Evolution and Existential Risk with Observation Selection Bias

Andrew E. Snyder-Beattie

Wolfson College
University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Trinity 2019

Abstract

If life emerges on a planet, is it likely to become intelligent? And if intelligent life does emerge, how long will it persist? These two questions motivate this thesis, and are linked by a common methodological challenge associated with *observer selection effects*. An observer selection effect is a particular type of sampling bias that occurs when a sample is not representative of all outcomes, but rather a subset of outcomes compatible with the existence of the observers. Naive approaches that fail to account for this potential bias could overestimate the probability that intelligent life emerges or underestimate the rate of catastrophes. An overview of these topics is presented in Chapter 1.

Using highly simplified models, I explore ways that observation selection effects could affect our understanding of the history of life on Earth. In some instances, these simple models enable us to conclude that observation selection effects are unlikely to be a major factor. For example, Chapter 2 investigates the rate of human extinction from natural sources, and finds that the rate is reassuringly low and unlikely to be heavily biased by observer selection effects.

In other instances, the simple models provide a reason to think that observation selection effects could have an overwhelming impact, opening up new hypotheses for future research. For example, Chapter 3 investigates the timing of key evolutionary transitions on Earth, and concludes that exceptionally rare events occurred much earlier than expected due to observation selection effects. Chapter 4 builds on this result, finding that catastrophes, when combined with observation selection effects, can accelerate hard evolutionary transitions that past literature may have overlooked.

Finally, in some cases the impact of observer selection bias may be modest but not overwhelming. In Chapter 5, I examine the rate at which catastrophes might destroy the biosphere, and conclude that although observation selection effects may play some role in underestimating the rate, it is unlikely to bias our estimates enough to be meaningful on the scale of human lifetimes. I finish with a short discussion of some general principles suggested by the work, and directions for future research.

Evolution and Existential Risk with Observation Selection Bias



Andrew E. Snyder-Beattie

Wolfson College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2019

Acknowledgements

First, I would like to thank my supervisor Mike Bonsall for his time, expertise, and encouragement. Finding a mentor willing to dive into crazy ideas and help impose some mathematical structure on them has been a huge blessing.

My collaborators have been a joy to work with. Thank you to Anders Sandberg, Toby Ord, and Eric Drexler for all of their positive energy and obscure knowledge. I also benefited greatly from feedback and conversations with countless people including Stuart Armstrong, Joe Carlsmith, Katja Grace, Sam Levin, Gregory Lewis, and Carl Shulman.

This journey started thanks to Doyne Farmer's encouragement and support, for which I am exceptionally grateful. I also owe a great deal to my previous academic mentors, Kevin Gross and Shawn Humphrey.

Finally, a huge thanks to Jaan Tallinn and the Open Philanthropy Project for funding this work.

Statement of Originality

I declare that this thesis was composed by myself and the work contained herein is my own except where explicitly stated in the text. This work has not been submitted for any other degree or professional qualification.

Andrew E. Snyder-Beattie
Trinity Term 2019

Statement of Authorship

Chapter 2 is my own work, supervised by Mike Bonsall. It was inspired by discussions with Toby Ord and his analysis that used Laplace's law of succession to achieve analogous results to the ones I develop using a likelihood framework. Toby Ord also proposed three of the ideas that I extended in the fourth section ('Testing the bound with indirect data').

Chapter 3 is my own work, supervised by Mike Bonsall. Anders Sandberg and Eric Drexler helped to draft the first paragraph of the 'priors for transition rates' section that I subsequently edited for clarity. Anders Sandberg helped with coding the final figure and Eric Drexler wrote a first draft of the section 'combinatorial models for upper bounds,' that I subsequently edited for clarity.

Chapter 4 is my own work, supervised by Mike Bonsall.

Chapter 5 is my own work, supervised by Mike Bonsall. Anders Sandberg helped find literature (about a third of the citations were those that he gave me to look into). Anders Sandberg also drafted about half of the final three paragraphs in the section 'Could intelligence have emerged earlier?'

Abstract

If life emerges on a planet, is it likely to become intelligent? And if intelligent life does emerge, how long will it persist? These two questions motivate this thesis, and are linked by a common methodological challenge associated with *observer selection effects*. An observer selection effect is a particular type of sampling bias that occurs when a sample is not representative of all outcomes, but rather a subset of outcomes compatible with the existence of the observers. Naive approaches that fail to account for this potential bias could overestimate the probability that intelligent life emerges or underestimate the rate of catastrophes. An overview of these topics is presented in Chapter 1.

Using highly simplified models, I explore ways that observation selection effects could affect our understanding of the history of life on Earth. In some instances, these simple models enable us to conclude that observation selection effects are unlikely to be a major factor. For example, Chapter 2 investigates the rate of human extinction from natural sources, and finds that the rate is reassuringly low and unlikely to be heavily biased by observer selection effects.

In other instances, the simple models provide a reason to think that observation selection effects could have an overwhelming impact, opening up new hypotheses for future research. For example, Chapter 3 investigates the timing of key evolutionary transitions on Earth, and concludes that exceptionally rare events occurred much earlier than expected due to observation selection effects. Chapter 4 builds on this result, finding that catastrophes, when combined with observation selection effects, can accelerate hard evolutionary transitions that past literature may have overlooked.

Finally, in some cases the impact of observer selection bias may be modest but not overwhelming. In Chapter 5, I examine the rate at which catastrophes might destroy the biosphere, and conclude that although observation selection effects may play some role in underestimating the rate, it is unlikely to bias our estimates enough to be meaningful on the scale of human lifetimes. I finish with a short discussion of some general principles suggested by the work, and directions for future research.

Contents

1 Background & Motivation **1**

2 An upper bound for the background rate of human extinction **15**

3 The timing of evolutionary transitions suggests intelligent life is rare **35**

4 Catastrophe risk can accelerate unlikely evolutionary transitions **61**

5 Bounding the destruction rate of Earth's biosphere **89**

6 Discussion **117**

Bibliography **121**

Chapter 1

Background & Motivation

If life emerges on a planet, is it likely to become intelligent? And if intelligent life does emerge, how long will it persist? These two questions motivate this thesis, and are linked by a common methodological challenge associated with *observer selection effects*. An observer selection effect is a special type of sampling bias that occurs due to our own existence as observers (Bostrom, 2002*a*). In the same way that a fisher's sample will contain only large fish if the net has holes that allow smaller fish through, intelligent observers will find themselves only in conditions that allowed such observers to emerge.

This concept was introduced by Carter (1974) as the *anthropic principle*, stating that what we can expect to observe must be restricted by the conditions necessary for our presence as observers. Since its introduction, the principle has become widely used in cosmology, seeking to explain why physical parameters are 'fine-tuned' for life (Barrow and Tipler, 1986; Carr and Rees, 1979; Weinberg, 1987).

The earliest discussion of observer selection effects may date back to the writing of Bernard Le Bovier de Fontenelle in 1686 (Ćirković, 2002). A year before Newton's *Principia*, he points to a possible observation selection effect in comet trajectories:

> In the next places, the reason why the planes of [the comet's] motions are not in the plane of the ecliptic, or any of the planetary orbits, is extremely evident; for had this been the case, it would have

> been impossible for the Earth to be out of the way of the comets tails. Nay, the possibility of an immediate encounter or shock of the body, of a comet would have been too frequent; and considering how great is the velocity of a comet at such a time, the collision of two such bodies must necessarily be destructive of each other; nor perhaps could the inhabitants of planets long survive frequent immersions in the tails of comets, as they would be liable to in such a situation.

Fontenelle's quote highlights the fact that observation selection effects not only concern fundamental physical parameters, but also the more local conditions of our own solar system and the risks of catastrophe therein. A small but growing literature has come to appreciate this, and is examining the ways in which observation selection bias can affect problems in fields ranging from biological evolution (Carter, 1983), planetary science (Waltham, 2019), and even risk assessment (Ćirković, 2016).

My focus is primarily on the way that observation selection effects can impact our estimates of catastrophic risks that could set back or end human civilisation, and our estimates of how likely intelligent life is to exist elsewhere in the universe. These topics draw on a diversity of scientific fields such as ecology, palaeontology, astronomy, planetary science, and even linguistics, to name a few. I discuss the relevant literature in each chapter, using this introduction to provide context and motivation for the problems I investigate.

Existential Risk

An *existential risk* is one that threatens the entire future of humanity (Bostrom, 2013*b*), or, more precisely defined, is one that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development (Bostrom, 2002*b*). Common examples of existential risks include asteroid impacts, nuclear war, and climate change (Bostrom and Cirkovic, 2011). A growing body of literature in moral philosophy has argued that the reduction of existential risk could be of

singular importance, since such risks jeopardise the lives of all future human generations (Beckstead, 2013; Parfit, 1984).

Broadly speaking, existential risks can be divided into those that are naturally occurring and those that are anthropogenic. Both categories include both well-understood and speculative risks. Examples of natural hazards include asteroid impacts, supervolcanic eruptions, gamma ray bursts, or a collapse of the false vacuum (Sandberg, 2018). Anthropogenic risks include nuclear war, climate change, and future risks posed by potential technologies such as biotechnology or artificial superintelligence (Bostrom and Cirkovic, 2011). There could also be risks in both categories of which we are still unaware. Chapter 2 aims to determine whether natural or anthropogenic risks pose a greater threat to our future, and also analyses the potential for 'anthropic shadows' affecting these estimates.

Anthropic Shadows

Estimating the frequency of existential risks is complicated in part due to observation selection effects. This was perhaps first discussed by Leslie (2002), who noted that humanity's track record of survival may be an unreliable guide:

> Notice that observational selection effects might help to explain our survival up to date. For instance, we couldn't observe that we were on a planet where disease or an asteroid impact had exterminated all intelligent life, even if such planets formed the majority of those on which such life had evolved. Intelligent living beings cannot find that they are in places devoid of living intelligence!

For some risks, such as asteroid impacts, one could attempt to get around this problem by examining the frequency of past catastrophes in Earth's history before humanity existed. Large asteroid impacts and extreme periods of volcanic activity leave traces in the geological record, which could give us a sense of how frequent such catastrophes are. However, the problems of observation selection effects crop up here too. For example, if human existence

required a 50 million year (Myr) period of evolution free from asteroid impacts, any human observers will necessarily find in their evolutionary history a period of 50 Myr that is free of asteroid impacts, regardless of the true impact rate. Inferring a rate based on crater data from those 50 Myr could therefore be misleading. The presence of an observer casts an 'anthropic shadow' over the data, so that the observer can only see a limited set of data that is compatible with their own existence (Ćirković, 2008; Ćirković et al., 2010).

For some risks, one can potentially find data that are less affected by observation selection bias. For example, examining craters on the moon or on Mars could be used to estimate the rate of impacts on Earth. This approach was used by Tegmark and Bostrom (2005*a*) when studying the rate of natural catastrophes that would destroy our planet altogether, such as a physical phase transition into 'strange matter' or a decay of a false vacuum. For some of these risks that would leave visible evidence, the fact no other planet in our solar system has been destroyed in this way over billions of years allows us to infer that the rate of such a catastrophe, if it exists, must be low. Another unbiased source of data can be found in the fact that Earth itself emerged late in the universe's history when compared to the majority of potentially habitable exoplanets (Lineweaver et al., 2004; Zackrisson et al., 2016). This enables one to put a bound on the rate of catastrophes such as vacuum decay that would destroy large regions of space, since if the rate were high we would expect find ourselves living on one of the first habitable planets (Tegmark and Bostrom, 2005*b*).

Although these data provide a way to bypass observer selection bias, they are only applicable to a limited set of catastrophes that would destroy a planet entirely, or otherwise leave evidence of catastrophe on neighbouring planets. Many existential risks are excluded from this category but could still be subject to observation selection effects, such as disease (Manheim, 2018) or volcanic eruptions. Perhaps most notably, the remarkable stability of the Earth's climate over billions of years is also likely a byproduct of observation selection

effects (Waltham, 2007, 2014). My goal in Chapter 2 is to estimate the rate of extinction risk for the human species in a way that accounts for observation selection effects when we do not have alternative data to draw on. Chapters 3 and 4 focus less on how long intelligent life is likely to last, but rather on how probable such intelligent life is to emerge in the first place. As it turns out, the two questions are inextricably linked by a mechanism known as the 'Great Filter.'

The Great Filter

There are hundreds of billions of stars in our galaxy, and hundreds of thousands of galaxies in our local supercluster. However, despite the abundance of potential locations in which intelligent life could emerge, we have found no evidence for intelligent life beyond Earth. The discrepancy between the vast number of potential locations for intelligent life to emerge and the apparent absence of such life has been dubbed the 'Fermi Paradox,' after Enrico Fermi famously asked 'where are they?' (although as noted by Gray (2015), the actual credit ought to go to Hart (1975)). Some have argued that the paradox is especially puzzling because it would only take one advanced civilisation expanding at a small fraction of the speed of light to create a signature detectable from Earth (Bradbury et al., 2011; Olson, 2016). Since there are trillions of stars but no civilisations advanced enough to produce such a signature, we can reason that one of the steps leading from a solar system to a large and observable interstellar civilisation is exceedingly improbable—perhaps on the order of less than one chance in 10^{24} per 10 billion years (Frank and Sullivan III, 2016). This exceedingly improbable step has been called the 'Great Filter' (Hanson, 1998*a*).

The Great Filter suggests that intelligent life is either exceptionally rare, or that intelligent life is exceptionally unlikely to establish an interstellar civilisation. This in turn has large implications for humanity's future and estimates of existential risk. If intelligent life is common, this would imply that the Great

Filter lies in humanity's future, and would imply that humanity is highly unlikely to achieve a level of technological maturity associated with interstellar civilisation, or that the level of technology required would necessitate some form of catastrophe that would short circuit our civilisation's development. Therefore if we find evidence that intelligent life is common, this would increase our credence in the possibility of existential risk in the centuries to come. If we discover that intelligent life is rare, then we would have no reason to be pessimistic—the absence of extraterrestrial civilisations would be explained by the absence of life or intelligent life itself, not some other factor that would prevent humanity from becoming an advanced interstellar civilisation (Bostrom, 2008).

How likely were the conditions and steps that led to intelligent life on Earth? Do we have reason to believe that the Great Filter lies in humanity's past, or does it lie in our future? Chapter 3 and Chapter 4 are in part motivated by this question, and focus primarily on the major evolutionary transitions in Earth's history and the Carter Argument, which I describe below.

The Major Evolutionary Transitions

Life on Earth has undergone a series of major evolutionary transitions (Smith and Szathmary, 1997), many of which were prerequisites for intelligent life. The canonical list of major evolutionary transitions from Szathmáry and Maynard Smith (1995) are:

1. Replicating molecules to populations of molecules in compartments
2. Unlinked replicators to chromosomes
3. RNA as gene and enzyme to DNA and protein (genetic code)
4. Prokaryotes to eukaryotes
5. Asexual clones to sexual populations

6. Protists to animals, plants, and fungi (cell differentiation)

7. Solitary individuals to colonies (non-reproductive castes)

8. Primate societies to human societies (language)

These transitions have many common features. For example, many of the transitions involve changes in how information is stored and transmitted between generations (e.g. the emergence of the genetic code and of sex) (Szathmáry, 2015). Others involve a transition from lower-level units that could previously replicate independently into a higher level organism that can only replicate as a whole (e.g. the emergence of eukaryotes and of multicellular life) (Queller, 2000; West et al., 2015).

Intelligent life may have required other key breakthroughs that are not on the canonical list of evolutionary transitions. For example, oxygen may be a critical prerequisite for any large and complex animal, meaning that intelligent life could require the earlier emergence of oxygenic photosynthesis (Catling et al., 2005; Lenton et al., 2011). Another example of a breakthough not included on the canonical list of evolutionary transitions may be the emergence of vertebrate life on land, since aquatic environments could preclude the acquisition of tools and thus a civilisation (Lingam and Loeb, 2018; Stern, 2016).

How unlikely were these evolutionary breakthroughs? Are any of them candidates for a Great Filter? Some transitions have independently emerged multiple times on Earth, ruling out the possibility that they are exceptionally rare (e.g. multicellular life is thought to have emerged over 40 times independently (Grosberg and Strathmann, 2007)). However, other transitions have only occurred once and took hundreds of millions of years to emerge, suggesting that they could be candidates for a Great Filter. It took 4 billion years (Gyr) for all of the evolutionary transitions to occur, and a line of reasoning known as the Carter Argument suggests that this timescale could be indicative of exceptionally difficult evolutionary transitions.

The Carter Argument

Intelligent life emerged on a timescale similar to that of Earth's lifetime. It took 4 Gyr for intelligent life to emerge, and it will take roughly the same amount of time for the Sun to enter its red giant phase, destroying all life on Earth. In perhaps less than 1 Gyr, the increasing luminosity of the Sun before the red giant phase will still likely destroy the Earth's ability to support complex life, due to increased surface temperatures (Franck et al., 2006) and an eventual breakdown in the carbon cycle (Lenton and von Bloh, 2001). Ultimately, intelligent life emerged on a timescale within an order of magnitude of our star's lifetime. This is puzzling, as the timescales associated with biological and stellar evolution are driven by fundamentally different processes and thus ought to be uncorrelated.

Carter (1983) noticed this coincidence and proposed a resolution to the puzzle based on observation selection effects. Letting $\tau_\odot$ denote the lifetime associated with our star, and τ_ℓ be the timescale it takes for evolution to produce intelligent life, one can analyse three possibilities: $\tau_\odot \gg \tau_\ell$, $\tau_\odot \approx \tau_\ell$, or $\tau_\odot \ll \tau_\ell$, denoting the situations in which the lifetime of the star either greatly exceeds the timescale associated with intelligent life, approximately equals it, or is greatly exceeded by it. Carter argues that *a priori*, the possibility that $\tau_\odot \approx \tau_\ell$ is exceptionally unlikely, leaving $\tau_\odot \gg \tau_\ell$ and $\tau_\odot \ll \tau_\ell$ as realistic alternatives. We can also rule out $\tau_\odot \gg \tau_\ell$ with high probability, given that intelligent life did not emerge exceptionally early when compared to the Sun's lifetime. This brings us to the possibility that $\tau_\odot \ll \tau_\ell$. This would mean that most stars will never support intelligent life, as the star will burn out before such life emerges. However, in the rare locations in which intelligent life does emerge, it will find itself emerging within the lifetime of the star, and moreover is most likely to observe $\tau_\odot \approx \tau_\ell$, consistent with our own observations. Observation selection effects therefore explain why we see these timescales tightly coupled, even if such an outcome is *a priori* unlikely.

In the same paper, Carter proposed a simple model of evolutionary transitions to describe the process of intelligent life emerging. The model proposes that intelligent life requires n 'critical steps,' each of which occurs at some rate λ. He further stipulates that $\lambda^{-1} > \tau_{\odot}$, so that the probability per unit time of the critical step is low enough that the time it takes for each critical step will typically exceed the lifetime of the star. A number of interesting properties follow from this model. First, the probability that the final transition occurs at time t is proportional to t^n, so that the final critical step is likely to occur towards the end of habitable time remaining. Second, the the amount of time remaining will be roughly equal to $\tau_{\odot}/(n+1)$, allowing one to estimate the number of critical steps that occurred in Earth's evolutionary history simply by knowing the amount of time left in Earth's habitable lifetime.

When Carter originally proposed the model, it was thought that the biosphere could last for another 4 Gyr, which in turn suggested that there were likely only one or two critical steps in our evolutionary history. Alternatively, Barrow and Tipler (1986) interpreted this as an argument for why the biosphere would not last for the entire 4 Gyr (an argument I quantify in Chapter 5). Subsequent improvements in climate models led to additional research that suggested that the time remaining is substantially shorter, on the order of 1 Gyr (Caldeira and Kasting, 1992). A number of researchers have returned to Carter's critical step model and re-estimated the number of critical steps predicted by the remaining lifetime of the biosphere. Watson (2008) found that the best fit was with four critical steps, while Carter (2008) suggested between five and six. Waltham (2017) went further to demonstrate that models up to 12 critical steps still fall within a 95% confidence interval.

Beyond estimating the total number of critical steps, other interesting results have emerged from the critical step model. The first is that the spacing of the critical steps will be roughly evenly distributed across the Earth's lifetime (Hanson, 1998*b*; Watson, 2008). This in turn enables one to test whether or not a particular evolutionary transition was a critical step. Watson (2008)

quantifies this by deriving probability density functions for the timing of each transition, and finds that many of the major evolutionary transitions are consistent with being critical steps. Notably, this includes at least one transition associated with the origin of life. Some have argued that the rapid emergence of life suggests that it is a fairly common event (Lineweaver and Davis, 2002), but if a number of critical steps were required afterwards to eventually get intelligent life, we should expect the first critical step (abiogenesis) to occur early (Flambaum, 2003). However, some transitions are not consistent with being critical steps. For example, Watson concludes that three of the first transitions related to abiogenesis could not have all been critical steps, given how closely they occurred. Similarly, Hanson (1998*b*) concludes that despite the staggering increase in biological complexity over the past 500 Myr, there were still only at most one or two critical steps during this time period.

