the type of harm is likely to be outweighed by benefits, as with the climate depression case, it is still important that scientists, policy-makers, and philosophers be aware that these harms exist, and where possible take efforts to minimise them.

More broadly, the existence of these harms seems to explicitly tell against the idea that research is always beneficial, and that we should always place moratoriums and bans on the implementation of science, rather than the study. Research can harm and, indeed, can harm in a wide number of importantly distinct ways.

Bibliography

[1] Risks and benefits of dual-use research. *Nature*, 435.

[2] Invasive alien species. *Convention on Biological Diversity, United Nations Decade on Biodiversity*, 2006.

[3] A call to protect food systems from genetic extinction technology: The global food and agriculture movement says no to release of gene drives. 2016.

[4] E Agazzi. Responsibility: The genuine ground for the regulation of a free science. *Scientists and their Responsibilities*, pages 203–219, 1989.

[5] UN General Assembly. Treaty on the prohibition of nuclear weapons. In *A/CONF*, volume 229, page 8, 2017.

[6] BURT Austin, Robert Trivers, and Austin Burt. *Genes in conflict: the biology of selfish genetic elements*. Harvard University Press, 2009.

[7] Luchuo Engelbert Bain and Gerald Chia Gwain. Cardiovascular disease and hiv infection in sub-saharan africa: misplaced priorities in the public health and research agendas? *Frontiers in Cardiovascular Medicine*, 6:35, 2019.

[8] Nicholas J Bax and Ronald E Thresher. Ecological, behavioral, and genetic factors influencing the recombinant control of invasive pests. *Ecological Applications*, 19(4):873–888, 2009.

[9] Andrea Beaghton, Pantelis John Beaghton, and Austin Burt. Gene drive through a landscape: reaction–diffusion models of population suppression and elimination by a sex ratio distorter. *Theoretical population biology*, 108:51–69, 2016.

[10] Camilla J Beech, J Nagaraju, SS Vasan, Robert I Rose, Rofina Yasmin Othman, Vilasini Pillai, and TS Saraswathy. Risk analysis of a hypothetical open field release of a self-limiting transgenic aedes aegypti mosquito strain to combat dengue. *Asia Pacific Journal of Molecular Biology and Biotechnology*, 17(3):97–108, 2009.

[11] Megan Blomfield. Geoengineering in a climate of uncertainty. *Climate change and justice*, pages 39–58, 2015.

[12] LB Bull and MW Mules. An investigation of myxomatosis cuniculi with special reference to the possible use of the disease to control rabbit populations in australia. *Journal of the Council for Scientific and Industrial Research, Australia*, 17(2), 1944.

[13] Katie Burke. An antidote to climate despair. 2021.

[14] George W Bush. President and prime minister blair discussed iraq, middle east. *November*, 12:20041112–5, 2004.

[15] Ken Caldeira and David W Keith. The need for climate engineering research. *Issues in Science and Technology*, 27(1):57–62, 2010.

[16] Daniel Edward Callies. The slippery slope argument against geoengineering research. *Journal of Applied Philosophy*, 36(4):675–687, 2019.

[17] Arthur L Caplan, Brendan Parent, Michael Shen, and Carolyn Plunkett. No time to waste—the ethical challenges created by crispr. *EMBO reports*, 16(11):1421–1426, 2015.

[18] Eli Carrami, Kolja N Eckermann, Hassan MM Ahmed, Héctor Sánchez, Stefan Dippel, John M Marshall, and Ernst A. Wimmer. Consequences of resistance evolution in a cas9-based sex conversion-suppression gene drive for insect pest management. *Proceedings of the National Academy of Sciences*, 115(24):6189–6194, 2018.

[19] Jackson Champer, Joanna Zhao, Sam Champer, Jingxian Liu, and Philipp W Messer. Population dynamics of underdominance gene drive systems in continuous space. *bioRxiv*, page 449355, 2018.