The Chapter 3 and Chapter 4 build heavily on the literature surrounding the Carter Argument. In the Chapter 3, I quantify the Carter Argument in Bayesian terms. Rather than hold λ fixed and estimate n (i.e. estimate the number of critical steps while assuming $\lambda^{-1} \gg \tau_{\odot}$), I hold n fixed and estimate λ (i.e. determine what the timing of each evolutionary transition says about its rate). This has the advantage of quantifying the data, priors, and/or assumptions that would be needed to overturn the Carter Argument. Quantifying the Carter Argument also helps highlight exactly how strong the argument holds. Frank and Sullivan III (2016) argue that as long as the odds that intelligent life emerges on a habitable planet are greater than 1 in 10^{24}, we will not be alone in the observable universe. But I find that for reasonable priors, the Carter Argument places high probability on the odds being less than 1 in 10^{24}. In Chapter 4, I question the extent to which critical steps must be evenly spaced. By allowing for catastrophes to 'undo' critical steps, I find that critical steps can become clustered or occur in an accelerating matter. This in turn opens up questions about whether the studies have overlooked some critical steps.

The Carter Argument has produced three interesting hypotheses that are also discussed in this thesis, all of which are related to the fact that the final transition probability increases as a power law with respect to time. This means that environments with longer habitable lifetimes are substantially more likely to host intelligent life than shorter ones. If we assume that we are typical among the set of intelligent observers (an assumption discussed in more detail in the subsequent section), then we can also infer that we happen to inhabit one of the longest lasting habitable environments[1]. This has bearing on three different hypotheses.

The first hypothesis is that life may have originated on Mars or another body beyond Earth, and was brought to Earth later on. If Mars was habitable for many millions of years before Earth, this hypothesis would be weakly supported by the Carter Argument, as it would create a longer period of time for evolutionary transitions to occur (Davies, 2003; McCabe and Lucas, 2010).

The second hypothesis is that planets orbiting red dwarf stars are inimical environments for intelligent life. Since red dwarf stars last substantially longer than the Sun, and the Carter Argument predicts that longer habitats are much more likely to be the location of intelligent life, then some other factor must be causing such planets to be inhospitable environments (Rushby et al., 2013; Waltham, 2011, 2017).

The third hypothesis is that Earth has remained continuously habitable for substantially longer than one might *a priori* expect. If there is observation selection bias for a long-lasting biosphere, we may in turn be underestimating the rate at which catastrophes could destroy the biosphere, connecting this third hypothesis to our earlier discussion of observation selection effects and

[1] A challenge to this hypothesis is that if the Earth were 16% further away from the Sun, the habitable lifetime of Earth might be extended by about 3 Gyr (Waltham, 2017). This suggests that anthropic selection for long habitable lifetimes cannot be overwhelming (Waltham, 2019), since we would otherwise find ourselves at that maximum point. Either the anthropic selection is modest (i.e. there were a small number of critical transitions), or perhaps other catastrophes could cut the Earth's habitability short before stellar evolution does so (Barrow and Tipler, 1986).

catastrophic risk. Chapter 5 focuses on this question.

There have been some criticisms and proposed alternatives to the Carter Argument. Some have suggested that we may expect $\tau_{\odot}$ to correlate with τ_{ℓ}, for example if a star's energy output affects the buildup of ozone in the atmosphere (Livio, 1999, 2018; Livio and Kopelman, 1990), but these theories have not been fully developed and are only suggestive of possible correlations. An alternative to the Carter Argument that I explore in the Chapter 4 has also been proposed by Ćirković et al. (2009), suggesting that the late emergence of intelligent life was the result of frequent gamma ray bursts. Outright criticism of the Carter Argument is less common, but a notable example exists from Wilson (1994), who rejects the notion that $\tau_{\odot} \approx \tau_{\ell}$ is *a priori* less likely than $\tau_{\odot} \gg \tau_{\ell}$ or $\tau_{\odot} \ll \tau_{\ell}$. I discuss in more detail the appropriate priors in Chapter 3, providing some reasons to believe that Wilson is mistaken.

Anthropic Reasoning

In the previous sections, I made two assumptions. The first is that we should reason as if we are a 'typical observer.' For example, when confronted with a model that implies that the vast majority of intelligent life will emerge around a red dwarf star, one should conclude that the model is likely wrong, rather than conclude that we are one of the exceedingly rare civilisations that happen to not live around a red dwarf star. In doing this, I am implicitly adopting the Self-Sampling Assumption (SSA) from Bostrom (2002*a*), which states:

> SSA: Observers should reason as if they were a random sample from the set of all observers in their reference class.

The second assumption I make is that we can derive no information from the fact of our own existence. In doing this, I am implicitly rejecting something known as the Self-Indication assumption (SIA), which in one formulation from Bostrom and Ćirković (2003) states:

> SIA: Given the fact that you exist, you should (other things being equal) favour hypotheses according to which many observers exist, over hypotheses on which few observers exist.

Our adoption of SSA and rejection of SIA is perhaps the most 'conventional' way forward. SSA is a straightforward extension of the principle of mediocrity (that if an item is drawn at random from one of several categories, it is more likely to come from the most numerous category), and is regularly used in cosmology and astrobiology (although SSA still does have some counterintuitive implications, most notably the Doomsday Argument from Leslie (1990), as well as others from Bostrom (2001)). When using SSA, we must define a reference class, which in this case could simply be the set of observers capable of conducting the kind of analysis in this thesis.

I reject SIA in part because it has deeply counterintuitive implications. For example, SIA essentially guarantees that the universe is exceptionally large or infinite even without any cosmological evidence (Bostrom and Ćirković, 2003), and will also give substantial probability to any bizarre theory, so long as it proposes a large enough population of observers to overwhelm the *a priori* implausibility of the theory (for example a theory that each planet has $10^{10^{100}}$ copies of itself on 'other planes' would seem hard to refute if one adopted SIA (Olum, 2002)). However, it is still worth noting that adopting SIA will undermine the Carter Argument, and therefore undermine much of the analysis in Chapter 3 and Chapter 4 of this thesis. This is because SIA will strongly favour the hypothesis that $\tau_{\odot} \approx \tau_{\ell}$ and that there were no critical steps at all, since these parameters result in much more intelligent life, and thus many more observers. I address this directly in Chapter 3 and examine how adopting SIA would change our results.

Roadmap

Chapter 2 places an upper bound on the background rate of human extinction. Using fossil record data on the age of *Homo sapiens* and a simple model

of 'observerhood,' I conclude A) that the risk of naturally occurring human extinction is low and B) that observation selection bias is unlikely to play a substantial role in these risk estimates.

Chapter 3 quantifies the Carter Argument and estimates the rates of key evolutionary transitions in a Bayesian framework. I find that the timing of key evolutionary transitions when combined with an uninformative prior supports the Carter Argument that the timing of intelligent life is due to anthropics and conclude that intelligent life elsewhere in the observable universe is rare.

Chapter 4 generalises the Carter Model and allows for catastrophes to undo evolutionary transitions. I find that under certain plausible conditions, the past literature based on the Carter Argument could be underestimating the number of critical steps needed for intelligent life.

Chapter 5 places an upper bound on the rate of more severe catastrophes that would entirely sterilise Earth's biosphere. As with Chapter 2, I conclude that the risk is low, but that observation selection effects could still cause us to inhabit an unusually long lasting biosphere.

Chapter 6 is a discussion section that offers some comments concerning the general principles and themes that recur in this work.

Chapter 2

An upper bound for the background rate of human extinction

Andrew E. Snyder-Beattie, Toby Ord, Michael B. Bonsall

Abstract

We evaluate the total probability of human extinction from naturally occurring processes. Such processes include risks that are well characterised such as asteroid impacts and supervolcanic eruptions, as well as risks that remain unknown. Using only the information that *Homo sapiens* has existed at least 200,000 years, we conclude that the probability that humanity goes extinct from natural causes in any given year is almost guaranteed to be less than one in 14,000, and likely to be less than one in 87,000. Using the longer track record of survival for our entire genus *Homo* produces even tighter bounds, with an annual probability of natural extinction likely below one in 870,000. These bounds are unlikely to be affected by possible survivorship bias in the data, and are consistent with mammalian extinction rates, typical hominin species lifespans, the frequency of well-characterised risks, and the frequency of mass extinctions. No similar guarantee can be made for risks that our ancestors did not face, such as anthropogenic climate change or nuclear/biological warfare.

Introduction

Out of all species that have existed, over 99% are now extinct (Raup, 1992). Although human activity is dramatically increasing extinction rates for many species (Barnosky et al., 2011), species extinctions were regular occurrences long before humanity emerged. Many of these extinctions were caused by gradual environmental shifts, evolutionary arms races, or local interspecific competition (Benton, 2009; Smith, 1989), while others were abrupt, being part of global mass extinctions caused by asteroid impacts, volcanism, or causes as of yet to be identified (Schulte et al., 2010; Wignall, 2001). Could such a catastrophe befall our own species? If so, are the risks greater from natural or anthropogenic sources?

Here, we evaluate the natural 'background' extinction rate for *Homo sapiens*. This means considerations of anthropogenic risks such as climate change and nuclear weapons are excluded from our estimates, although these clearly pose existential threats to our own species as well as others. Indeed, it has been hypothesised that the great majority of human extinction risk comes from anthropogenic sources (Bostrom, 2013*b*; Sandberg, 2018). But by limiting our analysis to natural risks that our predecessors also faced, we can draw on data spanning many thousands (or millions) of years. Obtaining bounds on natural extinction rates also enables an indirect and partial test of the hypothesis that anthropogenic risks are greater than natural ones, as sufficiently low natural extinction risk will imply higher relative risks from anthropogenic sources.

Estimating such an extinction rate directly is impossible. We have no examples of *Homo sapiens* extinction, so the most directly relevant data are non-existent. An alternative approach would be to enumerate the different types of naturally occurring hazards (e.g. asteroids, supervolcanoes), estimate their independent probability of causing extinction, and then use these probabilities to derive an aggregate extinction rate. However, this method has its own shortcomings. Beyond the great uncertainties around the probabilities of

each risk, there could also be unknown risks that fail to be included. It would be hard to say with confidence that any list of risks had captured all natural hazards to humanity.

We can bypass these problems by instead considering the length of time that humanity has survived so far (Gott III, 1993; Pisaturo, 2009). This survival time can be used to estimate an upper bound on the extinction rate from all natural sources combined, including from sources for which we remain unaware. However, this approach could be subject to a particular form of sample bias known as an observation selection bias. These observer selection biases occur when a sample is not representative of all outcomes, but rather a subset of outcomes that are compatible with the existence of the observers (Bostrom, 2013*a*). For example, if human existence required a 10 million year (Myr) period of evolution free from asteroid impacts, any human observers will necessarily find in their evolutionary history a period of 10 Myr that is free of asteroid impacts, regardless of the true impact rate. Inferring a rate based on those 10 Myr could therefore be misleading, and methods must to be used to correct for this bias (Ćirković et al., 2010).

Using data from archeological and fossil records, we place an upper bound on the natural rate of human extinction. We then test this model against possible forms of observer selection bias, and demonstrate that the data are unlikely to be severely biased due to these effects. We finally cross-check our conclusions against alternative forms of data, including mammalian extinction rates, the temporal ranges of other hominin species, and the frequency of potential catastrophes and mass extinctions.

Bounding the extinction rate based on age of humanity

Anatomically modern human fossils in Ethiopia have been dated to 195 ± 5 thousand years ago (kya) (McDougall et al., 2005). A more recent fossil

discovery in Morocco of an anatomically modern human has been dated to 315 ± 34 kya (Hublin et al., 2017; Richter et al., 2017) (though the fossil may exhibit more primitive neurocranial and endocranial morphology). Given that *Homo sapiens* has existed for hundreds of thousands of years, what can we infer about our background rate of extinction?

Assuming that we share a common extinction rate with our predecessors, we can rule out rates that are too high to be compatible with this track record of survival. As our aim is to construct an upper bound, we can set aside the possibility that modern human technology, habitat range, and population size have reduced a number of natural extinction risks. The upper bound is only violated if we have reason to believe current extinction rates are higher than those our predecessors faced. Since we exclude anthropogenic risks from our analysis, we also set aside the majority of the ways in which this could be the case, although we acknowledge there exist boundary cases between purely natural and anthropogenic risks (e.g. a naturally emerging disease could be spread further by modern technology). Ultimately the scope of the upper bound is limited to all risks that have remained constant (or have been reduced) over the past few hundred thousand years.

Likelihood of extinction rates

Analysis of taxonomic survivorship curves and temporal ranges for a wide variety of taxa suggest that extinction probabilities can be approximated well by assuming a constant risk of extinction over time (Alroy, 1996; Foote and Raup, 1996; Van Valen, 1973). Under this model, extinction can be represented by the exponential distribution with constant extinction rate μ. The probability that humanity goes extinct before time t is given by the cumulative distribution function $P(T \leq t) = 1 - e^{-\mu t}$, where T is the random variable denoting the longevity of our species. Conversely, the probability that humanity makes it beyond time t is $P(T \geq t) = e^{-\mu t}$.

We want to evaluate the likelihood of an extinction rate μ, given the observation that humanity has lasted up to time t (so we know that the total longevity of humanity $T \geq t$). This can be evaluated as the likelihood function $\mathcal{L}(\mu|T \geq t) = e^{-\mu t}$. We compute the likelihood of extinction rates between 10^{-8} and 10^{-4} per year given a number of different plausible starting dates for *Homo sapiens* outlined in Figure 2.1 and Table 2.1.

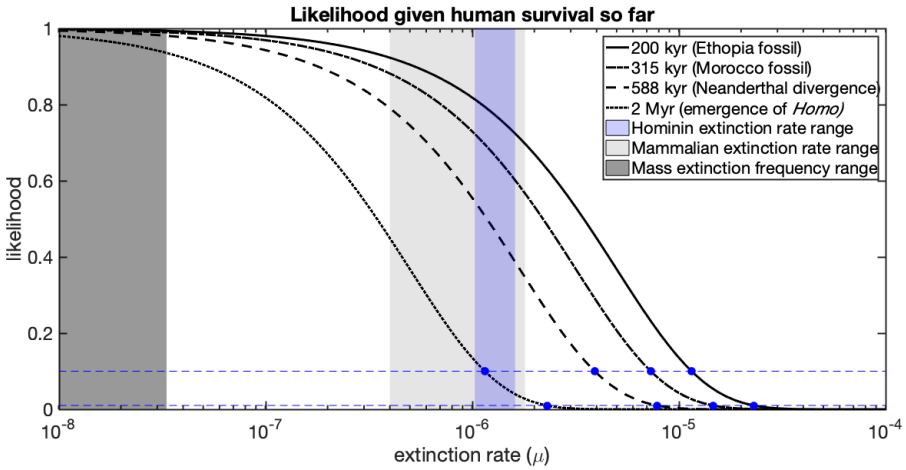


Figure 2.1: Likelihood of extinction rates given our track record of survival so far, with estimated ranges of Hominin extinction rates, mammalian extinction rates, and mass extinction frequency included for reference. Blue horizontal lines indicate likelihood of 10% and 1%. Rates exceeding 6.9×10^{-5} per year are ruled out even with the most conservative data. Extending humanity's track record of survival to match older fossils, the divergence with *Homo neanderthalensis*, or the origin of *Homo* creates even stricter bounds.

Table 2.1: Survival times and resulting upper bounds (risk per year)

Track Record (t)	Starting Point	Value of μ with relative likelihood of... 10^{-1}	10^{-2}	10^{-6}
200 kyr	Anatomically modern humans McDougall et al. (2005)	1.2×10^{-5}	2.3×10^{-5}	$\mathbf{6.9 \times 10^{-5}}$
315 kyr	Anatomically modern humans (Hublin et al., 2017; Richter et al., 2017)	7.3×10^{-6}	1.5×10^{-5}	4.4×10^{-5}
588 kyr	Neanderthal divergence (Mendez et al., 2016; Meyer et al., 2016)	3.9×10^{-6}	7.8×10^{-6}	2.4×10^{-5}
2 Myr	Emergence of *Homo* (Robinson et al., 2018; Wood and K. Boyle, 2016)	$\mathbf{1.2 \times 10^{-6}}$	2.3×10^{-6}	6.9×10^{-6}

Assuming a 200 thousand year (kyr) survival time, we can be exceptionally confident that rates do not exceed 6.9×10^{-5} per year. This corresponds to an annual extinction probability below roughly 1 in 14,000. The relative likelihood

for such high extinction rates are below 10^{-6} (one in a million) per year when compared to a rate of 10^{-8} per year. If we assume that our track record extends further, this upper bound becomes stronger. Using the fossil dated to 315 kya as a starting point for humanity gives an upper bound of $\mu < 4.4 \times 10^{-5}$ per year, corresponding to an annual extinction probability below 1 in 22,800. Using the emergence of *Homo* as our starting point pushes the initial bound back a full order of magnitude, resulting in an annual extinction probability below 1 in 140,000.

We can also relax the one in million relative likelihood constraint and derive less conservative upper bounds. An alternative bound would be rates with relative likelihood below 10^{-1} (1 in 10) when compared to the baseline rate of 10^{-8}. If we assume humanity has lasted 200 kyr, we obtain a bound of $\mu < 1.2 \times 10^{-5}$ per year, corresponding to an annual extinction probability below 1 in 87,000. Using the 2 Myr origin of *Homo* strengthens the bound by an order of magnitude in a similar way and produces annual extinction probabilities below 1 in 870,000.

Observation selection effects

The data on humanity's survival time could be subject to survivorship bias. If early *Homo sapiens* requires a long period of time to develop the intellectual machinery needed to make scientific observations, then such observations could not include short evolutionary histories, regardless of the extinction rate. The amount of information we could derive from a long track record of survival would therefore be limited due to this observation selection effect. Such a track record could indicate a low extinction rate, or be the byproduct of lucky ancestors surviving high extinction rates long enough to beget progeny capable of making scientific observations. One might therefore object that the bounds on the extinction rate we have estimated are too low (Ćirković et al., 2010; Manheim, 2018). Here, we examine and respond to this concern.

Models to quantify potential sample bias

To model observation selection bias, let us assume that after *Homo sapiens* first arises another step must be reached. This could represent the origin of language, writing, science, or any relevant factor that would transition early humans into the reference class of those capable of making observations[†] (we call this step 'observerhood'). Let this step be a random variable denoted S, with cumulative distribution function $F_S(t)$. As we are examining natural risks, we assume that S and T are independent. The probability that humanity survives long enough to reach observerhood status (via intelligence, language, writing, science, etc) can be found with the following integral:

$$P(T > S) = \int_0^\infty f_T(t) F_S(t) dt \tag{2.1}$$

where $f_T(t) = \mu e^{-\mu t}$, the probability of extinction at time t. We evaluate an adjusted likelihood function $\mathcal{L}^\star(\mu | T > t)$, denoting that we are taking the likelihood of an extinction rate μ given that humanity has survived to time t, and the fact that we are conditioning on the existence of observers such that $T > S$. This results in the adjusted likelihood function:

$$\mathcal{L}^\star(\mu | T > t) = P(T > t | T > S, \mu) \tag{2.2}$$

$$= \frac{1}{c} \int_t^\infty f_T(\xi) F_S(\xi) d\xi \tag{2.3}$$

where $c = P(T > S)$ is a normalising constant and ξ is an integration variable. We evaluate a model with four variations for the observerhood step: a model in which observerhood occurs as a single event that has a constant rate over

[†]We could more crisply define this observerhood property as the ability for a species to collect reliable data on its own track record of survival (e.g. via fossil dating). When correcting for observation selection effects, we are simply conditioning on the fact that our species has developed the ability to conduct this analysis. The observerhood property need not invoke consciousness or be the property of a biological species—a machine estimating a parameter would need to account for observer selection bias if its ability to make such estimates were correlated with the parameter in question.

time, a model with an increasing rate over time, a model with multiple steps, and a model where observerhood simply requires a fixed amount of time.

Model 1: Single step, constant rate

Our first model assumes that observerhood has a constant rate of occurrence θ, so that S is exponentially distributed with cumulative distribution function: $F_S(t) = 1 - e^{-\theta t}$. This model describes a process in which the transition from early humans into observers occurs by chance as a single step. This could represent the hypothesis that hierarchical language emerged in humans as the byproduct of a chance mutation (Bolhuis et al., 2014). With this model, the probability that observers arrive before extinction is $P(T > S) = \theta(\theta + \mu)^{-1}$. Our likelihood function can be analytically derived:

$$\mathcal{L}^\star(\mu|T > t) = \left(\frac{\theta + \mu}{\theta}\right) \int_t^\infty \mu e^{-\mu\xi}(1 - e^{-\theta\xi})d\xi \tag{2.4}$$

$$= \left(\frac{\theta + \mu}{\theta}\right) e^{-\mu t} - \left(\frac{\mu}{\theta}\right) e^{-(\mu+\theta)t} \tag{2.5}$$

Model 2: single step, increasing rate

Our second model similarly assumes that a single step is needed but that the rate of observerhood increases over time. This model could represent increasing population size or population density, which could in turn drive cultural evolution and increase the probability of such a step (Powell et al., 2009). We represent this with a Weibull distribution with cumulative distribution function $F_S(t) = 1 - e^{-(\theta t)^k}$ where $k > 1$ indicates increasing rate over time (when $k = 1$, this is the same as the exponential in Model 1). We use numerical integration to evaluate the likelihood function.

Model 3: multiple steps, constant rate

Our third model assumes that there are multiple steps that need to occur in a sequence in order to get observers. This could represent more incremental

development of tools, culture, or language. We assume that each step is exponentially distributed with rate θ, so that the timing of the final kth step follows an Erlang distribution with cumulative distribution function:

$$F_S(t) = 1 - \sum_{n=0}^{k-1} \frac{1}{n!} e^{-\theta t} (\theta t)^n. \tag{2.6}$$

Note that when $k = 1$, the distribution is the same as the exponential in Model 1. We use numerical integration to evaluate the likelihood function.

Model 4: fixed time requirement

Our final model assumes that it takes a fixed amount of time τ to reach observerhood. This is an extreme model that allows for no chance, but could represent a gradual and deterministic accumulation of traits. The probability that observerhood has been reached before time t is therefore $F_S(t) = \mathbb{1}_{[t>\tau]}$, the characteristic function that takes the value 1 when $t > \tau$ and 0 otherwise. The probability that humanity survives past time τ is $1 - F_T(\tau) = e^{-\mu\tau}$. Our likelihood function of μ is:

$$\mathcal{L}^\star(\mu | T > t) = \frac{1}{e^{-\mu\tau}} \int_t^\infty \mu e^{-\mu\xi} \mathbb{1}_{[\xi>\tau]} d\xi \tag{2.7}$$

$$= e^{-\mu(t-\tau)}. \tag{2.8}$$

This likelihood expression can also be derived using the memoryless property of the exponential. It is worth noting that the fixed time model is a limiting case for both the increasing rate model and the multiple steps model. Taking the limit of Model 2 as $k \to \infty$ results in a fixed time model with $\tau = \theta^{-1}$. Similarly, Model 3 converges to a fixed time model as the number of steps increases and the expected time of each step decreases (having infinitely many steps in the limit, each of which is infinitely short).

Results of sample bias models

We evaluate the likelihood of extinction rates between 10^{-8} and 10^{-2} per year, given a human survival time of 200 kyr and a wide range of different rates at which observers could originate (Figure 2.2). The first thing to note about the first three models is that when the observerhood rates are sufficiently rapid, the likelihood function converges to the unbiased version in the previous section. This can be verified by taking limits: for all of the models as $\theta \to \infty$ (or $\tau \to 0$ in the case of the fixed time model), $\mathcal{L}^\star(\mu|T > t) \to e^{-\mu t}$. If observerhood is expected to occur quickly, then we can take a 200 kyr track record of survival at face value and estimate the extinction rate without observation selection bias.

However, as the observerhood rates decrease to the point where the expected observerhood time approaches an order of magnitude close to 200 kyr, observer selection bias emerges. Rates that were previously ruled out by our track record of survival are assigned higher likelihoods, since a portion of the track record is a necessity for observers (Figure 2.2). For example in Model 1, when $\theta = 2 \times 10^{-4}$ per year (corresponding to an expected observerhood time of 20 kyr), the relative likelihood of $\mu = 6.9 \times 10^{-5}$ per year is increased by a factor of 2.3 (from 10^{-6} to 2.3×10^{-6}). To get a likelihood of 10^{-6} (corresponding to the most conservative upper bound), the rate must be set at 7.3×10^{-5} per year (see all edited bounds in Table 2.2). Interestingly though, this effect is limited. Even as observerhood rates slow to the point where expected observerhood time greatly exceeds 200 kyr (for example exceeding 20 billion years), the revised upper bounds remain within a factor of 2 of the original bounds. The stricter the bound, the weaker the potential bias: for example the 10^{-6} likelihood bound is only changed by a factor of about 1.2 in the limit as $\theta \to 0$. Although there would be some sample bias, there is a hard ceiling on how much our track record of survival can be distorted by observation selection effects.

Table 2.2: Upper bounds of μ with model 1 bias (in risk per year)

θ	$\mathcal{L} = 10^{-1}$	$\mathcal{L} = 10^{-2}$	$\mathcal{L} = 10^{-6}$
Unbiased	1.2×10^{-5}	2.3×10^{-5}	6.9×10^{-5}
2×10^{-4}	1.3×10^{-5}	2.5×10^{-5}	7.3×10^{-5}
2×10^{-5}	1.7×10^{-5}	3.1×10^{-5}	8.1×10^{-5}
$\theta \to 0$	1.9×10^{-5}	3.3×10^{-5}	8.3×10^{-5}

The reason slow rates of observerhood have a limited impact on our estimates is because if the extinction rate were exceptionally high, the lucky humans that do successfully survive to observerhood will have achieved such a status unusually quickly, and therefore will still observe a very short track

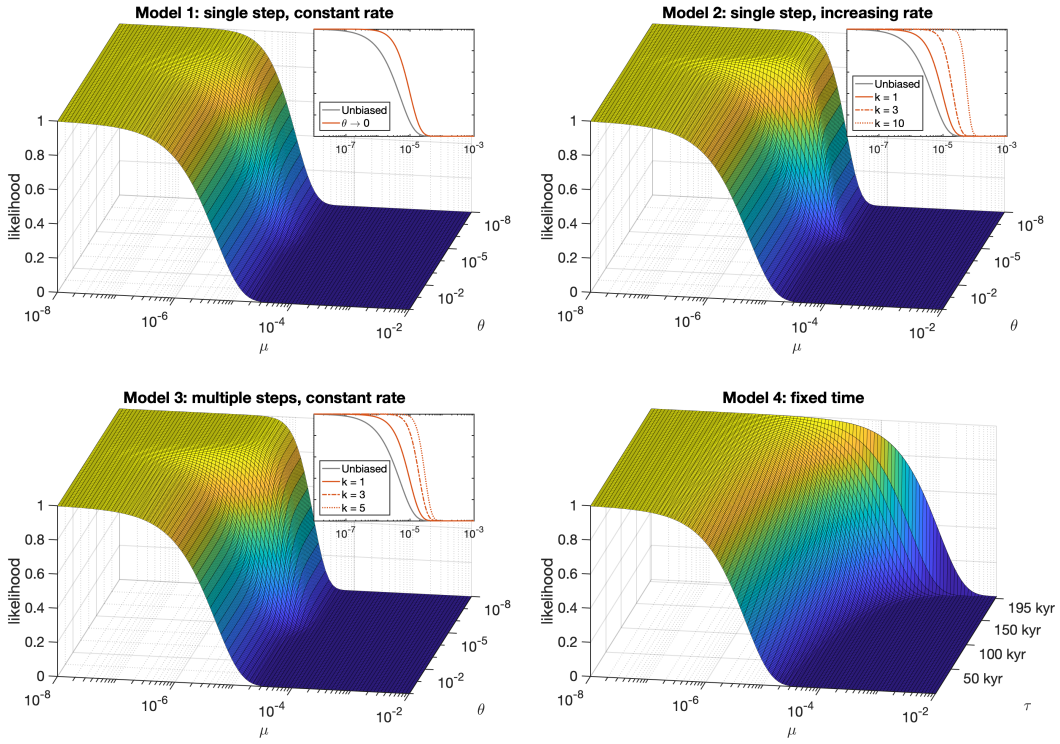


Figure 2.2: Models of observer selection bias. Surface plots show likelihood for combinations of μ and θ (where $k = 3$ for Models 2 and 3) or τ in Model 4. Upper righthand plots show how likelihood shifts when $\theta \to 0$ in Model 1, and for a variety of k values in Models 2 and 3. For the first three models, the unbiased model is recovered for large θ, and results start to become biased as the expected observerhood time approaches humanity's track record of survival. However, even as $\theta \to 0$, the bias is limited, and the likelihood of rates exceeding 10^{-4} per year remains at zero. This is only violated in the final fixed time model, or in models 2 and 3 when k is sufficiently large.

record of survival. A long track record of survival is therefore still sufficient to rule out high extinction rates paired with low observerhood rates. We can demonstrate this by examining the typical time it takes for lucky survivors to reach observerhood, assuming a high extinction rate and a low observerhood rate. For example, in the single step constant rate model when $\theta = 10^{-6}$ per year (corresponding to an expected observerhood time of 1 Myr) and $\mu = 10^{-3}$ per year (corresponding to a typical extinction time of 1000 years), the expected observerhood time conditional on these high extinction rates is 1000 years. A typical observer will thus still have a very short track record of survival. Models with increasing rates or multiple steps exhibit the same property, although the bias is larger depending on parameter k. For both model 2 and 3 with $\theta = 10^{-6}$ per year, $\mu = 10^{-3}$ per year, and $k = 2$ (parameters normally corresponding to an expected observerhood time of 830 kyr for Model 2 and 2 Myr for model 3), the high extinction rates will still result in a typical observer emerging unusually early and having only about a 2000 year track record of survival. This can be also seen in Figure 2.2 where for Models 1, 2, and 3, the likelihood of high extinction rates exceeding 10^{-4} per year are still assigned low likelihood regardless of θ.

However, severe observer selection bias can occur in Models 2 and 3 as k becomes larger, shaping the observerhood distribution such that early observerhood is vanishingly unlikely and late observerhood almost guaranteed. In the most extreme case this is represented by the fixed time model, where the probability of observerhood jumps from 0 to 1 when $t = \tau$ (the fixed time model is also the limiting case when $k \to \infty$). If that fixed amount of time is long enough (say, exceeding 190 or 195 kyr), a 200 kyr track record of survival is no longer sufficient to rule out extinction rates greater than 10^{-4} per year. This result occurs as the fixed time model prohibits any possibility of observerhood occurring unusually quickly. Any lineage of *Homo sapiens* lucky enough to survive long enough to obtain observer status must necessarily have a survival

time greater than τ, which means that being an observer with a survival time of τ conveys zero information about the extinction rate.

For numerous reasons, we find the fixed time model to be implausible. Virtually all biological and cultural processes involve some degree of contingency, and there is no fundamental reason to think that gaining the ability to make scientific observations would be any different. To illustrate a comparison, let us consider a world in which the extinction rate is 10^{-4} per year (averaging one extinction every 10,000 years), but observerhood status takes a fixed 200 kyr. Under this model, humanity successfully surviving long enough to reach observer status is an event with 1 in 200 million chance. Given observation selection bias, we cannot rule out the possibility of rare events that are required for our observations. But we could ask why a 1 in 200 million chance event could not also include the possibility that modern human observers would emerge unusually rapidly. Language, writing, and modern science are perhaps highly unlikely to develop within ten thousand years of the first modern humans, but it seems exceptionally overconfident to put the odds at fewer than 1 in 200 million.

A similar line of reasoning can be applied to determine whether the increasing rate and multiple step models with high k are reasonable. We test this by asking what parameters would be needed to expect a 200 kyr track record of survival with an extinction rate at our conservative upper bound of $\mu = 6.9 \times 10^{-5}$ per year. For the increasing rate model, observerhood is expected after 203 kyr with $\theta = 10^{-7}$ per year and $k = 14$ and for the multiple step model, observerhood is expected after 190 kyr with $\theta = 10^{-7}$ per year and $k = 16$. Although these models do not assign strictly zero probability to early observerhood times, the probabilities are still vanishingly small. With an increasing rate and these parameters, observerhood has less than a one in a trillion chance of occurring within 10,000 years (3.4×10^{-14} per year), and about 1% chance of occurring within 100,000 years. With multiple steps and these parameters, observerhood has less than one in a trillion chance of

occurring within 10,000 years (5.6×10^{-17} per year), and less than a 0.02% chance of occurring within 100,000 years. In a similar fashion to the fixed time model, we feel that these models exhibit unrealistic levels of confidence in late observerhood times.

Although the plausibility of the fixed time (or nearly fixed time) models is hard to test directly, the wide variance in the emergence of modern human behavior across geography offers one source of data that can test their plausibility. The Upper Palaeolithic transition occurred about 45 kya in Europe and Western Asia, marked by the widespread emergence of modern human behaviour (Powell et al., 2009) (e.g. symbolic artwork, geometric blades, ornamentation). But strong evidence exists for the sporadic appearance of this modern human behaviour much earlier in parts of Africa (Henshilwood et al., 2002; McBrearty and Brooks, 2000), including evidence of artwork and advanced tools as early as 164 kya (Marean et al., 2007). Although numerous factors could have prevented the Upper Palaeolithic transition from occurring quickly, the fact that some human communities made this transition more than 100 kyr earlier than the rest of humanity indicates that a much earlier development trajectory is not entirely out of the question.

In summary, observer selection effects are unlikely to introduce major bias to our track record of survival as long as we allow for the possibility of early observers. Deceptively long track records of survival can occur if the probability of early observers is exceptionally low, but we find these models implausible. The wide variance in modern human behavior is one source of data that suggests our track record is unlikely to be severely biased. We can also turn to other sources of indirect data to test for observer selection bias.

Testing the bound with indirect data

We cross check our upper bound against four other sources of data: mammalian extinction rates, survival times of other human species, rates of potential catas-

trophes, and mass extinction rates. Although these alternative data do not directly predict the background extinction rate of *Homo sapiens* per se, the rates of extinction are likely generated by similar processes and thus enable an indirect test of the upper bound. If our upper bound is sound (not biased by observer selection effects or otherwise flawed), we can make testable predictions that it will be A) broadly consistent with the extinction rates for similar species, and B) not exceeded by the rate of potential catastrophes or mass extinctions. As the extinction rate of other species and catastrophes many millions of years ago have little bearing on our ability to make scientific observations, these data are also less subject to potential observer selection bias.

Mammalian extinction rates

We first evaluate whether the upper bound is consistent with extinction rates for a typical mammalian species. Using fossil record data, median extinction rates for mammals have been estimated as high as 1.8 extinctions per million species years (E/MSY) (Barnosky et al., 2011), or equivalently $\mu = 1.8 \times 10^{-6}$ per year. Other estimates using fossil record data range from 0.165 extinctions per million genus years (Alroy, 1996) to 0.4 E/MSY for Cenozoic mammals (Foote and Raup, 1996). Alternative methods using molecular phylogeny suggest a much lower rate of 0.023 E/MSY for mammals (De Vos et al., 2015) and rates of 0.219-0.359 E/MSY for primates (Arbour and Santana, 2017), although these methods have been criticised (Rabosky, 2010). All of these estimated background rates are consistent with our upper bound. It is worth noting that *Homo sapiens* may be at lower extinction risk than a typical mammalian species due to a large habitat range, large population size, and having a generalist diet, which are all traits that militate against extinction risk (whereas long generation times and large body mass are sometimes correlated with increased extinction risk) (Fritz et al., 2009; Purvis et al., 2000).

Hominin survival times

Next, we evaluate whether the upper bound is consistent with the broader hominin fossil record. There is strong evidence that *Homo erectus* lasted over 1.7 Myr and *Homo habilis* lasted 700 kyr (Wood and K. Boyle, 2016), indicating that our own species' track record of survival exceeding 200 kyr is not unique within our genus. Fossil record data indicate that the median hominin temporal range is about 620 kyr, and after accounting for sample bias in the fossil record this estimate rises to 970 kyr (Robinson et al., 2018). Although it is notable that the hominin lineage seems to have a higher extinction rate than those typical of mammals, these values are still consistent with our upper bound. It is perhaps also notable that some hominin species were likely driven to extinction by our own lineage (Banks et al., 2008), suggesting an early form of anthropogenic extinction risk.

Individual sources of extinction risk

The upper bound can also be evaluated against the frequency of events that could pose extinction risks. If any particular risk (such as those from asteroid impacts) is known to have a higher rate than our bound of 6.9×10^{-5} per year, this could undermine and potentially falsify our hypothesis. We evaluate the frequencies of four types of potential disasters for which credible quantitative estimates exist: asteroid impacts, supervolcanic eruptions, stellar explosions, and vacuum collapse. All of these risks have been estimated to occur with a frequency well below our bound (Table 2), with the exception of smaller supervolcanic eruptions. Recent work has suggested the frequency of eruptions ejecting $> 10^4$ km^3 of material exceeds our upper bound of 6.9×10^{-5} per year with a recurrence time of 17 kyr (Rougier et al., 2018).

However, it is important to note that the smaller eruptions within this category do not necessarily have a high probability of causing human extinction. The most severe eruption of the past 2 million years occurred just 74 kya, and

Table 2.3: Catastrophe frequency estimates

Risk	Recurrance time (source)	Frequency
Asteroid $\geq$ 1km	500 kyr (Napier, 2008)	2×10^{-6}
Asteroid $\geq$ 5km	6 Myr (Chapman and Morrison, 1994)	1.7×10^{-7}
Supervolcano 10^4km^3	1.1 Myr (Mason et al., 2004)	9×10^{-7}
Supervolcano 10^4km^3	17 kyr (Rougier et al., 2018)	5.9×10^{-5}
Supervolcano 10^6km^3	32 Myr (Rampino and Stothers, 1988)	3.1×10^{-8}
Gamma Ray Burst	170 Myr (Melott et al., 2004)	5.9×10^{-9}
Supernovae	100 Myr (Beech, 2011)	1×10^{-8}
Vacuum collapse	1 Gyr (Tegmark and Bostrom, 2005a)	$< 1 \times 10^{-9}$

the existing humans in Africa and India survived seemingly unscathed (Smith et al., 2018). Moreover, most members of the Hominidae family did not suffer population bottlenecks around the time, with the possible exception of Eastern chimpanzees and Sumatran orangutans (Prado-Martinez et al., 2013). The lack of dramatic evidence suggesting other species extinctions or bottlenecks also undercuts the possibility that humanity's survival was highly improbable and is observed only due to observation selection effects. However, a handful of substantially larger flood basalt events have taken place over the past 250 Myr, some of which have been linked to mass extinctions (Bond and Wignall, 2014; Rampino and Stothers, 1988). These events occur with a frequency of roughly once every 20-30 Myr, much more infrequently than smaller eruptions. If we assume that human extinction is threatened only from larger volcanic eruptions well exceeding 10^4 km^3, then none of the risk frequencies we have catalogued come within an order of magnitude of the conservative upper bound.

Similarly, impacts from smaller asteroid around 1km in diameter may not have a high probability of causing human extinction. Although it is hard to estimate the consequences of such impacts, some researchers have argued that such impacts would fall below the threshold for a global catastrophe (Toon et al., 1997). Impacts that disperse enough dust and sulphites to significantly disrupt photosynthesis occur much more rarely, with an estimated frequency of about 15 Myr years (Toon et al., 1997). If we assume human extinction is

only threatened by these more severe impacts exceeding 5 km, each of these catastrophe frequencies falls well below even our most optimistic bound of 1 in 870,000 chance of extinction per year.

Mass extinction frequency

A mass extinction is marked by substantially increased extinction of multiple geographically widespread taxa over a relatively short period of time (Sepkoski, 1986). There have been five major mass extinctions in the past 541 Myr (Jablonski, 1994; Raup and Sepkoski, 1982), with many arguing that human activity is currently causing a sixth (Barnosky et al., 2011). In a similar way to our previous analysis of catastrophe rates, we should expect our upper bound to be consistent with the frequency of non-anthropogenic mass extinctions. Using only the big five extinctions produces a frequency of less than one per 100 Myr, far below our upper bound. In addition to the big five, there have been 13 other mass extinctions in the fossil record (Bambach, 2006). Using these numbers for 18 mass extinctions over 541 Myr still results in a frequency of about one per 30 Myr, many orders of magnitude below our upper bound.

Conclusions

Using the fact that humans have survived at least 200 kyr, we can infer that the annual probability of human extinction from natural causes is less than 1 in 87,000 with modest confidence (0.1 relative likelihood) and less than 1 in 14,000 with near certainty (10^{-6} relative likelihood). These are the most conservative bounds. Estimates based on older fossils such as the ones found in Morocco dated to 315 kya result in annual extinction probabilities of less than 1 in 137,000 or 1 in 23,000 (for relative likelihood of 0.1 and 10^{-6}, respectively). Using the track record of survival for the entire lineage of *Homo*, the annual probability of extinction from natural causes falls below 1 in 870,000 (relative likelihood of 0.1). We also conclude that these data are unlikely to be biased by

observer selection effects, especially given that the bounds are consistent with mammalian extinction rates, the temporal range of other hominin species, and the frequency of potential catastrophes and mass extinctions.

The bounds are subject to important limitations. Most importantly, they only apply to extinction risks that have either remained constant or declined over human history. Our 200 kyr track record of survival cannot rule out much higher extinction probabilities from modern sources such as nuclear weapons or anthropogenic climate change. Some naturally occurring risks could be also be worsened by anthropogenic factors: a minor asteroid impact could be interpreted as a nuclear attack and lead to retaliation (Baum, 2018), or a naturally occurring disease which previously may have only been a local extinction risk could spread much further due to modern travel (Manheim, 2018). In the cases where a natural risk is amplified by modern conditions, we can still derive some partial information from the upper bound by evaluating how much the risk would need to change from the purely natural baseline. For example, the claim that a natural disease poses a greater than 1 in 1,000 chance of extinction per year would require that anthropogenic conditions have increased the risk of natural disease by a factor of more than 14 to 870 (under our most conservative and optimistic upper bounds, respectively). In general, for a naturally occurring risk to violate our upper bounds via human activity by more than a factor of two, the majority of the risk would still need to come from anthropogenic circumstances.