[20] MC Chang, S Pickworth, and RW McGaughey. Experimental hybridization and chromosomes of hybrids. In *Comparative Mammalian Cytogenetics*, pages 132–145. Springer, 1969.

[21] Katelyn Cioffi, Victoria Adelmant, Christiaan van Veen, Laura Bingham, Ed Deluca, Sarbjot Kaur Dhillon, and Bianca Evans. Response to request for information. *Center for Human Rights and Global Justice*, 2022.

[22] Miguel Clavero and Emili García-Berthou. Invasive species are a leading cause of animal extinctions. *Trends in ecology & evolution*, 20(3):110, 2005.

[23] Catherine Clifford. Climate change is radicalizing young people — here's what that means and how to combat despair. 2021.

[24] MN Clout. Biodiversity loss caused by invasive alien vertebrates. *Zeitschrift für Jagdwissenschaft*, 48(1):51–58, 2002.

[25] Le Cong, F Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D Hsu, Xuebing Wu, Wenyan Jiang, and Luciano A Marraffini. Multiplex genome engineering using crispr/cas systems. *Science*, 339(6121):819–823, 2013.

[26] Conservapedia. Best arguments against homosexuality. 2022.

[27] Brian Cooke, Randall Jones, and Wendy Gong. An economic decision model of wild rabbit oryctolagus cuniculus control to conserve australian native vegetation. *Wildlife Research*, 37(7):558–565, 2011.

[28] Brian D Cooke and Frank Fenner. Rabbit haemorrhagic disease and the biological control of wild rabbits, oryctolagus cuniculus, in australia and new zealand. *Wildlife Research*, 29(6):689–706, 2002.

[29] P Cowan and B Warburton. Animal welfare and ethical issues in island pest eradication. *Island Invasives: Eradication and Management'.(Eds CR Veitch, MN Clout and DR Towns.) pp*, pages 418–421, 2011.

[30] William Craig. A christian perspective on homosexuality. 2019.

[31] Robert A Dahl. *Democracy and its Critics*. Yale university press, 2008.

[32] Bonnie Docherty. Making the case: The dangers of killer robots and the need for a preemptive ban. 2016.

[33] Peter C Doherty. *Pandemics: What Everyone Needs to Know®*. Oxford University Press, 2013.

[34] Douglas W Drury, Amy L Dapper, Dylan J Siniard, Gabriel E Zentner, and Michael J Wade. Crispr/cas9 gene drives in genetically variable and nonrandomly mating wild populations. *Science advances*, 3(5):e1601910, 2017.

[35] Victor Arnold Dyck, Jorge Hendrichs, and Alan S Robinson. *Sterile insect technique: principles and practice in area-wide integrated pest management*. Springer, 2006.

[36] Steve Ebbert and Kathy Burek-Huntington. Anticoagulant residual concentration and poisoning in birds following a large-scale aerial application of 25 ppm brodifacoum bait for rat eradication on rat island, alaska. In *Proceedings of the Vertebrate Pest Conference*, volume 24, 2010.

[37] Philip A Eckhoff, Edward A Wenger, H Charles J Godfray, and Austin Burt. Impact of mosquito gene drive on malaria elimination in a computational model with explicit spatial and temporal dynamics. *Proceedings of the National Academy of Sciences*, 114(2):E255–E264, 2017.

[38] John T Edsall. Scientific freedom and responsibility: Report of the aaas committee on scientific freedom and responsibility. *Science*, 188(4189):687–693, 1975.

[39] Jon Elster. *Explaining social behavior: More nuts and bolts for the social sciences*. Cambridge University Press, 2015.

[40] Dawn Ennis. The 'gay gene' is a myth but being gay is 'natural,' say scientists. 2019.

[41] Markus I Eronen. Psychopathology and truth: A defense of realism. In *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine*, volume 44, pages 507–520. Oxford University Press US, 2019.

[42] Kevin M Esvelt and Neil J Gemmell. Conservation demands safe gene drive. *PLoS biology*, 15(11):e2003850, 2017.

[43] John H Evans. *Playing god?: human genetic engineering and the rationalization of public bioethical debate*. University of Chicago Press, 2002.

[44] Terence R Flotte. Senior gene therapy scientists take a stand against human embryo gene editing, 2019.

[45] John EC Flux. Relative effect of cats, myxomatosis, traditional control, or competitors in removing rabbits from islands. *New Zealand Journal of Zoology*, 20(1):13–18, 1993.

[46] Matt Frost. After climate despair. 2019.

[47] Andrea Ganna, Karin JH Verweij, Michel G Nivard, Robert Maier, Robbee Wedow, Alexander S Busch, Abdel Abdellaoui, Shengru Guo, J Fah Sathirapongsasuti, 23andMe Research Team 16, Paul Lichtenstein, Sebastian Lundström, Niklas Langstrom, Adam Auton, Kathleen Mullan Harris, Gary W. Beecham, Eden R. Martin, Alan R. Sanders, John R. B. Perry, Benjamin M. Neale, and Brendan P. Zeitsch. Genetics of sexual behaviour - why we did this study. 2019.

[48] Andrea Ganna, Karin JH Verweij, Michel G Nivard, Robert Maier, Robbee Wedow, Alexander S Busch, Abdel Abdellaoui, Shengru Guo, J Fah Sathirapongsasuti, 23andMe Research Team 16, Paul Lichtenstein, Sebastian Lundström, Niklas Langstrom, Adam Auton, Kathleen Mullan Harris, Gary W. Beecham, Eden R. Martin, Alan R. Sanders, John R. B. Perry, Benjamin M. Neale, and Brendan P. Zeitsch. Large-scale gwas reveals insights into the genetic architecture of same-sex sexual behavior. *Science*, 365(6456):eaat7693, 2019.

[49] Peter Gärdenfors. Is there anything we should not want to know? In *Studies in Logic and the Foundations of Mathematics*, volume 126, pages 63–78. Elsevier, 1989.

[50] Stephen M Gardiner and Augustin Fragniere. The tollgate principles for the governance of geoengineering: Moving beyond the oxford principles to an ethically more robust approach. *Ethics, Policy & Environment*, 21(2):143–174, 2018.

[51] P Genovesi. Are we turning the tide? eradications in times of crisis: how the global community is responding to biological invasions. *Island invasives: eradication and management*, pages 5–8, 2011.

[52] Simona Giordano and John Harris. *The freedom of scientific research: Bridging the gap between science and society*. Manchester University Press, 2020.

[53] Andrew Hammond, Roberto Galizi, Kyros Kyrou, Alekos Simoni, Carla Siniscalchi, Dimitris Katsanos, Matthew Gribble, Dean Baker, Eric Marois, Steven Russell, Austin Burt, Nikolai Windbichler, Andrea Crisanti, and Tony Nolan. A crispr-cas9 gene drive system targeting female reproduction in the malaria mosquito vector anopheles gambiae. *Nature biotechnology*, 34(1):78–83, 2016.

[54] Andrew M Hammond, Kyros Kyrou, Marco Bruttini, Ace North, Roberto Galizi, Xenia Karlsson, Nace Kranjc, Francesco M Carpi, Romina D'Aurizio, Andrea Crisanti, and Tony Nolan. The creation and selection of mutations resistant to a gene drive over multiple generations in the malaria mosquito. *PLoS genetics*, 13(10):e1007039, 2017.

[55] Matthew Hanser. The metaphysics of harm. *Philosophy and Phenomenological Research*, 77(2):421–450, 2008.

[56] Tim Harvey-Samuel, Thomas Ant, and Luke Alphey. Towards the genetic control of invasive species. *Biological Invasions*, 19(6):1683–1703, 2017.