In general, we conclude that anthropogenic extinction risks are likely greater than natural ones. We do not have a long track record of data for anthropogenic risks, so evaluating this relies far more on speculation. But despite the paucity of data, the little evidence we do have seems to be indicative of rates greatly exceeding our upper bounds. During the Cuban Missile Crisis of 1962, John F Kennedy put the odds of nuclear war at 'somewhere between one out of three and even' (Kennedy, 2011). If 0.1% of nuclear wars result in human extinction via nuclear winter, taking Kennedy's odds that year would surpass

our most conservative bound by more than a factor of 4 (and surpass our most optimistic bound by a factor of more than 250). Anthropogenic climate change could pose existential risks as well if warming is much worse than expected. A ballpark suggestion for the probability of 20 degrees of anthropogenic climate change was placed at 1% (Weitzman, 2009), which would make the planet largely uninhabitable for humans due to heat stress (Sherwood and Huber, 2010). And these are not the only risks we may face. One century ago, the existential risks posed by nuclear weapons or climate change may have seemed extremely implausible. We should therefore be cautious before dismissing the potential risks that future centuries of technological development could bring, such as those stemming from biotechnology (Millett and Snyder-Beattie, 2017) or artificial general intelligence (Bostrom, 2014).

Despite the low probability of human extinction from natural causes, it may still be prudent to reduce these risks. Existential risks jeopardise not only the lives of those currently present, but also the existence of all future generations. Depending on how much value we place on such generations, it may still be cost-effective to reduce existential risks from natural sources (Matheny, 2007). However, given limited resources to spend on reducing existential risks, one may be better off focusing on greater risks from our own design.

Chapter 3

The timing of evolutionary transitions suggests intelligent life is rare

Abstract

It is unknown how abundant extraterrestrial life is, or whether such life might be complex or intelligent. On Earth, the emergence of complex intelligent life required a preceding series of evolutionary transitions such as abiogenesis, eukaryogenesis, and the evolution of sexual reproduction, multicellularity, and intelligence itself. Some of these transitions could have been extraordinarily improbable, even in conducive environments. The emergence of intelligent life late in Earth's lifetime is thought to be evidence for a handful of rare evolutionary transitions, but the timing of other evolutionary transitions in the fossil record has yet to be analysed in a similar framework. Using a simplified Bayesian model that combines uninformative priors and the timing of evolutionary transitions, we demonstrate that expected evolutionary transition times likely exceed the lifetime of Earth, perhaps by many orders of magnitude. Our results corroborate the original argument suggested by Brandon Carter that intelligent life in the universe is exceptionally rare, assuming that intelligent life elsewhere requires analogous evolutionary transitions. Arriving at the opposite conclusion would require exceptionally conservative priors, evidence for much earlier transitions, multiple instances of transitions, or an alternative

model that can explain why evolutionary transitions took hundreds of millions of years without appealing to rare chance events. Although the model is simple, it provides an initial basis for evaluating how varying biological assumptions and fossil record data impact the probability of evolving intelligent life, and also provides a number of testable predictions, such as that some biological paradoxes will remain unresolved and that planets orbiting M dwarf stars are uninhabitable.

Introduction

Life on Earth has undergone a number of major evolutionary transitions (Szathmáry and Maynard Smith, 1995). These include abiogenesis, as well as the emergence of increasingly complex forms of life such as eukaryotic, multicellular, and intelligent life. Some transitions seem to have occurred only once in Earth's history, suggesting a hypothesis reminiscent of Gould's remark that if the 'tape of life' were to be re-run, 'the chance becomes vanishingly small that anything like human intelligence' would occur (Gould, 1990). Here, we explore this hypothesis.

Given that we cannot re-run the 'tape of life,' it is difficult to derive the probability of these major evolutionary transitions. An alternative would be to examine the timing and frequency of the transitions. The fact that eukaryotic life took over a billion years to emerge from prokaryotic precursors suggests it is a far less probable event than the development of multicellular life, which is thought to have originated independently over 40 times (Grosberg and Strathmann, 2007). The early emergence of abiogenesis is one example that is frequently cited as evidence that simple life must be fairly common throughout the universe (Lineweaver and Davis, 2002). By using the timing of evolutionary transitions to estimate the rates of transition (probability per unit of time), we can derive information about the likelihood of a given transition even if it occurred only once in Earth's history.

However, an additional methodological challenge arises in estimating these rates from their timing in our evolutionary history, given that the timings are subject to a sample bias. In particular, we can only observe evolutionary transitions that occurred rapidly enough to fit within Earth's habitable lifetime. It is estimated that the increased luminosity of the Sun will make complex eukaryotic life impossible on Earth in approximately 0.8 to 1.3 billion years (Ga) (Caldeira and Kasting, 1992; Franck et al., 2006). If a long period of time is required for intelligence to evolve, any intelligent observers on an Earth-like planet may observe early evolutionary transitions occurring regardless of the true transition rate.

A previous analysis that accounted for these sample biases found that early abiogenesis was still consistent with life being rare, as long as one did not presuppose a particular order of magnitude for the transition rate (Spiegel and Turner, 2012). Here, we generalise this Bayesian analysis to a chain of multiple evolutionary transitions, drawing inspiration from the work of Carter (1983) and others (Carter, 2008; Hanson, 1998*b*; McCabe and Lucas, 2010; Watson, 2008) that sought to explain why intelligent life emerged so late in Earth's history. The transitions we consider include abiogenesis, eukaryogenesis, the emergence of sexual reproduction, and the emergence of language and intelligence, although the model can be applied more broadly. These particular transitions are selected because large scientific uncertainty remains around how frequent such transitions are, and because they were prerequisites for the existence of intelligent life on Earth. In the next section we introduce a simple model for evolutionary transitions, and in the subsequent sections determine how evolutionary transition times and varying assumptions can impact our estimates for the probability of these transitions.

A Simplified Model of Evolutionary Transitions

The generalised Carter Model

Let us assume that intelligent life requires a sequence of n evolutionary transitions. We assume that the transitions must occur sequentially, so that the second transition cannot occur until after the first one, the third not until after the second, and so on. Let x_i be the timing of the ith transition, and let t_i be the time it takes between the ith transition and the previous one, so that $t_i = x_i - x_{i-1}$. We set $x_0 = 0$, representing the earliest possible time that the first transition could have occurred. We assume that once an evolutionary transition is possible (i.e. once the previous transition has occurred), it occurs at a constant average rate λ_i, so that each t_i is exponentially distributed with an expected transition time of $\beta_i = 1/\lambda_i$. The joint probability density function for the transition times is the product of exponentials:

$$f_{T_1,\ldots,T_n}(t_1,\ldots,t_n|\beta_1,\ldots,\beta_n) \;=\; \prod_{i=1}^{n} \frac{e^{-t_i/\beta_i}}{\beta_i}. \tag{3.1}$$

The probability that n transitions successfully occur before a certain point in time (for example, the lifetime of Earth) can be found by computing the cumulative distribution function of the final transition time, $F_{X_n}(x_n|\beta_1,\ldots,\beta_n)$. This is found using the hypoexponential distribution:

$$F_{X_n}(x_n|\mathbf{\Theta}) \;=\; 1 - \alpha e^{x_n \mathbf{\Theta}} \mathbf{1} \tag{3.2}$$

where $\alpha = (1, 0, 0, \ldots)$ is a row vector denoting the probability at time zero of being in any given transition state (here, we have probability of 1 that the model starts without having made any evolutionary transitions), $\mathbf{1}$ indicates a column vector of 1's, and $\mathbf{\Theta}$ is a bidiagonal matrix of transition rate parameters:

$$\mathbf{\Theta} \;=\; \begin{bmatrix} -\beta_1^{-1} & \beta_1^{-1} & 0 & \ldots & 0 & 0 \\ 0 & -\beta_2^{-1} & \beta_2^{-1} & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & -\beta_{n-2}^{-1} & \beta_{n-2}^{-1} & 0 \\ 0 & 0 & \ldots & 0 & -\beta_{n-1}^{-1} & \beta_{n-1}^{-1} \\ 0 & 0 & \ldots & 0 & 0 & -\beta_n^{-1} \end{bmatrix}. \tag{3.3}$$

When all transition rates are equal, F_{X_n} follows an Erlang distribution, and closed form solutions also have been derived for hypoexponential distributions with mixes of equal and unequal rates (Amari and Misra, 1997). However, given that the rates of different evolutionary transitions are unlikely to be exactly equal, we assume that $\beta_i \neq \beta_j$ for all i, j in order to simplify our calculations.

A Bayesian analysis of transition times

Our objective is to estimate evolutionary transition rates, given how long it took to complete each transition. This can be found using a Bayesian update:

$$P(\boldsymbol{\Theta}|\boldsymbol{t}) \quad \propto \quad P(\boldsymbol{t}|\boldsymbol{\Theta})P(\boldsymbol{\Theta}) \tag{3.4}$$

where $\boldsymbol{t}$ is the sequence of transition times $t_1, \ldots, t_n$ and $P(\boldsymbol{\Theta})$ is a prior density over the expected transition time parameters. The term $P(\boldsymbol{t}|\boldsymbol{\Theta})$ is equivalent to the likelihood function:

$$\mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{t}) = f_{T_1,\ldots,T_n}(\boldsymbol{t}|\boldsymbol{\Theta}). \tag{3.5}$$

However, this likelihood function needs to be adjusted to account for the fact that we can only observe these data if all evolutionary transitions occurred before the end of Earth's lifetime. Accounting for this sample bias produces the adjusted likelihood function:

$$\mathcal{L}^\star(\boldsymbol{\Theta}|\boldsymbol{t}) = \frac{f_{T_1,\ldots,T_n}(\boldsymbol{t}|\boldsymbol{\Theta})\mathbb{1}_{\{\sum_{i=1}^n t_i < L\}}}{F_{X_n}(L|\boldsymbol{\Theta})} \tag{3.6}$$

where L is the lifespan of Earth and $\mathbb{1}$ is an indicator function that takes the value 1 when the total transition time is less than the lifetime of Earth (i.e. $\sum_{i=1}^n t_i < L$) and 0 otherwise.

Limiting behaviour of the likelihood

We can use limits to evaluate the likelihood of expected transition times that are arbitrarily large. Setting all rate parameters equal, so that all $\beta_i = \beta$, one

can show that for a model with n transitions, taking the limit as $\beta \to \infty$ results in a likelihood of:

$$\lim_{\beta \to \infty} \mathcal{L}^{\star}(\mathbf{\Theta}|t_1, \ldots, t_n) = \frac{n!}{L^n}. \tag{3.7}$$

A more detailed proof of this limit is provided in the appendix. This nonzero constant highlights our first key result, which is that the likelihood does not diminish to zero even as the expected transition times go to infinity. Extremely improbable evolutionary transitions will still be observed by intelligent life if such transitions are crucial to the existence of such intelligence, and this holds true no matter how improbable the transition event.

The Data and Priors

The model requires data for the starting and ending points for Earth's habitability, as well as the timing of the four transitions we evaluate (the origin of life, eukaryotes, sex, and intelligence). The Earth's oceans are thought to have formed between 4.4 and 4.2 billion years ago (Gya) (Cavosie et al., 2005; Peck et al., 2001), which we take to be the earliest point at which life could form on Earth. Our earliest evidence of life dates between 4.1 and 3.5 Gya (Awramik, 1992; Bell et al., 2015), while eukaryotic life has been dated between 1.6 Gya and 2.7 Gya (Betts et al., 2018; Brocks et al., 1999). Sexual reproduction has been hypothesised to have emerged concurrently with eukaryotic life, with firm evidence of its existence at 1.2 Gya (Butterfield, 2000). We date the arrival of human language, symbolic reasoning, and intelligence to be approximately present day, with *Homo sapiens* arrival at 200,000 years ago and the earliest artwork around 80,000 years ago (Pääbo, 2014). We assume the habitability of Earth will end in 0.8-1.3 Gyr due to the increasing luminosity of the Sun. We summarise this data in the Table 1.

Transition	Date (Gya)	Source	Method
Abiogenesis	4.10 ± 0.01	(Bell et al., 2015)	Carbon isotope ratio
Abiogenesis	> 3.9	(Betts et al., 2018)	Molecular clock
Abiogenesis	3.86 ± 0.01	(Mojzsis et al., 1996)	Carbon isotope ratio
Abiogenesis	3.77–4.28	(Dodd et al., 2017)	Microfossils, isotope ratio
Abiogenesis	> 3.5	(Awramik, 1992)	Microfossils, stromatolites
Eukaryotes	2.7	(Brocks et al., 1999)	Sterane presence
Eukaryotes	< 1.84	(Betts et al., 2018)	Molecular clock
Eukaryotes	> 1.87–1.68	(Parfrey et al., 2011)	Molecular clock
Sex	> 1.2	(Butterfield, 2000)	Fossils of red algae
Intelligence	≈ 0	(Pääbo, 2014)	Oldest artwork
Boundary	Timing	Source	Method
Oceans form	4.40 Gya	(Peck et al., 2001)	Oxygen isotope ratio
Oceans form	4.3–4.2 Gya	(Cavosie et al., 2005)	Oxygen isotope ratio
Biosphere ends	0.8–1.2 Ga	(Franck et al., 2006)	Climate model
Biosphere ends	0.9–1.5 Ga	(Caldeira and Kasting, 1992)	Climate model
Biosphere ends	≈ 1.0 Ga	(Kasting, 1988)	Climate model
Biosphere ends	$1.8 - 3.3$ Ga	(Rushby et al., 2013)	Climate model

Table 3.1: Biosphere end dates are given for eukaryotic life. Extremophiles may persist beyond the dates given.

The priors for transition rates

The prior distributions express subjective beliefs and uncertainties about evolutionary transition times, which are then updated to a posterior distribution based on the observed data. Given the large scientific uncertainties surrounding the 'true' rate for each transition (abiogenesis, eukaryogensis, sexual reproduction, evolution of intelligence), we begin by considering an uninformative prior: a log-uniform distribution for each transition rate. Following Spiegel and Turner (2012), this is equivalent to saying we have no prior information that informs us of even an order-of-magnitude estimate of each transition time (β_i). In addition to being uninformative, the log-uniform prior is thought to be most appropriate given that it is also invariant to the choice of parameterisation of event frequencies such as the mean waiting time for an event (β) or the mean number of events per unit time (β^{-1}). In contrast, a uniform prior on mean waiting time per se would, for an interval bounded by 10^{-10} and 10^{10} Ga, imply ≈ 0.9 confidence in values near the upper bound, $> 10^8$ Ga, while a uniform prior over the same interval expressed in frequency terms would imply ≈ 0.9

confidence in values near the lower bound on mean waiting times, $< 10^{-8}$ Ga.

To be well-defined, each log-uniform distribution needs upper and lower bounds. To create an exceptionally conservative lower bound, we assume each expected evolutionary transition time cannot be faster than a rapid bacterial doubling time (roughly 10^{-14} Ga). Appropriately conservative upper bounds on expected transition times are more difficult to produce without controversy. Although some evolutionary transitions could be the result of incremental and deterministic processes, they could also require a precise combination of extremely rare events.

Combinatorial models for upper bounds

In general, if an evolutionary transition requires a specific combination of N binary elements, transition rates to any particular state decline as 2^{-N}. Protein folding is one example that can serve as a more general analogy for why extremely long transition times should be considered. Folding of a 300 residue sequence can be naively modelled as a random search through a space of over 10^{285} conformational states (the bond between a given pair of residue is described by ϕ and ψ torsional angles, each typically regarded as occupying one of 3 low-energy conformations). Given this, it would take $\gg 10^{200}$ times the present age of the universe for a particular folding to occur, even assuming a sampling rate of one trillion conformational states per molecule per second and a volume of concentrated protein solution the size of Earth's oceans. As protein sequences have evolved to fold reliably, most proteins typically fold within seconds, driven by a so-called 'funnel' in the free-energy landscape (Dill and Chan, 1997). But a prebiotic world would have no evolutionary process to shape such a funnel (i.e. variation with selection), and perhaps no reliable mechanism to assemble polymeric components of the requisite kinds. If so, then expected transition times for abiogenesis could be truly immense.

The transition to eukaryotic life also involves similar 'chicken and egg' difficulties, with uncertainty on how an archaeon acquired a proto-mitochondrion,

since endocytosis requires complex machinery only present in eukaryotes (Lane, 2011). A second potential hurdle for eukaryogenesis was the survival of the first prokaryotic host with a bacterial symbiont. Without the protection of spliceosomes and a nucleus, the prokaryotic host would be disrupted by extensive intron transfer from the lysis of its symbionts, resulting in few functional proteins (Koonin, 2011). The chimaera cell would need to evolve these complex defences faster than mutation ratchet effect driving the (already tiny) population to extinction, which could have also required a rare specific outcome among a vast combinatorial space.

Combinatorial models could apply to the evolution of language and intelligence as well. Human language is thought to be fundamentally different than other forms of animal communication, and fundamental to our general intelligence via the human usage of the merge operation (the ability to combine two items into an unordered set) and the resulting ability to construct hierarchal and recursive expressions (Berwick and Chomsky, 2016). If hierarchal and recursive language only results after obtaining a specific combination of neutral alleles, the probability could be extremely low that each allele would spread to fixation and combine to result in intelligence. Some argue that language acquisition was subject to selection pressure, similar to the gradual evolution of the eye (Pinker and Bloom, 1990), but others argue that language arose suddenly (Bolhuis et al., 2014), and that its origin was a consequence of biological spandrels or exaptations (Tattersall, 2008). Even more pessimistic models might incorporate fitness costs associated with large brains, both metabolically and in terms of high levels of parental care (Lineweaver, 2009; Mayr, 1994).

In the presence of this wide uncertainty we proceed by initially setting the upper bound of each prior to 10^{10} Ga and calculating the posterior distribution. We subsequently examine what happens if we change the priors to be even more conservative and discuss whether such conservative priors are plausible.

Results of the Model, Data, and Priors

Our Bayesian calculation is repeated for a variety of combinations of evolutionary transitions (Figure 3.1 and Table 3.2). The posterior probability of expected transition times is maximised around the transition time found in the fossil record (e.g. if abiogenesis took 1 Ga, then posterior probability is maximised along values of $\beta_1 = 1$ Ga). Dramatically fast rates are assigned very low posterior probability (e.g. we can confidently rule out that an evolutionary transition that took 1 Ga does not have an expected transition time of 0.1 Ga). However, the calculation produces an interesting asymmetry, since dramatically slow rates are not assigned low probability in the same way. In fact, the fossil record data are consistent with expected transition times that exceed observed transition times, even by many orders of magnitude (with posterior probability levelling off at a non-zero value when all β_i become sufficiently large). Expected transition times can become arbitrarily large while still being consistent with observed transition times, as suggested by the earlier results of nonzero likelihood in the limit when transition times went to infinity. This phenomenon is caused by observation selection effects, since even astronomically rare transitions will be observed if they were prerequisites for intelligent observers. Failing to account for this selection bias produces results that almost guarantee that all evolutionary transitions will occur within 100 Ga.

`\medmuskip=0mu:`

The importance of this result is that the vast majority of the parameter range consistent with the fossil record is inconsistent with expecting the final transition to occur within the window of Earth's habitability, suggesting that intelligent life is highly improbable. For a model with two transitions, over 90% of the posterior probability is assigned to rate parameters that would result in less than a 1% chance of achieving the final transition within Earth's lifetime (Table 3.2). This proportion of parameter space increases to 99% in a three step model and a four step model (Table 3.2). A substantial amount of posterior

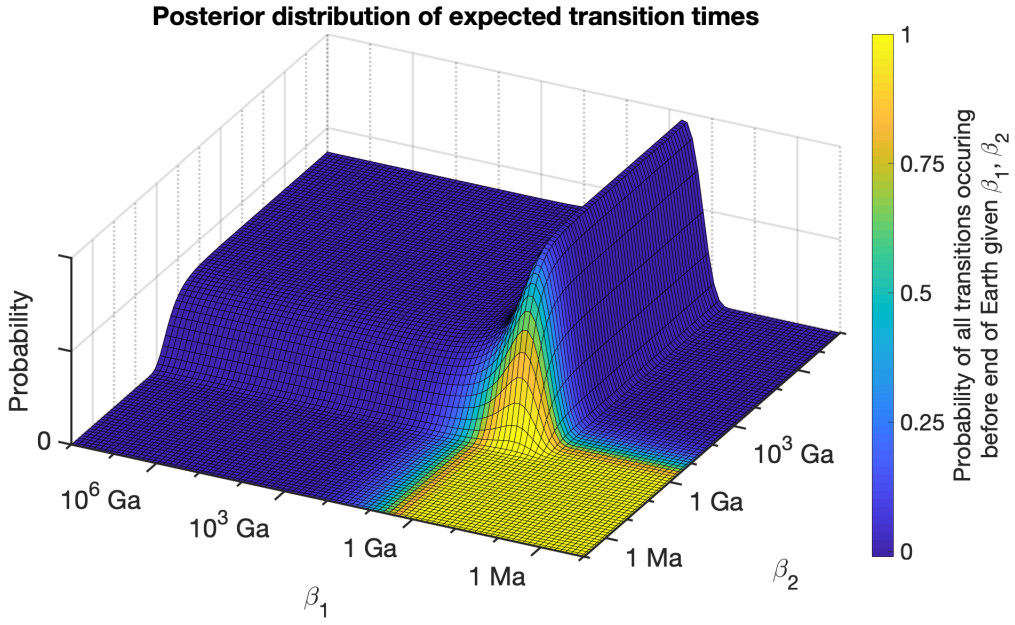


Figure 3.1: Posterior distribution of expected transition times in a two step model (abiogenesis with expected time β_1 and the emergence of human intelligence with expected time β_2). Parameter combinations of rapid rates of transition, such as those resulting in expected transition times of 1 Ma, are inconsistent with the fossil record data and thus have posterior probability close to zero. Conversely, expected transition times exceeding 10^6 Ga are compatible with the data, with posterior probability asymptoting along a nonzero constant as the transition times approach infinity. As a result, parameter combinations resulting in intelligent life before the end of the Earth constitute a very narrow slice of posterior probability (marked in yellow). For example, less than one percent of posterior probability is assigned to parameters that result in the final transition occurring before the end of Earth's lifetime with a greater than one percent chance (first row in Table 3.2).

probability is even assigned to combinations of transition rates that have a less than 10^{-12} or 10^{-24} chance of reaching the final transition within the time that Earth is habitable (corresponding to on the order of 10^{-12} stars in our galaxy or 10^{-24} stars in the observable universe). For example, a three transition model has over 90% of posterior probability assigned to rates that have less than a 10^{-12} chance of reaching the final transition (Table 3.2). If complex or intelligent life beyond Earth requires analogous evolutionary transitions, then the fossil record combined with uninformative priors suggests that such life is exceptionally rare.