[57] Evridiki Hatziandreu, John D Graham, and Michael A Stoto. Aids and biomedical research funding: comparative analysis. *Reviews of infectious diseases*, 10(1):159–167, 1988.

[58] Carl G Hempel. Fundamentals of concept formation in empirical science, vol. ii. no. 7. 1952.

[59] Wendy R Henderson and Elaine C Murphy. Pest or prized possession? genetically modified biocontrol from an international perspective. *Wildlife Research*, 34(7):578–585, 2008.

[60] Chad L Hewitt. Marine biosecurity issues in the world oceans: global activities and australian directions. *Ocean Yearbook Online*, 17(1):193–212, 2003.

[61] Joshua B Horton, Jesse L Reynolds, Holly Jean Buck, Daniel Callies, Stefan Schäfer, David W Keith, and Steve Rayner. Solar geoengineering and democracy. *Global Environmental Politics*, 18(3):5–24, 2018.

[62] Philip E Hulme. Trade, transport and trouble: managing invasive species pathways in an era of globalization. *Journal of applied ecology*, 46(1):10–18, 2009.

[63] Philip E Hulme, Sven Bacher, Mark Kenis, Stefan Klotz, Ingolg Kühn, Dan Minchin, Wolfgang Nentwig, Sergej Olenin, Vadim Panov, Jan Pergl, P. Pysek, A. Roques, D. Sol, W. Solarz, and M. Vilà. Grasping at the routes of biological invasions: a framework for integrating pathways into policy. *Journal of Applied Ecology*, 45(2):403–414, 2008.

[64] Scott E Hygnstrom, Robert M Timm, Paul D Curtis, Dale L Nolte, Mark E Tobin, and Kurt C VerCauteren. Prevention and control of wildlife damage. In *Proceedings of the Vertebrate Pest Conference*, volume 26, 2014.

[65] John Innes and Gary Barker. Ecological consequences of toxin use for mammalian pest control in new zealand—an overview. *New Zealand Journal of Ecology*, pages 111–127, 1999.

[66] Deborah G Johnson. Forbidden knowledge and science as professional activity. *The Monist*, 79(2):197–217, 1996.

[67] Deborah G Johnson. Reframing the question of forbidden knowledge for modern science. *Science and Engineering Ethics*, 5(4):445–461, 1999.

[68] Immanuel Kant. *Perpetual peace*. Wildside Press, 1903.

[69] Samuel Karlin and James McGregor. Application of method of small parameters to multi-niche population genetic models. *Theoretical population biology*, 3(2):186–209, 1972.

[70] Peter J Kerr and SM Best. Myxoma virus in rabbits. *Revue scientifique et technique-Office international des epizooties*, 17:256–264, 1998.

[71] Philip Kitcher. An argument about free inquiry. *Noûs*, 31(3):279–306, 1997.

[72] Samuel Koranteng-Pipim. Born a gay and born again? adventism's changing attitutde toward homosexuality. *Journal of the Adventist Theological Society*, 10(1):11, 2000.

[73] Kyros Kyrou, Andrew M Hammond, Roberto Galizi, Nace Kranjc, Austin Burt, Andrea K Beaghton, Tony Nolan, and Andrea Crisanti. A crispr–cas9 gene drive targeting doublesex causes complete population suppression in caged anopheles gambiae mosquitoes. *Nature biotechnology*, 36(11):1062, 2018.

[74] Eric S Lander, Françoise Baylis, Feng Zhang, Emmanuelle Charpentier, Paul Berg, Catherine Bourgain, Bärbel Friedrich, J Keith Joung, Jinsong Li, David Liu, , Luigi Naldini, Jing-Bao Nie, Renzong Qiu, Bettina Schoene-Seifert, Feng Shao, Sharon Terry, Wensheng Wei, and Ernst-Ludwig Winnacker. Adopt a moratorium on heritable genome editing, 2019.

[75] Robert Jan Lebbink, Dorien de Jong, Femke Wolters, Elisabeth M Kruse, Petra M van Ham, Emmanuel JHJ Wiertz, and Monique Nijhuis. A combinational crispr/cas9 gene-editing approach can halt hiv replication and prevent viral escape. *Scientific reports*, 7(1):1–10, 2017.

[76] Alexander C Lees and Diana J Bell. A conservation paradox for the 21st century: the european wild rabbit oryctolagus cuniculus, an invasive alien and an endangered native species. *Mammal Review*, 38(4):304–320, 2008.

[77] Caroline M Leitschuh, Dona Kanavy, Gregory A Backus, Rene X Valdez, Megan Serr, Elizabeth A Pitts, David Threadgill, and John Godwin. Developing gene drive technologies to eradicate invasive rodents from islands. *Journal of Responsible innovation*, 5(sup1):S121–S138, 2018.

[78] Agnieszka Lekka-Kowalik. “on freedom and limits of scientific inquiry”, in: Z. zdybicka (ed.), freedom in contemporary culture, lublin: Rw kul, 1999, 547-561. 08 1999.

[79] Albert C Lin. Does geoengineering present a moral hazard. *Ecology LQ*, 40:673, 2013.

[80] Marc Lipsitch and Thomas V Inglesby. Moratorium on research intended to create novel potential pandemic pathogens, 2014.

[81] Sarah Lowe, Michael Browne, Souyad Boudjelas, and Maj De Poorter. *100 of the world's worst invasive alien species: a selection from the global invasive species database*, volume 12. Invasive Species Specialist Group Auckland, 2000.

[82] John M Marshall and Bruce A Hay. Confinement of gene drive systems to local populations: a comparative analysis. *Journal of Theoretical Biology*, 294:153–171, 2012.

[83] Hugh R McCrimmon. Carp in canada. 1968.

[84] William McGucken. On freedom and planning in science: The society for freedom in science, 1940-46. *Minerva*, pages 42–72, 1978.

[85] Alison McIntyre. Doctrine of double effect. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2019 edition, 2019.

[86] Catriona McKinnon. Sleepwalking into lock-in? avoiding wrongs to future people in the governance of solar radiation management research. *Environmental Politics*, 28(3):441–459, 2019.

[87] John Stuart Mill. On the definition and method of political economy. *The philosophy of economics*, pages 41–58, 1836.

[88] John Min, Charleston Noble, Devora Najjar, and Kevin Esvelt. Daisy quorum drives for the genetic restoration of wild populations. *BioRxiv*, page 115618, 2017.

[89] Albert Mohler. Is your baby gay? what if you could know? what if you could do something about it? 2007.

[90] Megan Molteni. How earnest research into gay genetics went wrong. 2019.

[91] Nancy A Moran and Tyler Jarvik. Lateral transfer of genes from fungi underlies carotenoid production in aphids. *science*, 328(5978):624–627, 2010.

[92] Carroll Muffett and Steven Feit. Fuel to the fire; how geoengineering threatens to entrench fossil fuels and accelerate the climate crisis. *Center for International Environmental Law (CIEL)*, 2019.

[93] Harald Müller and Jonas Wolff. Democratic peace: Many data, little explanation? In *Democratic Wars*, pages 41–73. Springer, 2006.

[94] Judith H Myers, Anne Savoie, and Ed van Randen. Eradication and pest management. *Annual review of entomology*, 43(1):471–491, 1998.

[95] Md National Commission for the Proptection of Human Subjects of Biomedicaland Behavioral Research, Bethesda. *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. Superintendent of Documents, 1978.

[96] Office of Science Policy National Institute for Health. Dual use research of concern. 2022.

[97] Dorothy Nelkin. Threats and promises: Negotiating the control of research. *Daedalus*, pages 191–209, 1978.