Transition times (Gya)							Posterior Mass $F(x_n) > ...$		
0th	1st	2nd	3rd	4th	L	n	10^{-2}	10^{-12}	10^{-24}
4.3	4.2	-	-	0	5.8	2	0.0682	0.3952	0.8330
4.2	3.8	-	-	0	5.2	2	0.0257	0.2490	0.6842
4.2	3.8	1.8	-	-	5.2	2	0.0412	0.2705	0.7005
4.2	-	1.8	-	0	5.2	2	0.0119	0.1835	0.6146
4.3	4.2	1.8	-	0	5.8	3	0.0064	0.1073	0.4293
4.2	3.8	1.8	-	0	5.2	3	0.0022	0.0592	0.3009
4.2	3.8	1.8	1.2	-	5.2	3	0.0056	0.0804	0.3442
4.2	3.8	1.8	1.2	0	5.2	4	0.0003	0.0142	0.1102
4.3	4.2	1.8	1.7	0	5.8	4	0.0027	0.0560	0.2631
						2	0.4174	0.6588	0.9371
						3	0.2686	0.5094	0.8366
						4	0.1678	0.3849	0.7146
							Prior Mass $F(x_n) > ...$		

Table 3.2: A table of evolutionary transition combinations and for each, a proportion of posterior probability assigned to parameters that achieve the final transition with probability greater than 10^{-2}, 10^{-12}, and 10^{-24}. As the number of included transitions increases, the prior and posterior probability of reaching the final transition falls. Grey rows indicate data used elsewhere in the paper.

Priors required to change the result

Our posterior estimates favoured the hypothesis that the selected evolutionary transitions were exceptionally rare. However, the strength of this result will change depending on the Bayesian prior. Increasing (or decreasing) the amount of prior probability assigned to long expected transition times (for example, by increasing or decreasing the upper bounds on the prior) will increase (or decrease, respectively) the amount of posterior probability assigned to long transition times. This is because the fossil record data are equally consistent with long transition times or extremely long transition times, and thus only the prior rather than the data determines to what extent such extremely long transition times should be considered. Bayesian priors are meant to capture wide scientific uncertainty and *a priori* assumptions, so any upper bound, including our arbitrary selection of 10^{10} Ga, will be subject to controversy. It is perhaps more instructive to determine how conservative the priors would need to be

to reach the conclusion that the evolutionary transitions should be expected within Earth's lifetime, and then ask whether limiting the prior in such a way would be reasonable.

We can calculate how low the upper bound of each prior would need to be in order to produce 10% or 50% of posterior mass on parameter values that predict intelligent life within Earth's lifetime with a greater than 1% or 10% chance (Table 2). To get such a result, the bounds need to be unrealistically small (for example, setting a maximum transition time of 40 Ga for a three step model to get 10% mass with a $> 10\%$ chance of life). Given the enormous uncertainties around the processes underpinning these evolutionary transitions, it seems excessively conservative to claim we should limit our priors in such a way.

	10% of Posterior such that		50% of Posterior such that	
n	P(life) $> 1\%$	P(life) $> 10\%$	P(life) $> 1\%$	P(life) $> 10\%$
2	$10^{4.8}$ Ga	$10^{3.1}$ Ga	$10^{1.8}$ Ga	$10^{1.2}$ Ga
3	$10^{2.4}$ Ga	$10^{1.6}$ Ga	$10^{1.2}$ Ga	$10^{0.8}$ Ga
4	$10^{1.8}$ Ga	$10^{1.2}$ Ga	$10^{1.0}$ Ga	$10^{0.6}$ Ga

Table 3.3: Priors needed to get high probability of intelligent life. Data used from grey highlighted rows from table in Figure 3.1.

Data required to change the result

Discovering a second independent instance of an evolutionary transition (e.g. a branch of life unrelated to our universal common ancestor) would be a dramatic development that would change our estimates of transition rates. If we assume the transition rate remains constant over time, we can include such information in a Bayesian update, multiplying the previous posterior distribution by a likelihood function that incorporates the new data. This likelihood function is $\mathcal{L}(\beta_i) = (1 - e^{-1/\beta_i})\tau$, the probability that one or more additional transitions occurred in time τ, where τ is the time window in which the transitions could have occurred (specifically, after the prerequisite transition but before present day).

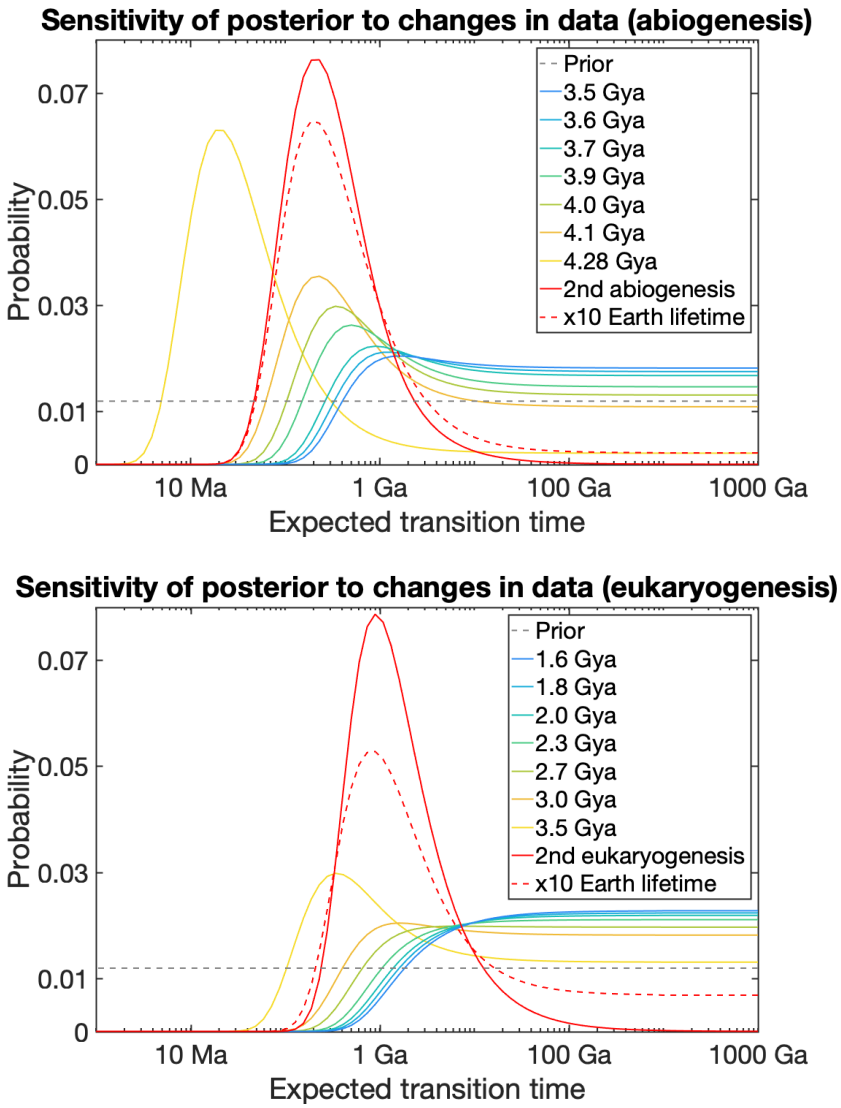


Figure 3.2: Posterior distributions of abiogenesis and eukaryogenesis transition rates for different timings of abiogenesis and eukaryogenesis. Earlier evidence of transitions pushes some posterior mass to faster transition rates, but excluding exceptionally long expected transition times requires finding a second independent instance of the transition or the discovery that Earth will remain habitable for much longer than expected. If we were to find evidence of life occurring within 5 Myr of the start of Earth's habitability, this would also be enough to conclude that abiogenesis is not extremely rare.

Notably, the likelihood of extremely large β_i goes to zero with evidence of additional transitions (Figure 3.2, red solid line). Discovering additional independent instances of a transition therefore rules out the possibility that a particular evolutionary transition is exceptionally unlikely. Indeed, multicellu-

lar life is thought to have emerged over 40 times, demonstrating that it is not an astronomically rare evolutionary transition. Finding even earlier evidence of a successful evolutionary transition could also change our estimates. We conduct a sensitivity analysis of our posterior estimates to changes in the fossil record data for abiogenesis and eukaryogenesis (Figure 3.2). The primary effect of earlier evidence is to adjust the maximum likelihood peak (Figure 3.2, yellow to blue spectrum). However, excluding the possibility of extremely long expected transition times requires an exceptionally rapid transition (on the order of 10s of millions of years). For example, finding evidence of eukaryotic life 3 billion years ago would still be insufficient to rule out expected transition times exceeding 1000 Ga (Figure 3.2). The reason for this is that the conclusion holds so long as the habitable lifetime of Earth is roughly within the same order of magnitude as the evolutionary transition time. If it turned out that Earth will naturally remain habitable far longer than current science predicts, this would also be sufficient to overturn the conclusion that any of the transitions are rare (Figure 3.2, red dotted line).

Given the wide uncertainty in the fossil record for when certain evolutionary transitions occurred (e.g. first eukaryotes between 2.7 and 1.6 Gya), we can also examine what happens if we update our priors based on an interval of possible transition times rather than a specific transition time. Let a_i be the lower end of the range of possible transition times for the ith transition (e.g. the fastest plausible transition time perhaps only tentatively supported by the fossil record), and b_i be the upper end of the range (e.g. a conservative estimate where fossil evidence is clear). To update on these intervals, we use the following likelihood function:

$$\mathcal{L}^{\star}(\mathbf{\Theta}|a_1,\ldots,a_n,b_1,\ldots,b_n) = \int_{a_n}^{b_n}\cdots\int_{a_1}^{b_1}\frac{f_{T_1,\ldots,T_n}(t_1,\ldots,t_n|\beta_1,\ldots,\beta_n)}{F_{X_n}(L|\beta_1,\ldots,\beta_n)}dt_1\ldots dt_n. \tag{3.8}$$

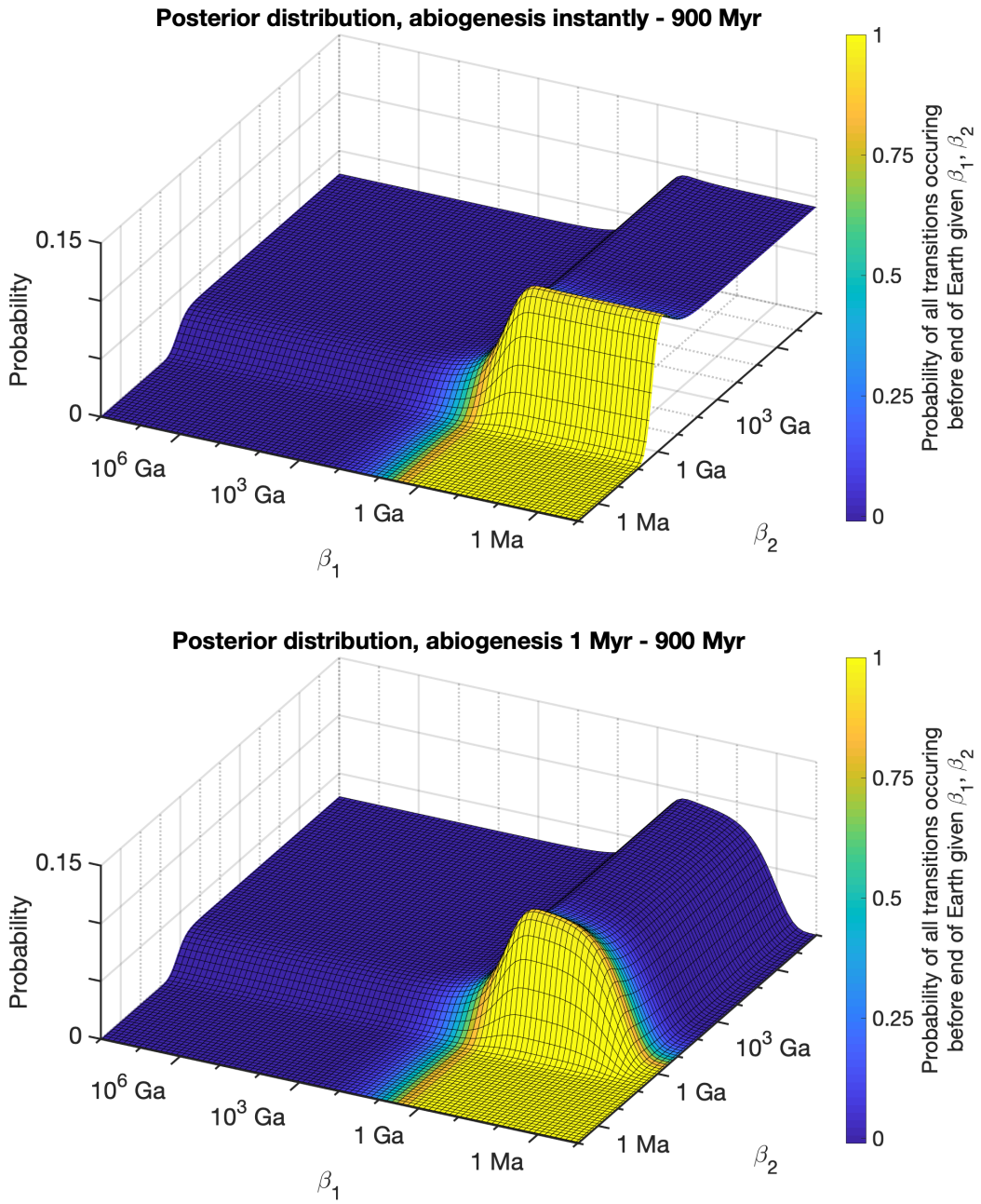


Figure 3.3: Posterior distributions updated on an interval of possible timings for the origin of life and eukaryotic life. On the top, we assume abiogenesis took between 1 Myr and 900 Myr to occur, and on the bottom we assume that abiogenesis may have taken an amount of time ranging from instantaneously to 900 Myr. Although more posterior probability mass is assigned to rates that have a high probability of intelligent life emerging (denoted in yellow), the majority of posterior mass is still on rates that are incompatible with intelligent life occurring within the habitable lifetime of Earth (denoted in blue).

In cases where the fossil record cannot definitively rule out an exceptionally short (or even instantaneous) transition time (e.g. the origin of life or sexual reproduction, see Pearce et al. (2018)), this can cause the posterior estimates to include more mass on rapid transiton rates (Figure 3.3, top). When abiogenesis could take between 0 and 900 Myr, our posterior is consistent with abiogenesis being common or rare, replicating the results of Spiegel and Turner (2012). However, even when abiogenesis is common, our posterior still suggests that intelligent life is rare, supporting the Rare Earth Hypothesis (Ward and Brownlee, 1999). For a two step model with abiogenesis taking between 0 and 900 Myr and eukaryotes taking between 800 and 2800 Myr, only 16% and 54% of posterior probability mass is assigned to parameters that would result in the transitions occurring successfully within the lifetime of Earth with greater than 1% and 10^{-12} chance, respectively.

Assumptions required to change the result

Our model makes a number of simplifying assumptions, primarily that each evolutionary transition has a constant probability of happening per unit time throughout Earth's history. Although this is crude, we believe this is the best model as it requires the fewest biological assumptions. Still, it is worth discussing where the constant probability assumption is likely to fall short and to what extent our results rely on this assumption.

There are a number of reasons why the probability of an evolutionary transition could change over time. Perhaps most importantly, some evolutionary transitions may have required high oxygen concentrations as a source of energy, and oxygen concentrations have changed dramatically over Earth's history (Holland, 2006). The fact that oxygen concentrations have became high enough to support humans only in the past 800 Myr or so has led to some speculation that a planetary oxygenation time is the primary rate-limiting step to intelligent life (Catling et al., 2005). Relatedly, complex life on land requires shielding from ultraviolet radiation, and the emergence of an ozone layer has

also been hypothesised to be a rate limiting step that is correlated with stellar evolution, undermining Carter's original argument (Livio, 1999).

To test this, we adjust our model so that the transition rates change over time. The most dramatic example of this is a model in which the final evolutionary transition to intelligent life has a probability of zero until vertebrates on land emerge (0.34 Gya), and that transition has probability zero until Phanerozoic oxygen concentrations are reached (0.8 Gya). This model essentially tells us that these transitions occurred fairly rapidly once oxygen concentrations were high enough, and the results show a much larger peak around fast rates, suggesting a higher probability of intelligent life emerging in the right conditions (Figure 3.4, Top). However, even these faster transition times are not enough to exclude extremely slow rates. For example, using the previous log-uniform prior with a lower bound of 10^{10} Gyr still results in over 60% of the posterior parameter mass on rates that result in intelligent life with probability less than a 1 in 10^{12}. Overall, accounting for a changing environment in terms of oxygen concentrations does not seem to be sufficient to overturn our key results.

There are numerous other ways in which the evolutionary transition rates could change over time. For example, the total biomass or the number of lineages relevant for a particular evolutionary transition could change over time (e.g. the concentration of potential symbionts for eukaryotic life, or the number of candidate animal lineages that could evolve intelligence) and the rate of catastrophes that could act as evolutionary setbacks could also change over time (e.g. asteroid impacts during the Late Heavy Bombardment or the rate of gamma ray bursts) (Ćirković et al., 2009). Although these examples make it clear that the assumption of a constant transition rate over time is oversimplified, creating a comprehensive model that incorporates all factors into the evolutionary transition rates would require many more assumptions. We can say in general that any an alternative model still needs to explain why certain evolutionary transitions took such a long period of time.