[98] Katerina Nikolouli, Hervé Colinet, David Renault, Thomas Enriquez, Laurence Mouton, Patricia Gibert, Fabiana Sassu, Carlos Cáceres, Christian Stauffer, Rui Pereira, and Kostas Bourtzis. Sterile insect technique and wolbachia symbiosis as potential tools for the control of the invasive species drosophila suzukii. *Journal of pest science*, 91(2):489–503, 2018.

[99] Eric Niler. How landing the first man on the moon cost dozens of lives. 2019.

[100] Charleston Noble, John Min, Jason Olejarz, Joanna Buchthal, Alejandro Chavez, Andrea L Smidler, Erika A DeBenedictis, George M Church, Martin A Nowak, and Kevin M Esvelt. Daisy-chain gene drives for the alteration of local populations. *Proceedings of the National Academy of Sciences*, 116(17):8275–8282, 2019.

[101] Ace North, Austin Burt, and H Charles J Godfray. Modelling the spatial spread of a homing endonuclease gene in a mosquito population. *Journal of Applied Ecology*, 50(5):1216–1225, 2013.

[102] Ngum Helen Ntonifor and Serophine Veyufambom. Assessing the effective use of mosquito nets in the prevention of malaria in some parts of mezam division, northwest region cameroon. *Malaria journal*, 15(1):1–8, 2016.

[103] Peter O Hara. The illegal introduction of rabbit haemorrhagic disease virus in new zealand. *Revue scientifique et technique-Office international des épizooties*, 25(1):119, 2006.

[104] US Department of Defence. Dir 3000.09, autonomy in weapons systems. 2012.

[105] US Department of Health and Human Services. United states government policy for institutional oversight of life sciences dual use research of concern, 2013.

[106] Future of Life Institute. An open letter to the united nations convention on certain conventional weapons. 2017.

[107] World Health Organization. Who priority research questions for tb/hiv in hiv-prevalent and resource-limited settings. 2010. *Available from: whqlibdoc. who. int/publications/2010/9789241 500302_eng. pdf*, 2012.

[108] Zeynep Pamuk. Risk and fear: Restricting science under uncertainty. *Journal of Applied Philosophy*, 38(3):444–460, 2021.

[109] Martin Paparella. Rodenticides-an animal welfare paradox? *ALTEX-Alternatives to animal experimentation*, 23(1):51–52, 2006.

[110] Mike Pearl. 'climate despair' is making people give up on life. 2019.

[111] Marcello Pera. Should science be supervised, and if so by whom. 1989.

[112] I Valentin Petrescu-Mag, Miklos Botha, and Claudiu Gavriloaie. Lepus× oryctolagus cuniculus hybrids: incompatibilities of behavioral and molecular nature. *Rabbit Genetics*, 8(1):23–25, 2018.

[113] Mike Picker. *Alien and invasive animals: A South African perspective.* Penguin Random House South Africa, 2013.

[114] Adriana Placani. When the risk of harm harms. *Law and Philosophy*, 36(1):77–100, 2017.

[115] Thomas AA Prowse, Phillip Cassey, Joshua V Ross, Chandran Pfitzner, Talia A Wittmann, and Paul Thomas. Dodging silver bullets: good crispr gene-drive design is critical for eradicating exotic vertebrates. *Proceedings of the Royal Society B: Biological Sciences*, 284(1860):20170799, 2017.

[116] Petr Pyšek and David M Richardson. Invasive species, environmental change and management, and health. *Annual review of environment and resources*, 35:25–55, 2010.

[117] James Lee Ray. Wars between democracies: rare, or nonexistent? *International Interactions*, 18(3):251–276, 1993.

[118] Sara Reardon. Viral research moratorium called too broad. *Nature News, 23rd October*, 2014.

[119] Thomas Risse-Kappen. Democratic peace—warlike democracies? a social constructivist interpretation of the liberal argument. *European journal of international relations*, 1(4):491–517, 1995.