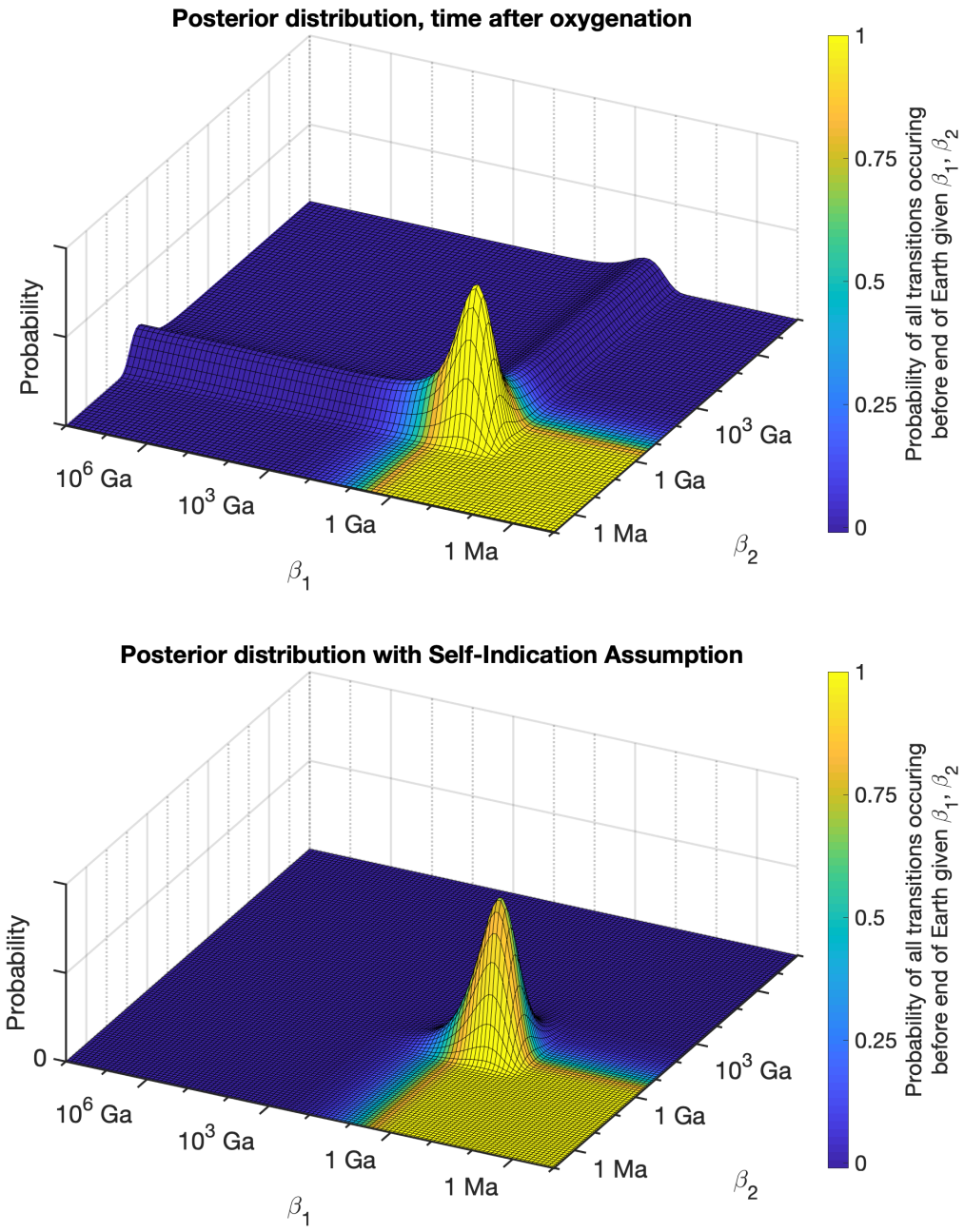


Figure 3.4: (Top) Posterior distribution if we assume two transitions that were made possible only after high oxygenation levels. Given the late oxygenation of Earth's atmosphere, these transition times are short, resulting in higher posterior probability on faster rates. However, arbitrarily slow rates are still not excluded. (Bottom) Posterior distribution if we adopt the Self-Indication Assumption, and weight all parameter combinations by their probability of obtaining intelligent life. Only parameters that are consistent with intelligent life are assigned high probability, and extremely slow rates are ruled out entirely.

In order to conclude that intelligent life is common, an alternative model would not only need to explain why intelligence emerged on roughly the same timescale of Earth's habitable lifetime, but also why eukaryotic life and other evolutionary transitions did so as well. Certainly with enough assumptions one could create a model that guarantees that each transition will occur at roughly the same time that it did on Earth, and then conclude that any Earth-like habitat lasting 5 Gyr will have a high probability of hosting intelligent life. But we ultimately think that the most parsimonious model is one in which the long transition times are a byproduct of contingency.

In addition to assumptions about how evolutionary transitions occur, we also consider assumptions around how to use the information that we exist as observers. So far, we have assumed that we can derive no information on the probability of intelligent life from our own existence, since any intelligent observer will inevitably find themselves in a location where intelligent life successfully emerged regardless of the probability. But another line of reasoning, known as the 'Self-Indication Assumption' (Bostrom, 2013*a*), essentially suggests that if there are different possible worlds with differing numbers of observers, we should weigh those possibilities in proportion to the number of observers. For example, if we posit only two possible universes, one with 10 human-like civilisations and one with 10 billion, SIA implies that all else being equal we should be 1 billion times more likely to live in the universe with 10 billion civilisations. If SIA is correct, this could greatly undermine the premises argued here, and under our simple model it would produce high probability of fast rates that reliably lead to intelligent life (Figure 3.4, bottom). However, embracing SIA leads to a number of other very counterintuitive results, for example essentially guaranteeing that the universe is exceptionally large or infinite even without accounting for cosmological evidence (Bostrom and Ćirković, 2003), or giving substantial probability to any bizarre theory that proposes a large enough population of observers to overwhelm the *a priori* implausibility of the theory (for example a theory that each planet has $10^{10^{100}}$ copies of itself

on 'other planes' would seem hard to refute if one adopted SIA (Olum, 2002)). Adopting SIA thus will undermine our results, but also undermine any other scientific result that would suggest a lower number of observers in the universe. The plausibility and implications of SIA remain poorly understood and outside the scope of our present work, but we proceed by proposing a set of testable predictions.

Testable Predictions

The model offers a number of testable predictions. First, we conclude that intelligent life is exceptionally rare and that we may possibly be the only intelligent civilisation within the observable universe, so long as we assume that intelligent life elsewhere requires similar evolutionary transitions. Although this may seem like a large assumption, there are good reasons to believe many evolutionary transitions have universal properties (Levin et al., 2017), and it also follows if we reason that our civilisation is typical. If there were substantially easier evolutionary pathways to intelligent life that did not require such evolutionary transitions, we should expect to observe this easier evolutionary history instead. Although it is hard to show beyond doubt the absence of extraterrestrial intelligence, so far all of our astronomical data are consistent with being alone (Tipler, 1980). A handful of other testable predictions follow from the model as well.

Exceptionally rare transitions

The unlikeliness of different evolutionary transitions can also be tested more directly as we learn more about the underlying physical and biological processes of different evolutionary transitions, or find evidence that a transition occurred more than once. Abiogenesis and eukaryogenesis both involve unexplained or partially unexplained paradoxes, which could be resolved through further research. Examples include Eigen's paradox, describing the mystery of how

systems for error correction evolved in the absence of error correction, with similar 'chicken and egg' problems in the case of eukaryogenesis. Our model predicts that these paradoxes and similar ones may only be resolved by allowing for exceptionally rare chance events. If subsequent research discovers relatively easy pathways for each of these transitions (as it has done for multicellular life, for example), this would falsify our prediction (Chen and Kipping, 2018).

Habitability of red dwarf stars

There is an interesting connection between the evolutionary transitions model and arguments for the habitability of planets orbiting red dwarf stars (Rushby et al., 2013; Waltham, 2017). A typical M dwarf star will last one trillion years, roughly two orders of magnitude longer than our Sun, and could also be host to habitable exoplanets (Haqq-Misra et al., 2017). However, the habitability of planets orbiting M dwarf stars remains an open question, as tidal locking, solar flares, or other harsh conditions could make these planets inimical to complex life (Zendejas et al., 2010). If we start with the assumption that environments around dwarf stars are just as habitable as our solar system, such that the transition parameters are the same, we can calculate the probability ratio of final evolutionary transitions occurring around such dwarf stars as opposed to a habitat lasting as long as Earth. These probability ratios are heavily skewed in favour of longer lasting environments, by between four and ten orders of magnitude depending on the number of transitions (Figure 3.5, top). Assuming we are typical for intelligent life and finding ourselves not orbiting a red dwarf star, our model brings us to a testable prediction that the transition parameters are not similar, and some other factor is reducing the habitability of dwarf star environments by a factor of over ten thousand when compared to Earth. The strength of this predicted factor is great enough that the prediction ought to be testable using climate models or other tools currently accessible, and coincide with eventual scientific consensus that dwarf stars are inimical for complex life.

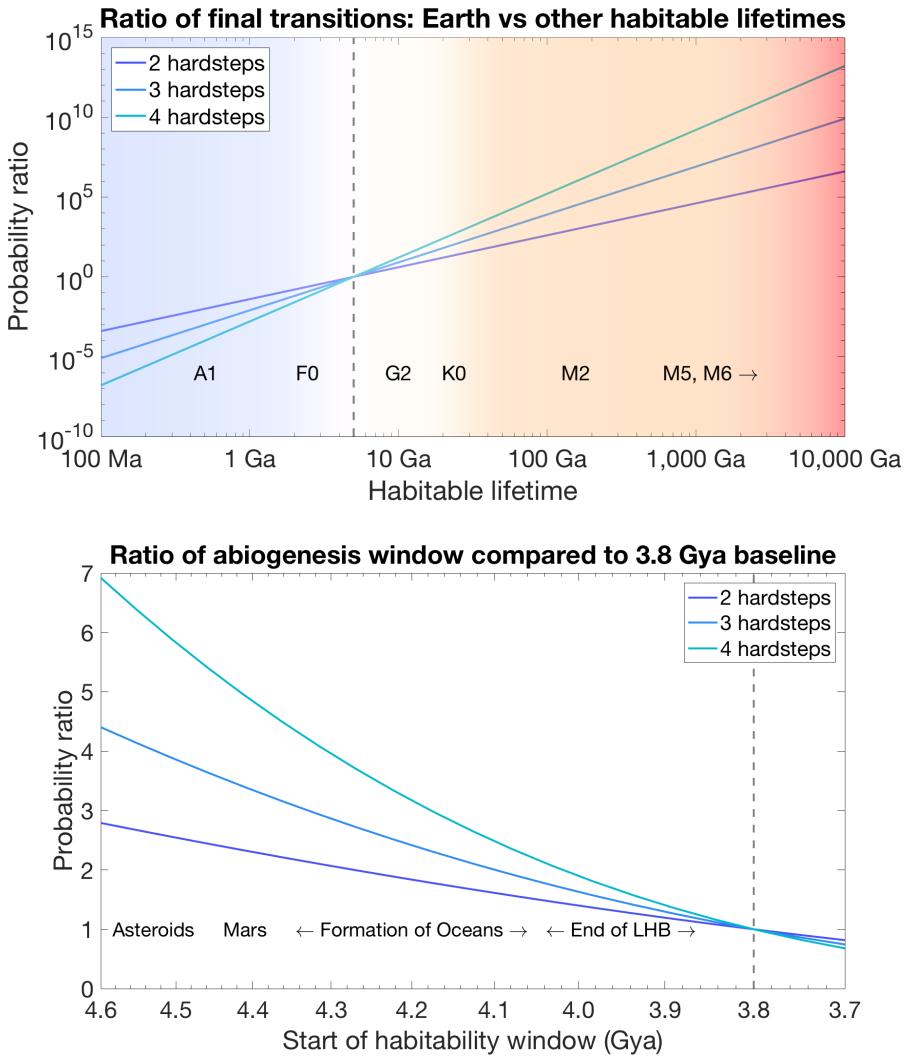


Figure 3.5: (Top) The ratio of final transitions that occur within a 5 Ga habitability window as opposed to a habitable window of another length of time, all else being equal. Habitable environments that last longer are more likely to support the final transition by many orders of magnitude, suggesting that other factors must reduce the habitability of red dwarf stars by many orders of magnitude. (Bottom) Probability ratios of earliest time abiogenesis could have occurred, as compared to a 3.8 Gya baseline. All else being equal, longer habitability windows have slightly higher probabilities, which increases the credence that the solar system was habitable from an early date (including an extraterrestrial origin of life).

Extraterrestrial Origin of Life?

When was the earliest that life could have emerged? Given our model, we can evaluate this by taking the likelihood ratio between the hypothesis of an early

conducive environment (say, with the formation of the oceans at 4.3 Gya) and the hypothesis that life was only possible relatively late in Earth's beginning (say, with the end of the Late Heavy Bombardment at 3.9 Gya). It has also been suggested that Mars was habitable 100 million years before Earth's oceans formed (Sleep and Zahnle, 1998), and that asteroids could have been habitable yet another 100 million years earlier (Abramov and Mojzsis, 2011). If the rate of material transfer between Mars and Earth is high enough to consider the two planets a single environment, we can compare the likelihood ratio between an extraterrestrial origin of life and origin of life on Earth as well. In general, the model favours hypotheses with earlier possible starting dates for the first transition (Figure 3.5, bottom), with the strength increasing with the number of transitions included (Davies, 2003; McCabe and Lucas, 2010). The effect is very modest though. For example, the likelihood ratio for a 4.4 Gya starting point versus a 3.8 Gya starting point differs only a factor of about two to seven. Perhaps more relevant is a prediction that if we were to find life on Mars, it would have emerged extremely early and have a common ancestor with life on Earth.

Conclusions

It took approximately 4.5 billion years for a series of evolutionary transitions resulting in intelligent life to unfold on Earth. In another billion years, the increasing luminosity of the Sun will make Earth uninhabitable for complex life. Intelligence therefore emerged late in Earth's lifetime. Together with the dispersed timing of key evolutionary transitions and plausible priors, one can conclude that the expected transition times likely exceed the lifetime of Earth, perhaps by many orders of magnitude. In turn, this suggests that intelligent life is likely to be exceptionally rare. Arriving at an alternative conclusion would require either exceptionally conservative priors, finding additional instances of evolutionary transitions, or adopting an alternative model that can explain

why evolutionary transitions took so long on Earth without appealing to rare stochastic occurrences. The model provides a number of other testable predictions, including that M dwarf stars are uninhabitable, that many biological paradoxes will remain unsolved without allowing for extremely unlikely events, and that, counterintuitively, we might be slightly more likely to find simple life on Mars.

Appendix

We use limits to analyse the likelihood of the evolutionary transition parameters as the transitions become arbitrarily unlikely. Here, we use the parameters $\lambda_i = 1/\beta_i$, and take the limit of the likelihood function assuming all $\lambda_i = \lambda$, where the likelihood function is $\mathcal{L}^\star(\lambda|\boldsymbol{t}) = g(\boldsymbol{t}|\lambda)/F(L|\lambda)$ where

$$g(\boldsymbol{t}|\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda t_i} \tag{3.9}$$

$$F(L|\lambda) = 1 - \sum_{\ell=0}^{n-1} \frac{1}{\ell!}(\lambda L)^\ell e^{-\lambda L}. \tag{3.10}$$

The limit is therefore

$$\lim_{\lambda\to 0} \frac{g(\boldsymbol{t}|\lambda)}{F(L|\lambda)} = \lim_{\lambda\to 0} \frac{\lambda^n \prod_{i=1}^{n} e^{-\lambda t_i}}{1 - e^{-\lambda L} - \sum_{\ell=1}^{n-1} \frac{1}{\ell!}(\lambda L)^\ell e^{-\lambda L}}. \tag{3.11}$$

$$\tag{3.12}$$

Applying L'Hospital's rule, we can find the kth derivative (for $k \leq n$) is:

$$\frac{d^k}{d\lambda^k} g(\boldsymbol{t}|\lambda) = \frac{n!\lambda^{n-k}}{(n-k)!}\left(\prod_{i=1}^{n} e^{-\lambda t_i}\right) + O(\lambda^{n-k+1}) \tag{3.13}$$

$$\frac{d^k}{d\lambda^k} F(L|\lambda) = \frac{L^n}{(n-k)!}\lambda^{(n-k)} e^{-\lambda L} + O(\lambda^{n-k+1}) \tag{3.14}$$

and the nth derivative is

$$\frac{d^n}{d\lambda^n} g(x) = n!\left(\prod_{i=1}^{n} e^{-\lambda t_i}\right) + O(\lambda) \tag{3.15}$$

$$\frac{d^n}{d\lambda^n} F(L) = L^n e^{-\lambda L} + O(\lambda) \tag{3.16}$$

such that

$$\lim_{\lambda \to 0} \frac{g(\boldsymbol{t}|\lambda)}{F(L|\lambda)} = \frac{n!}{L^n}. \tag{3.17}$$

Since $n! > 0$, this demonstrates that the likelihood of arbitrarily difficult steps is positive.

Chapter 4

Catastrophe risk can accelerate unlikely evolutionary transitions

Andrew E. Snyder-Beattie and Michael B. Bonsall

Abstract

Intelligent life has emerged late in Earth's habitable lifetime, and required a preceding series of key evolutionary transitions. A simple model developed by Carter (1983) explains the late arrival of intelligent life by positing that a handful of these evolutionary transitions were exceptionally unlikely 'critical steps.' This model has generated a number of additional predictions, such that 1) intelligent life beyond Earth is rare, 2) the transitions that constitute critical steps will be evenly spaced throughout the Earth's habitable lifetime and, using this evenly spaced property, that 3) there were about four critical steps for the emergence of intelligent life. An alternative model, known as the Neocatastrophism Hypothesis, has been proposed and argues that intelligent life in our galaxy was delayed by frequent catastrophes (particularly gamma ray bursts) that served to set back evolutionary innovation. Here, we generalise the Carter Model and explore this hypothesis by including catastrophes that can 'undo' an evolutionary transition. We find that although a model of the Neocatastrophsim Hypothesis can explain the late emergence of intelligent life on Earth, it cannot explain the long timescales associated with the emergence

of eukaryotic life or other evolutionary transitions that could occur in environments relatively unaffected by gamma ray bursts. However, using catastrophes to supplement rather than replace the Carter Model does result in a set of new predictions that open up additional hypotheses. In general, introducing catastrophes or evolutionary dead ends to the Carter Model can create situations in which critical steps occur rapidly or in clusters, suggesting that past estimates of the number of critical steps could be underestimating the true number. If catastrophes affect complex life more than simple life, the critical steps will also exhibit a pattern of acceleration towards the end of Earth's habitable lifetime, suggesting that the recent increase in biological complexity over the past 500 Myr could reflect previously overlooked critical steps.

Introduction

Intelligent life has emerged late in Earth's lifetime. In about 1 billion years (Gyr), the increasing luminosity of the Sun in its later stages of life will destroy the Earth's ability to support complex life (Franck et al., 2006; Lenton and von Bloh, 2001), a comparatively short amount of time compared to the 4 Gyr that it has taken for intelligence to emerge. The fact that intelligent life emerged on a timescale within an order of magnitude of our star's lifetime is puzzling, as the timescales associated with biological and stellar evolution are driven by fundamentally different processes and thus ought to be uncorrelated.

Our work builds on literature inspired by Carter (1983), who was the first to notice this coincidence and proposed the following resolution to the puzzle based on observation selection effects. Let L denote the habitable lifetime of Earth, and τ be the timescale it takes for evolution to produce intelligent life. Given that *a priori* these timescales may differ by many orders of magnitude, the possibility that $\tau \ll L$ or $\tau \gg L$ should be heavily favoured over the possibility that $L \approx \tau$. We can also rule out the possibility that $\tau \ll L$, since intelligent life did not emerge in the early stages of Earth's lifetime. However,

due to observation selection effects, we can not rule out that $\tau \gg L$, the possibility that intelligent life typically takes much longer than Earth's lifetime. Although most biospheres will expire long before intelligent life emerges, for the rare instances in which it does emerge, such life will inevitably still find itself emerging within the habitable lifetime of its environment. Moreover, this emergence will occur on roughly the same timescale as the environment's lifetime, consistent with our observation of $\tau \approx L$. This line of reasoning has become known as the Carter Argument.

There could be a number of reasons for why intelligence in general takes a long time to emerge or is a rare event. On Earth, intelligent life required a preceding series of evolutionary transitions (Smith and Szathmary, 1997), such as abiogenesis, eukaryogenesis, and the emergence of sexual reproduction, multicellular life, and intelligence itself. Although some of these transitions such as multicellularity have occurred multiple times throughout Earth's history (Grosberg and Strathmann, 2007), other transitions have only occurred once and may have been exceptionally unlikely, even in conducive environments. Built around Carter's original argument, a small body of research has examined a model where a small number of rare 'critical steps' are required for intelligent life (Carter, 2008; Flambaum, 2003; Hanson, 1998*b*; McCabe and Lucas, 2010; Watson, 2008). Here, a *critical step* is defined as one that has an exceptionally low probability per unit time, so that the expected time for the critical step greatly exceeds the lifetime of Earth.

The work inspired by Carter's critical step model has produced a number of interesting predictions, many of which seem broadly consistent with fossil record data. If one assumes that all critical steps must occur within the habitable lifetime of Earth, the timing of the critical steps will become evenly spaced out across Earth's history (Carter, 2008; Hanson, 1998*b*), a pattern that matches the timing of some key evolutionary transitions (Watson, 2008). Interestingly, this holds even if the critical steps have vastly different probabilities per unit time, so long as the probabilities for all critical steps are low enough

for each event to be unlikely to occur within Earth's lifetime (Hanson, 1998*b*). Another key prediction is that the fraction of time remaining between the final critical step and the end of the biosphere will be inversely proportional to the number of critical steps, so that intelligent life is likely to find itself towards the end of Earth's biosphere. The fact that we have emerged in the final fifth of Earth's lifetime can then be used to predict that the number of critical steps in our evolutionary history was likely somewhere between three and eight (Carter, 2008; Watson, 2008), although models with up to 12 critical steps are still broadly consistent (i.e. within a 95% confidence interval (Waltham, 2017)).

Given that the model predicts evenly spaced critical steps, one can also use this fact to evaluate whether a particular evolutionary transition is likely to be a critical step (in the sense of having an exceptionally low probability per unit time). If a transition happens quickly or a cluster of transitions occurs in a short period of time, this model would predict that the rapid transitions would not be a critical step. The rapid origin of life has been cited as possible evidence that it is not a critical step (Lineweaver and Davis, 2002), and subsequent analysis using the critical step model predicted that some of the key evolutionary transitions related to the origin of life as defined by Smith and Szathmary (1997) (emergence of replicating molecules, chromosomes, and the central dogma of molecular biology) could not have been critical steps as all three of them occurred within the first few hundred million years of Earth's history (Watson, 2008). Similarly, despite the stunning increase in biological complexity over the past 500 Myr, the even spacing property of the Carter Model has led some to conclude that at most one or two critical steps could have occurred in this time period. This is perhaps best encapsulated in the words of Hanson (1998*b*):

> For example, textbooks tend to favor a sequence of major transitions which become more and more closely spaced in time as one approaches the present day. This is natural since the fossil record improves with time, but it also strongly suggests that the three

> most recent transitions in a typical textbook series are unlikely to all be critical transitions. This is because any two particular critical steps are unlikely to be very closely spaced in time, and three critical steps are especially unlikely to be very close.

Here, we generalise the critical steps model by allowing for catastrophes that 'undo' evolutionary transitions, as well as allowing for evolutionary 'dead-ends' that could preclude a critical step. We show that these conditions can result in unevenly spaced critical steps, including critical steps that occur rapidly, in tight clusters, or that accelerate over time. In addition to opening up new hypotheses about whether the previous literature has underestimated the number of critical steps, this work also enables us to test an alternative hypothesis to the Carter Argument, known as the Neocatastrophism Hypothesis (Ćirković et al., 2009).

The Neocatastrophism Hypothesis

The Neocatastrophism Hypothesis seeks to explain why we do not see any signs of other intelligent civilisations in our galaxy, despite the hundreds of billions of stars that could potentially be hosts to habitable planets (Vukotic and Cirkovic, 2008). Many of the conditions for habitable planets in our galaxy seem to have been in place many billions of years ago (Lineweaver et al., 2004), suggesting that if intelligent life is common, we should expect to find evidence of many civilizations in our galaxy far older than our own. Conversely, the Neocatastrophism Hypothesis argues that our failure to observe ancient intelligent civilisations could still be reconciled with intelligent life being common, if intelligent life were just now emerging simultaneously across our galaxy. This correlation in emergence times for intelligent life could arise if intelligent life was suppressed beforehand by galaxy-wide catastrophes such as gamma ray bursts (GRBs) (Annis, 1999; Ćirković et al., 2009). If intelligent life requires a period of catastrophe-free time to emerge successfully, then we may expect intelligent life to emerge in a correlated fashion across the galaxy in the first

sufficiently long window of catastrophe-free time. Given that the frequency of such catastrophes is declines over time (some models suggest the exponential function $\mu(t) = \mu(0)e^{-t/t_\gamma}$, e.g. Ćirković et al. (2009)), the windows of catastrophe-free time will increase over time until their present day rate of approximately 1 every 220 million years (Myr) (Annis, 1999). It is striking that this rate roughly corresponds to the amount of time that vertebrate life has successfully survived on land (about 340 Myr) (Long and Gordon, 2004). The Neocatastrophism Hypothesis thus predicts that intelligent life will take a long period of time to emerge successfully on Earth, in a similar way as the Carter Argument. However, the hypotheses differ greatly in their proposed resolutions to the apparently lifeless galaxy. The Neocatastrophism Hypothesis resolves the question by suggesting that intelligent life is just getting started across the galaxy and is not yet detectable, while the Carter Argument predicts that intelligent life required a small set of exceptionally rare critical steps, and thus is itself exceptionally rare.

Methods

Let us assume that intelligent life requires n evolutionary transitions, each of which is exponentially distributed with a constant rate λ_i. We assume that the transitions must occur in sequence, so that the second transition can only occur after the first, the third after the second, and so on. This can be modelled as a continuous time Markov Chain, for which a substantial amount of theory already exists (Anderson, 2012). Let the Markov Chain be denoted $X(t)$ with $n + 1$ states $i = 0, 1, \ldots, n$, where the state $i = 0$ is the state before the first evolutionary transition and the ith state represents the time after the ith transition but before the next one. We assume $X(0) = 0$, so that Earth starts with no evolutionary transitions having occurred. Let q_{ij} indicate the rate at which the chain moves from state i to state j, producing a generator matrix $\mathbf{Q}$:

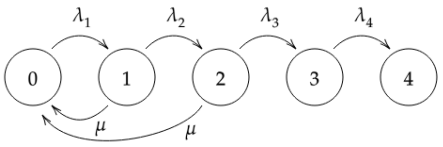
$$\mathbf{Q} = \begin{bmatrix} -q_0 & q_{01} & \cdots & q_{0n} \\ q_{10} & -q_1 & \cdots & q_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ q_{n0} & q_{n1} & \cdots & -q_n \end{bmatrix} \tag{4.1}$$

where $q_i = \sum_{i \neq j} q_{ij}$, so that all rows sum to 0. In the case of n evolutionary transitions occurring at rate λ_i, our generator matrix is:

$$\mathbf{Q} = \begin{bmatrix} -\lambda_1 & \lambda_1 & 0 & \dots & 0 \\ 0 & -\lambda_2 & \lambda_2 & & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -\lambda_n & \lambda_n \\ 0 & \dots & 0 & 0 & 0 \end{bmatrix} \tag{4.2}$$

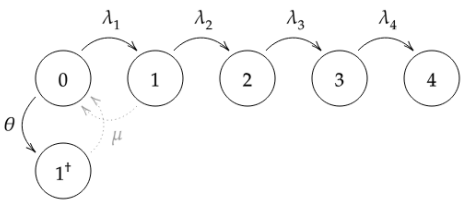
where the final row of zeros indicate that the nth state is an absorbing state, representing the emergence of intelligent life. This reflects our interest in studying the transient behaviour of $X(t)$ before intelligence emerges, rather than an assumption that no further evolutionary transitions are possible once intelligent life emerges.

We provide two modifications to this basic model. The first variation is to allow some states on the chain to transition to earlier states, representing catastrophes or other setbacks that could 'undo' evolutionary transitions. The second variation is to create additional absorbing states that represent evolutionary 'dead-ends.' We define μ to be the catastrophe rate, and θ to be the evolutionary transition rate to an evolutionary dead-end. Visual representations of Markov chains with these characteristics are given in Table 4.1, along with their generator matrices.



$$\mathbf{Q} = \begin{bmatrix} -\lambda_1 & \lambda_1 & 0 & 0 & 0 \\ \mu & -(\lambda_2+\mu) & \lambda_2 & 0 & 0 \\ \mu & 0 & -(\lambda_3+\mu) & \lambda_3 & 0 \\ 0 & 0 & 0 & -\lambda_4 & \lambda_4 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Example of evolutionary transitions model with catastrophes that 'undo' the first two transitions with rate μ.



$$\mathbf{Q} = \begin{bmatrix} -(\lambda_1+\theta) & \theta & \lambda_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\lambda_2 & \lambda_2 & 0 & 0 \\ 0 & 0 & 0 & -\lambda_3 & \lambda_3 & 0 \\ 0 & 0 & 0 & 0 & -\lambda_4 & \lambda_4 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Example of evolutionary transitions model with an evolutionary 'dead-end' that can preclude the first evolutionary transition.

Table 4.1: A diagram of the Markov Chain models for a catastrophe model and an evolutionary dead end model, along with their transition matricies.

By varying the structure and rates of $X(t)$, we can investigate different hypotheses about evolutionary transitions on Earth. To incorporate observation selection effects, we condition all probabilities on the eventual emergence of intelligent life before the end of Earth's lifetime. Let L denote the habitable lifetime of Earth. Our goal is to calculate $P(X(t) = i|X(L) = n)$ for all i, the state probabilities throughout Earth's history conditional on intelligence (the final transition) emerging before the end of Earth's lifetime. Using Bayes Rule, we obtain:

$$P(X(t) = i|X(L) = n) = \frac{P(X(L) = n|X(t) = i)\ P(X(t) = i)}{P(X(L) = n)}. \tag{4.3}$$

Let $\pi_i(t)$ be defined as $P(X(t) = i)$, the probability of being in the ith state at time t and $\boldsymbol{\pi}(t)$ be the row vector of these state probabilities $\boldsymbol{\pi}(t) = [\pi_0(t)\ \pi_1(t)\ \ldots\ \pi_n(t)]$. The state probabilities can be calculated using the equation:

$$\boldsymbol{\pi}(t) = \boldsymbol{\pi}(0)\mathbf{P}(t) \tag{4.4}$$

where the matrix $\mathbf{P}(t)$ is found from the Kolmogorov backward equations:

$$\mathbf{P}'(t) = \mathbf{Q}\mathbf{P}(t) \tag{4.5}$$

where the entries of $\mathbf{P}(t)$ are $p_{ij}(t)$, denoting the probability of moving from state i to state j within time t, and the entries of $\mathbf{P}'(t)$ are the derivative of each $p_{ij}(t)$ with respect to time. Let $\pi_i^\star(t)$ denote $P(X(t) = i|X(L) = n)$, the probability of being in state i at time t conditional on the final transition occurring before the end of Earth's lifetime. The calculation for each $\pi_i^\star(t)$ can be written using Bayes Rule:

$$\pi_i^\star(t) = p_{in}(L-t)p_{0i}(t)p_{0n}^{-1}(L) \tag{4.6}$$

where the first term is the probability that the process moves from state i to state n in the remaining time left on Earth, $L-t$, the second term denotes the probability of having moved from the initial condition of state 0 to state i within time t, and the final term is a normalizing constant denoting the probability that the final transition is reached by the end of Earth's lifetime.

When the generator matrix $\mathbf{Q}$ is constant, the solution of the Kolmogorov backward equations takes the form of a matrix exponential:

$$\mathbf{P}(t) = e^{\mathbf{Q}t} = \sum_{k=0}^{\infty} \frac{t^n}{n!}\mathbf{Q}^n. \tag{4.7}$$

However, in order to test the Neocatastrophism Hypothesis, we also want to investigate what happens if the catastrophe rate declines over time, producing a system of differential equations of the form $\mathbf{P}'(t) = \mathbf{Q}(t)\mathbf{P}(t)$ that lacks a closed form solution. We instead adopt a numerical integration scheme from Shampine and Reichelt (1997) based on an explicit Runge-Kutta method with adaptive step size control.

Results

We begin by evaluating the simplest model with n evolutionary transitions, no catastrophes, and no evolutionary dead ends. We normalise the lifetime of Earth to $L = 1$, and set all the transition rates to $\lambda_i = 10^{-3}$, so that the expected time for each transition greatly exceed the habitable lifetime of Earth

by three orders of magnitude. In doing so, this model replicates the results of Watson (2008) and Carter (2008), in producing probability distribution functions for each critical step that are evenly spaced throughout the habitable lifetime of the planet (Figure 4.1). For a model with one critical step, the probability of the transition occurring at a particular time becomes uniform over the habitable interval. As the number of critical steps increases, the probability distributions for each transition become compressed and evenly spaced, so that the kth critical step should occur roughly at the point in the planet's lifetime $k/(n+1)$.

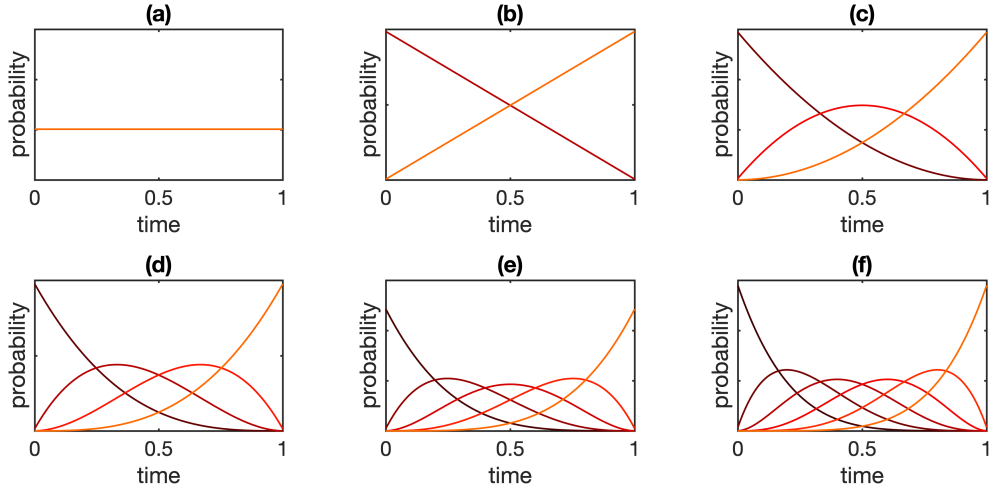


Figure 4.1: Probability distributions for the timing of each critical step, in a model with between (a) one and (f) six critical steps with each expected transition time greatly exceeding the lifetime of Earth. The steps are evenly spaced out across Earth's lifetime, identical to the results by Watson (2008) and Carter (2008).

We can also examine what happens when a transition rate becomes more rapid, so that the expected transition time occurs within Earth's lifetime (first row of Figure 4.2). Unsurprisingly, faster transitions produce probability density functions that are skewed towards earlier transitions. When the rate is fast enough, a transition will occur almost immediately after the preceding transition, and the remaining difficult transitions will become spaced evenly over

the habitable interval. This can be seen in Figure 4.2 in a model with four transitions, where the first transition occurs more rapidly than the other three.

If all of the transition rates become more rapid, so that the expected transition times are approximately equal to $1/n$th the lifetime of Earth ($\lambda_i = n/L$), the distributions will all skew more towards the start of Earth's lifetime. However, there will still remain a substantial probability that the final few steps will occur in the latter half of Earth's lifetime (Figure 4.2.e). When all rates are rapid enough to be expected well within Earth's lifetime ($\lambda_i = 10n/L$), the transition distributions converge to a sum of exponential distributions (Figure 4.2.f). Since the probability of successfully achieving intelligence within the Earth's lifetime is so high with such parameters, the need to condition on the final transition occurring before the end of Earth becomes redundant and the distributions converge to the unconditional distributions without observer selection effects.

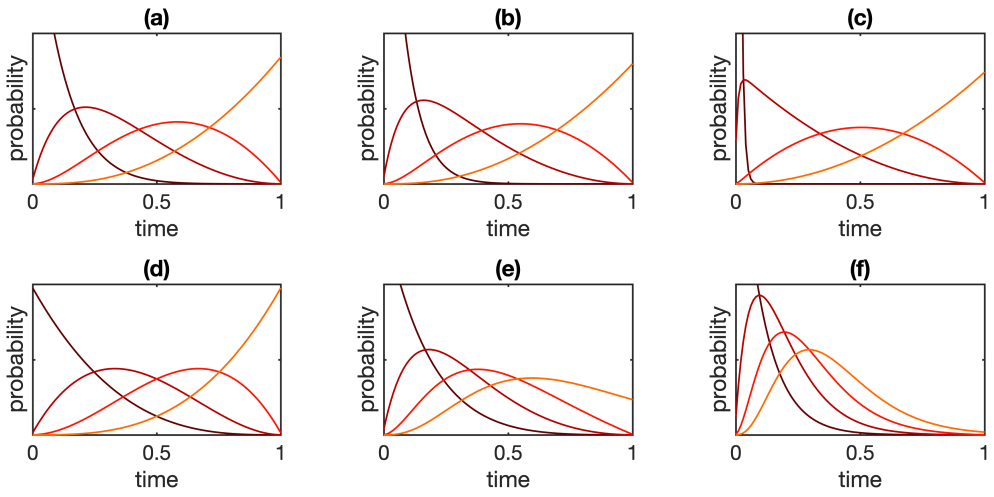


Figure 4.2: Probability density functions for transition times with faster rates in a four step model. In the first row, only the first transition becomes faster, starting with $\lambda_1 = 5$ (a), then moving to $\lambda_1 = 10$ (b) and $\lambda_1 = 100$ (c). In the second row, all transitions become faster, starting with $\lambda = 1$ (d), then moving to $\lambda = 5$ (e) and $\lambda = 10$ (f).

With an understanding of how the model behaves without any catastrophes

or evolutionary dead ends, we can now investigate these possibilities.

Modelling catastrophes

We now include catastrophes in the model. Each catastrophe is modelled as reverting the Markov chain to an earlier state, effectively undoing one or more of the preceding evolutionary transitions. We can represent different scenarios or hypotheses depending on which transitions are vulnerable to catastrophes, and how many transitions are undone with each catastrophe.

Broadly speaking, there are two types of model combinations that we focus on. The first type of model describes a situation in which more complex forms of life are more vulnerable to catastrophes, whereas simple life forms are resistant to catastrophes. Many mass extinctions have disproportionately affected terrestrial animals with large body size (Raup, 1986), and catastrophes stemming from cosmic radiation will also disproportionately affect complex life on land (Lsor and Shaviv, 1997). Runaway climate feedback, such as in a snowball Earth episode, could destroy many forms of complex life or forms of life that depend on photosynthesis, but leave extremophiles unscathed (Cockell, 2003; Hoffman et al., 1998). Perhaps in general, increased biological complexity is inexorably tied to increased fragility (Carlson and Doyle, 2002). In the critical steps model, this can be represented by having the first few critical steps immune to catastrophes, with the later critical steps being vulnerable (or increasingly vulnerable, by having higher catastrophe rates for each subsequent step). Since we assume that the simpler forms of life are immune to the catastrophes, each catastrophe will move the Markov chain from the complex form of life back to the simpler form of life that is immune, rather than start the entire process over from the first transition state.

The second type of model describes a situation in which the earliest life is fragile, but subsequent evolution provides protection from catastrophes. This could represent a hypothesis that the earliest forms of life may have been limited to specific niches with a limited geographic range and population size (or

if the conditions for abiogensis required particular local conditions (Deamer, 2018)), and thus more vulnerable to extinction. Subsequent evolutionary transitions may have then allowed life to greatly expand its range and the number of evolutionary niches that it inhabited, therefore reducing the risk that all life could be destroyed by a single catastrophe. In the critical steps model, this can be represented by having the first few transition states vulnerable to catastrophes, followed by transition states that are all immune (the first Markov chain diagram in the previous section describes a model of this type).

Results for both model types are given in Figure 4.3. The primary result is that critical steps that are vulnerable to catastrophes will occur on a faster timescale that is driven by the rate of catastrophe rather than the rate of evolutionary transition. The higher the catastrophe rate, the faster the vulnerable transitions. If multiple transitions are vulnerable to a catastrophe, they can become tightly clustered and occur on a short timescale.

This can be seen most clearly in the second type of model that represents fragile early life, shown in the first row of Figure 4.3. Here, we show results for a six step model, where the first two transitions are vulnerable to a catastrophe but the subsequent four transitions are immune. All evolutionary transition rates are set to $\lambda = 10^{-3}$, and we assume that a catastrophe reverts the chain to the 0th state, representing a disaster that would destroy fragile early life entirely. When the catastrophe rate is low ($\mu = 5$), the effect is minor and the first three transitions will occur slightly faster than the baseline model without catastrophes. However, as the catastrophe rate increases (to $\mu = 20$ or $\mu = 100$), the first transitions will occur back-to-back in a tight cluster. Moreover, the rapid early transitions provide more time for the subsequent transitions to occur, so that the transitions which are immune to catastrophes will still occur in an evenly spaced manner across the remaining habitable lifetime of the planet. In a six step model with the first two transitions vulnerable and a high enough catastrophe rate, the probability density functions for the transition times closely resemble those of a model with only four critical steps (where

the vulnerable steps occur almost simultaneously with the first non-vulnerable step).

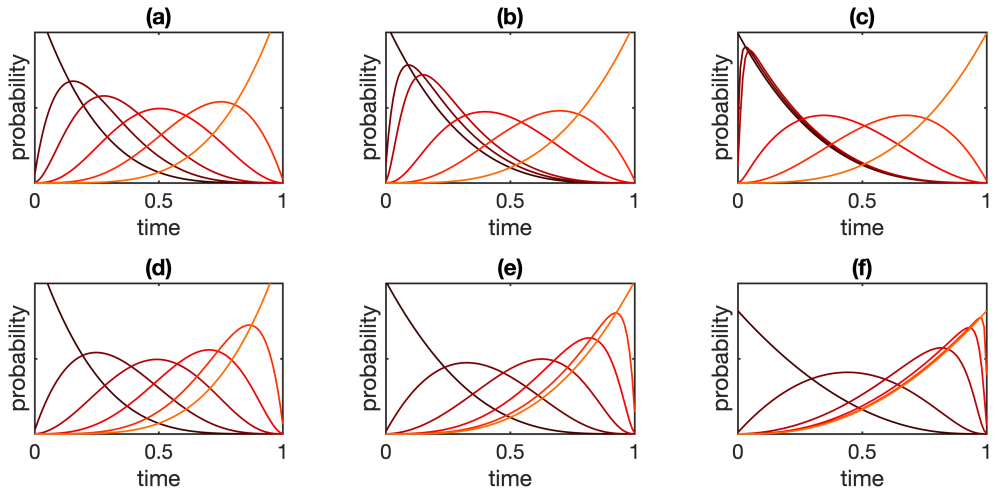


Figure 4.3: A model of six critical steps with catastrophes occurring at a constant rate. First row describes a model in which catastrophes affect the first two transitions, and sets the chain back to the 0th state. Rates of catastrophe move from $\mu = 5$ (a), to $\mu = 20$ (b), and $\mu = 100$ (c). Second row describes a model in which life becomes increasingly vulnerable to catastrophes as it becomes more complex, with the final transition vulnerable to a catastrophe with rates of $\mu = 5$ (d) to $\mu = 20$ (e) and $\mu = 100$ (f). The two transitions preceding the final one have catastrophe rates one third and one ninth of the final catastrophe rate (for the 5th and 4th transition, respectively). Note that the spacing between evolutionary transitions decreases over time.