[120] Andrew Roberts, Paulo Paes De Andrade, Fredros Okumu, Hector Quemada, Moussa Savadogo, Jerome Amir Singh, and Stephanie James. Results from the workshop "problem formulation for the use of gene drive in mosquitoes". *The American journal of tropical medicine and hygiene*, 96(3):530–533, 2017.

[121] Nicolas O Rode, Arnaud Estoup, Denis Bourguet, Virginie Courtier-Orgogozo, and Florence Débarre. Population management using gene drive: molecular design, models of spread dynamics and assessment of ecological risks. *Conservation Genetics*, pages 1–20, 2019.

[122] Holmes Rolston. *Genes, genesis, and God: Values and their origins in natural and human history.* Number 76. Cambridge University Press, 1999.

[123] J Ross and MF Sanders. The development of genetic resistance to myxomatosis in wild rabbits in britain. *Epidemiology & Infection*, 92(3):255–261, 1984.

[124] James C Russell, David R Towns, Sandra H Anderson, and Mick N Clout. Intercepting the first rat ashore. *Nature*, 437(7062):1107, 2005.

[125] Elmer Eric Schattschneider and Sidney A Pearson. *Party government: American government in action*. Routledge, 2017.

[126] Paul Schliekelman and Fred Gould. Pest control by the release of insects carrying a female-killing allele on multiple loci. *Journal of Economic Entomology*, 93(6):1566–1579, 2000.

[127] Kai Schönig, Tillmann Weber, Ariana Frömmig, Lena Wendler, Brigitte Pesold, Dominik Djandji, Hermann Bujard, and Dusan Bartsch. Conditional gene expression systems in the transgenic rat brain. *BMC biology*, 10(1):77, 2012.

[128] Dane Scott. The technological fix criticisms and the agricultural biotechnology debate. *Journal of agricultural and environmental ethics*, 24(3):207–226, 2011.

[129] Hanno Seebens, Tim M Blackburn, Ellie E Dyer, Piero Genovesi, Philip E Hulme, Jonathan M Jeschke, Shyama Pagad, Petr Pyšek, Marten Winter, Margarita Arianoutsou, Sven Bacher, Bernd Blasius, Giuseppe Brundu, César Capinha, Laura Celesti-Grapow, Wayne Dawson, Stefan Dullinger, Nicol Fuentes, Heinke Jäger, John Kartesz, Marc Kenis, Holger Kreft, Ingolf Kühn, Bernd Lenzner, Andrew Liebhold, Alexander Mosena, Dietmar Moser, Misako Nishino, David Pearman, Jan Pergl, Wolfgang Rabitsch, Julissa Rojas-Sandoval, Alain Roques, Stephanie Rorke, Silvia Rossinelli, Helen E. Roy, Riccardo Scalera, Stefan Schindler, Kateřina Štajerová, Barbara Tokarska-Guzik, Mark van Kleunen, Kevin Walker, Patrick Weigelt, Takehiko Yamanaka, and Franz Essl. No saturation in the accumulation of alien species worldwide. *Nature communications*, 8:14435, 2017.

[130] Morgan Shaw. Geoengineering and future generations- responsible technological development under climate change. *COOLEST STUDENT PAPERS AT FINLAND FUTURES RESEARCH CENTRE 2017–2018*, page 108, 2019.

[131] Daniel Simberloff. We can eliminate invasions or live with them. successful management projects. In *Ecological Impacts of Non-Native Invertebrates and Fungi on Terrestrial Ecosystems*, pages 149–157. Springer, 2008.

[132] Potter Stewart. Jacobellis v ohio. *US Rep*, 378:184, 1964.

[133] Jack Stilgoe, Richard Owen, and Phil Macnaghten. Developing a framework for responsible innovation. In *The Ethics of Nanotechnology, Geoengineering and Clean Energy*, pages 347–359. Routledge, 2020.

[134] Telford Taylor. *Final Report to the Secretary of the Army on Nuernberg War Crimes Trials Under Control Council Law No. 10*, volume 10. US Government Printing Office, 1950.

[135] Lewis Thomas. The hazards of science, 1977.

[136] Ronald E Thresher. Genetic options for the control of invasive vertebrate pests: prospects and constraints. 2007.

[137] Ronald E Thresher, Michael Jones, and D Andrew R Drake. Evaluating active genetic options for the control of sea lamprey (petromyzon marinus) in the laurentian great lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, (999):1–17, 2018.

[138] David R Towns, Ian AE Atkinson, and Charles H Daugherty. Have the harmful effects of introduced rats on islands been exaggerated? *Biological invasions*, 8(4):863–891, 2006.

[139] Robert L Unckless, Andrew G Clark, and Philipp W Messer. Evolution of resistance against crispr/cas9 gene drive. *Genetics*, 205(2):827–841, 2017.

[140] Eric Vance. What to do about climate despair. 2021.

[141] Stacey Vanek Smith. *Machiavelli for Women: Defend Your Worth, Grow Your Ambition, and Win the Workplace*. Gallery Books, 2021.

[142] Adrian Veres, Bridget S Gosis, Qiurong Ding, Ryan Collins, Ashok Ragavendran, Harrison Brand, Serkan Erdin, Chad A Cowan, Michael E Talkowski, and Kiran Musunuru. Low incidence of off-target mutations in individual crispr-cas9 and talen targeted human stem cell clones detected by whole-genome sequencing. *Cell stem cell*, 15(1):27–30, 2014.

[143] Human Rights Watch. Advancing the debate on killer robots: 12 key arguments for the preemptive ban on fully autonomous weapons. 2014.

[144] Bruce L Webber, S Raghu, and Owain R Edwards. Opinion: Is crispr-based gene drive a biocontrol silver bullet or global conservation threat? *Proceedings of the National Academy of Sciences*, 112(34):10565–10567, 2015.

[145] Max Weber. " objectivity" in social science and social policy. *The methodology of the social sciences*, pages 49–112, 1949.

[146] Lawrence M Wein and Yifan Liu. Analyzing a bioterror attack on the food supply: the case of botulinum toxin in milk. *Proceedings of the National Academy of Sciences*, 102(28):9984–9989, 2005.

[147] Frederick G Whelan. Prologue: Democratic theory and the boundary problem. *Nomos*, 25:13–47, 1983.

[148] Lynn White. The historical roots of our ecologic crisis. *Science*, 155(3767):1203–1207, 1967.

[149] Torsten Wilholt. Scientific freedom: its grounds and their limitations. *Studies in History and Philosophy of Science Part A*, 41(2):174–181, 2010.

[150] Woodrow Wilson. War message to congress. *]- : http://www. heritage. org/initiatives/first-principles/primary-sources/woodrow-wilsons-war-message-tocongress*, 1917.

[151] Eric Winsberg. A modest defense of geoengineering research: A case study in the cost of learning. *Philosophy & Technology*, 34(4):1109–1134, 2021.

[152] Andrea I Woody. Re-orienting discussions of scientific explanation: A functional perspective. *Studies in History and Philosophy of Science Part A*, 52:79–87, 2015.

[153] Hui Zhang, Jinshan Zhang, Pengliang Wei, Botao Zhang, Feng Gou, Zhengyan Feng, Yanfei Mao, Lan Yang, Heng Zhang, Nanfei Xu, and Jian-Kang Zhu. The crispr/c as9 system produces specific and homozygous targeted gene editing in rice in one generation. *Plant biotechnology journal*, 12(6):797–807, 2014.

[154] Xiao-Hui Zhang, Louis Y Tee, Xiao-Gang Wang, Qun-Shan Huang, and Shi-Hua Yang. Off-target effects in crispr/cas9-mediated genome engineering. *Molecular Therapy-Nucleic Acids*, 4:e264, 2015.

[155] Sarah Zielinski. The invasive species we can blame on shakespeare. *Smithsonian. com*, 4, 2011.