This opens up an intriguing hypothesis. Watson (2008) argued that since the first three evolutionary transitions described by Smith and Szathmary (1997) had occurred closely together, it was unlikely that more than one of them had an associated timescale longer than the habitability of Earth. Conversely, our model suggests that all three associated timescales could exceed the lifetime of Earth and still be consistent with the data, so long as all three transitions were needed to produce life capable of surviving a stochastic event with a rapid enough rate.

The reason for the accelerated transitions is simple. Instead of just racing against the lifetime of the planet, the vulnerable transitions are also racing

against the possibility of being undone by a catastrophe. The faster the catastrophes happen, the faster the vulnerable transitions must occur in order to avoid such a fate. Once a resilient state has been reached, then the remaining transitions are racing only against the lifetime of the planet, and will conform to the original critical steps model by being evenly spaced out throughout the remaining habitable time left.

We can also examine what happens in a model that treats more complex life as being more vulnerable to catastrophes. In a six step model, we assume that the first two transitions are entirely immune to catastrophes, but that the final transitions are subject to catastrophes of different magnitudes and rates. The most minor catastrophes will only impact the 6th state of the chain, and revert it to the 5th state (representing perhaps a large mass extinction that destroys large land vertebrates, but leaves plenty of smaller invertebrate species unharmed). More severe catastrophes will impact the 5th and 6th state, reverting it to the 4th state, while the most severe catastrophes will revert the 4th, 5th, and 6th states all the way to the 3rd state (perhaps representing a catastrophic event that leaves only extremophiles surviving). We assume that more severe catastrophes occur less frequently than the minor catastrophes, with rates reduced by a factor of five between each step. The transition matrix associated with this model is:

$$\mathbf{Q} = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & 0 & 0 & 0 \\ 0 & -\lambda & \lambda & 0 & 0 & 0 & 0 \\ 0 & 0 & -\lambda & \lambda & 0 & 0 & 0 \\ 0 & 0 & \mu_1 & -(\lambda+\mu_1) & \lambda & 0 & 0 \\ 0 & 0 & \mu_1 & \mu_2 & -(\lambda+\sum_{i=1}^{2}\mu_i) & \lambda & 0 \\ 0 & 0 & \mu_1 & \mu_2 & \mu_3 & -(\lambda+\sum_{i=1}^{3}\mu_i) & \lambda \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{4.8}$$

Note that we still treat the final transition as an absorbing state, reflecting the fact that we are analyzing only the transient behaviour of the Makrov Chain

(since we are interested in the processes leading up to intelligent life that can make scientific observations, as opposed to how long such observers will exist).

The model with $\lambda = 10^{-3}$, $\mu_1 = \mu_3/9$, and $\mu_2 = \mu_3/3$ produces the results in the second row of Figure 4.3, with $\mu_3 = 5$ (d), $\mu_3 = 20$ (e) and $\mu_3 = 100$ (f). As before, when the catastrophe rate is exceptionally high, the transitions that are vulnerable will become clustered closely together and occur right before the final transition. More interesting is what occurs at intermediate catastrophe rates. Since later transitions are vulnerable to more catastrophes, the catastrophe rate effectively increases with each critical step. As each critical transition is racing against the catastrophe rate it is subjected to, the speed of each transition will be greater than the previous one. This leads to a general acceleration in the frequency of critical steps, so that any intelligent observers will see an evolutionary history with many more critical steps occurring in their recent history as opposed to their ancient history. The spacing between transitions will also decrease over time.

This opens up a second intriguing hypothesis. Although life on Earth has existed for over 3.5 Gyr (Awramik, 1992), a staggering amount of complexity has increased in the final 15% of that time (Carroll, 2001), with major changes in the biota of Earth including the Cambrian Explosion ($<$ 550 Mya (Erwin et al., 2011)), the first plants on land ($<$ 520 Mya, (Morris et al., 2018)), the first vertebrates on land ($<$ 350 Mya (Long and Gordon, 2004)), greater encephalization ($<$ 10 Mya (Ruff et al., 1997)), and the emergence of hierarchical language and intelligence itself ($<$ 1 Mya Berwick and Chomsky (2016)). Although past models of critical steps suggested that at most one or two of these transitions (or others around the same time) could be critical steps, the possibility that catastrophes could undo these levels of complexity may account for the sense of accelerating transitions. The catastrophes that could undo all encephalised lineages are a small subset of the catastrophes that could destroy all land vertebrates, which is in turn a subset of the catastrophes that would destroy all complex multicellular life. This in turn means that the effective

catastrophe rate will increase with each step towards intelligence, and thus that each critical step will occur more rapidly than the previous one. This acceleration comes not due to the actual realisation of a catastrophe, but rather the risk of a catastrophe driving the rates faster when subject to observation selection effects. However, we also explore a possible way that catastrophes that occur could accelerate evolution, in part motivated by the theories of punctuated equilibrium versus gradualism (Gould and Eldredge, 1993).

Modelling niche incumbency

We now include evolutionary dead ends into the model. These are represented by adding additional absorbing states to the Markov chain. An evolutionary transition can be precluded if a transition to the dead end state occurs before a transition to the next step towards intelligent life. We assume that transitions to an evolutionary dead end occur at constant rate θ.

As with the catastrophes, the ultimate effect of these evolutionary dead ends is to accelerate any difficult transitions that are competing with an evolutionary dead end. The results are shown in Figure 5.2. In a model with four critical steps where the final step is competing with an evolutionary dead end, the final step will tend to dovetail closely with the third step. If the first two transitions are competing with evolutionary dead ends, they will occur much more rapidly.

The reason is similar to why catastrophes can accelerate a critical step. The critical step needs to occur before falling into an evolutionary dead end, and so is therefore dictated by the timescales associated with the evolutionary dead end rather than the timescales associated with the remaining habitable time on the planet.

A property of the exponential distribution can further clarify this. The time it takes to either complete a critical step or fall into an evolutionary dead end is the minimum of two exponential distributions (one for the critical step, one for the evolutionary dead end). If the distributions have rates λ and θ, then the minimum of the two distributions is also exponentially distributed with rate

$\lambda+\theta$. Since $\lambda \ll \theta$ in this model, the speed at which a critical transition occurs that is competing with an evolutionary dead end will be driven at roughly the rate θ.

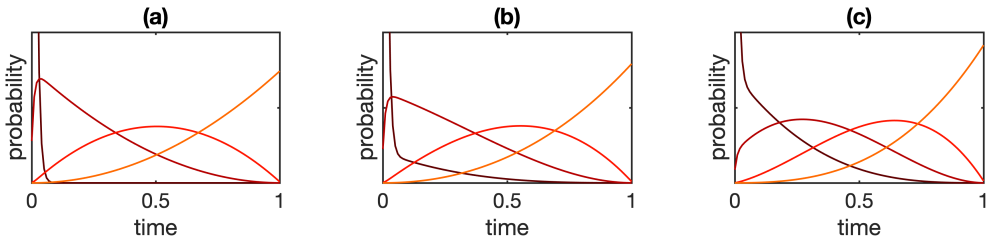


Figure 4.4: A model with four transitions, where the first transition is competing with an evolutionary dead end with rate $\theta = 100$ (a). Subfigures (b) and (c) indicate a catastrophe that can undo the evolutionary dead end with rate $\mu = 1$ and $\mu = 10$, respectively. Without catastrophes, this is close to the situation where the first transition has a rate $\lambda_1 = 100$, but as the catastrophe rate increases, the effect of evolutionary dead ends decreases.

This opens a third hypothesis that critical steps could occur more rapidly if they are competing against a faster evolutionary transition that leads to a dead end. Although speculative, there have been some suggestions that alternative genetic coding systems (e.g. ones that used only two bases instead of four, or those that had punctuation between codons) may have been easier to obtain from periodic chemistry, but with slower downstream rates of evolution or a ceiling on the amount of biological complexity (Sandberg et al., 2018).

We can also add catastrophes that can undo the evolutionary dead ends, reminiscent of various niche incumbency hypotheses (e.g. see Rosenzweig and McCord (1991)), with the canonical example being the hypothesis that an asteroid was needed to remove dinosaurs and open the ecological niche for mammals to eventually evolve intelligence (Cockell and Bland, 2005).

When evolutionary dead ends can be reversed, an evolutionary history that eventually leads to intelligent life may spend time in that dead end state, and increases in the catastrophe rate will reduce the probability of spending a sub-

stantial amount of time in the dead end state. If the catastrophe rate becomes greater than the transition rate compared to evolutionary dead ends (i.e. $\mu \gg \theta$), then the evolutionary dead ends will cease to play a large role and the critical steps will become evenly spaced throughout the Earth's lifetime (Figure 5.2). This only holds true if the catastrophe affects the evolutionary dead ends but not the critical steps. If the catastrophe affects both the evolutionary dead ends and the critical step (which is perhaps a more realistic model), then increasing the catastrophe rate will reduce the amount of time spent in the vulnerable states, as highlighted in the previous section.

Modelling neocatastrophism

We now evaluate the Neocatastrophism Hypothesis using a model in which the rate of catastrophes declines exponentially over time. We run three models, each with five transitions. The first model we treat the first two transitions as being vulnerable, and the final three transitions as being immune to the catastrophes. This model might represent a scenario in which life becomes less vulnerable to GRBs once photosynthesis emerges and a protective ozone layer is established (it could also generally represent threats to early life on Earth, such as the late heavy bombardment). The final two models assume that the first two transitions are immune to catastrophes, but that the final transitions are vulnerable. These models more realistically resemble the impact of GRBs, which would be highly damaging to complex life on land while leaving primitive marine life relatively unscathed (Lsor and Shaviv, 1997; Sloan et al., 2017).

In the first two of these three models, we assume each evolutionary transition has an associated timeframe exceeding the lifetime of Earth. In the final model, we assume that each transition occurs rapidly, with an expected transition time of one 10th of the Earth's lifetime. This final model corresponds to a situation in which intelligent life has a high probability of successfully evolving within Earth's lifetime, with only the catastrophes holding intelligent life back. This model ultimately describes a situation in which the Carter Argument is

rejected, and serves as our primary point of comparison to the Carter Argument and its associated models with exceptionally unlikely critical steps.

We set the initial catastrophe rate to $\mu(0) = 100$, and vary the decay rate parameter with values $t_\gamma = 0.1$, 0.3, and 1. The results are shown in Figure 4.5. These are arbitrary parameter choices that are selected to highlight different model behaviours. Rate estimates for biologically meaningful GRBs vary a substantial amount, ranging from once every 170 Myr for mass extinction level events (Melott et al., 2004) to one every 13.8 Gyr (as calculated from a Poisson model in Olson (2017) using analysis from Piran and Jimenez (2014)). GRBs that would be powerful enough to sterilise the biosphere are thought to be substantially more rare, with probability estimates below 10^{-7} per Gyr (Sloan et al., 2017).

We start by analysing the first model, in which the first two transitions are vulnerable to catastrophes. When the catastrophe rate declines quickly, these transitions will be delayed until the rate falls to a low enough level (to an average of about one catastrophe every 5% of the Earth's lifetime). Evolutionary transitions are then evenly spaced out throughout the remainder of the Earth's habitable lifetime (Figure 4.5.a). When the catastrophe rate declines more slowly, the early vulnerable transitions can be pushed even further back, and the time between vulnerable transitions is reduced (Figure 4.5.b). However, when the catastrophe rate declines so slowly that the rate remains high throughout Earth's entire lifetime, the first three transitions will occur in a tight cluster with higher probabilities towards the beginning of Earth's lifetime (Figure 4.5.c). This can be explained by the balance between the relative improbability of vulnerable transitions occurring early (when the catastrophe rate is high), as opposed to the improbability of final transitions occurring in the time remaining. When the decline in catastrophe rate is fast, the vulnerable transitions can occur later on when the catastrophe rate is low, and still leave time for subsequent transitions. When the decline in the catastrophe rate is

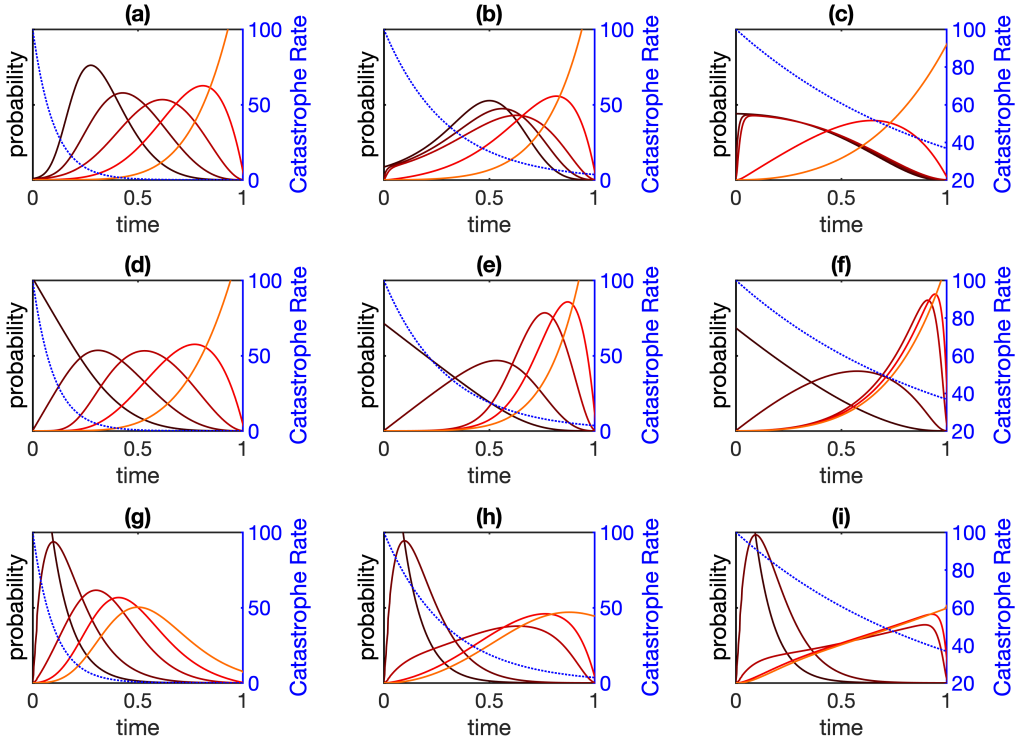


Figure 4.5: Model with five critical steps and a catastrophe rate (denoted by blue curves) that declines exponentially over time as the function $\mu(0)e^{-t/t_\gamma}$. $\mu(0)$ is set to 100 for all models, and $t_\gamma = 0.1$ (all left panels), 0.3 (all centre panels), and 1 (all right panels), resulting in a slower decline of rates as one moves from left to right. In the first row, the first two transitions are vulnerable to catastrophes, with each catastrophe sending the chain to the 0th state. In the bottom two rows, the final three transitions are vulnerable, with each catastrophe sending the chain back to the 2nd state. In the first two rows, evolutionary transitions have rates of $\lambda = 10^{-3}$, in the bottom row, $\lambda = 10$, representing fast rates.

slow, the increased probability that comes with delayed vulnerable transitions is not enough to offset the time needed for the later transitions to occur.

When the first few transitions are unaffected by the catastrophes, but the final transitions are, the dynamics can become substantially different. The first obvious difference is that the first few transitions can occur even when the catastrophe rate is at its highest, and so therefore are not prevented from occurring early. When the catastrophe rate declines quickly, this results in a

model that is hard to distinguish from a standard critical steps model (Figure 4.5.d), since the catastrophe rate will be negligible by the time the vulnerable transitions start to become possible. When the catastrophe rate declines more slowly, a more interesting pattern emerges, with hard evolutionary transitions occurring more towards the end of the interval, in a more tightly spaced manner (Figure 4.5.e). Moreover, the transitions that are unaffected by catastrophes become spaced further apart during the high-catastrophe rate period. When the catastrophe rate declines extremely slowly, the model becomes similar to one with constant catastrophe rates, with a very tight cluster of transitions occurring towards the end of the habitable interval (Figure 4.5.f).

It is important to note that the models we have discussed so far represent a combination of the Neocatastrophism Hypothesis with the Carter Argument, since we still assume that intelligent life requires a handful of critical steps, each of which has an expected transition time greatly exceeding the lifetime of Earth. This is in contrast to the Neocatastrophism Hypothesis as an alternative to the Carter Argument, which would describe a situation in which the timescales associated with intelligent life are substantially shorter, so that the primary bottleneck is the time between catastrophes rather than the total time of planetary habitability. This model results in a high probability of intelligent life emerging once the catastrophe rate has become low enough.

We represent this hypothesis (that there were no critical steps, but rather catastrophes that delayed intelligent life), by adjusting the previous five step model to have fast evolutionary transition rates ($\lambda = 10$, corresponding to each transition taking roughly 10% of the Earth's lifetime). In the absence of catastrophes, intelligence would therefore have a high probability of emerging within Earth's lifetime (about 97%, derived from the Erlang cumulative distribution function evaluated with rate λ and shape parameter n, giving a mean of λn). We then investigate by adding catastrophes, assuming as previously that the first two transitions are unaffected by catastrophes but the final transitions are.

As before, the presence of catastrophes will delay the onset of the final evolutionary transitions. However, there are two noticeable differences with this model and the previous one that had hard critical steps, and may provide avenues for comparing the Carter Argument against the Neocatastrophism Hypothesis. First, the transitions unaffected by catastrophes occur much earlier, and second, the variance in the timing of the vulnerable transitions is higher. The first difference (of early transitions occurring rapidly) is a straightforward consequence of fast transition rates and the fact that these transitions are unaffected by catastrophes. The second difference (of increased variance in the timing of the remaining transitions) is in part due to having more time left since the initial transitions occur quickly. Another key factor is the fact that since the transition rates are more rapid, there is a higher probability that all transitions can occur before a catastrophe, and changes in the catastrophe rate have a smaller impact on the relative probability of making the transitions earlier or later. This results in a substantially greater probability of intelligent life emerging earlier in Earth's history.

We can evaluate this in more detail by comparing how likely a series of transitions is to occur early on under a catastrophe rate μ_1 as opposed to later on under a lower catastrophe rate μ_2. The probability that n transitions with rate λ will occur before a catastrophe with rate μ is $\lambda^n/(\mu+\lambda)^n$. The odds that the transitions occur under a higher catastrophe rate as opposed to a lower one can therefore be analysed as the ratio $(\lambda+\mu_1)^n/(\lambda+\mu_2)^n$. When $\lambda \ll \mu_1, \mu_2$, this ratio is roughly $(\mu_1/\mu_2)^n$. Given this power law relationship, the differences in probability become amplified when λ is small. For example if μ_1 is twice as large as μ_2, then the odds that three transitions will occur under the higher catastrophe rate are only 1-to-8. However, if λ is not small enough to ignore, the odds ratio behaves as $(\lambda+\mu_1)^n/(\lambda+\mu_2)^n$, reducing the effect of differences between μ_1 and μ_2 (e.g. if $\lambda = \mu_1 = 2\mu_2$, the odds ratio from our previous example is reduced to about 1-to-3 rather than 1-to-8). The overall

effect is that when transition rates are faster, the differences in catastrophe rates over time have a weakened impact on delaying evolutionary transitions.

Neocatastrophism vs. Carter Argument

Given that the Neocatastrophism Hypothesis has been suggested as an alternative to the Carter Argument, it is worth comparing these models and and their predictions in light of our findings. The Carter Argument predicts that a small set of critical steps were required for intelligent life on Earth, and that these steps will be spaced out roughly evenly across Earth's lifetime. It also predicts that intelligent life is exceptionally rare, but when it does emerge it does so in the late stages of a planet's lifetime.

If we reject the Carter Argument and instead rely on our Neocatastrophism model to explain the apparent lifelessness of our galaxy, a number of other predictions emerge that seem to fit poorly with some data. The first is that any evolutionary transitions that are unaffected by the catastrophes in question should occur rapidly and early in a planet's lifetime. In our model, this is seen with the first few transitions happening quickly. But given that GRBs will primarily affect life on land or shallow water (Melott et al., 2004), a declining rate of GRBs fails to explain why it took so long for other evolutionary transitions that occurred within the oceans. Eukaryotes took well over one billion years to emerge (Knoll, 2014), and it took another billion years for eukaryotic life to develop into the complex multicellular life found in the Cambrian Explosion (Lenton et al., 2011). The Carter Argument provides a parsimonious explanation for why some of these evolutionary transitions may have taken such a long fraction of Earth's history, beyond explaining why intelligent life emerged in the late stages of Earth's history. In summary, explaining why intelligent life emerged on roughly the same timescale as the lifetime of Earth is insufficient for an alternative hypothesis to the Carter Argument. Any alternative hypothesis would also need to explain why a number of other evolutionary transitions seem to have occurred on roughly the same timescale as the lifetime of Earth,

and the hypothesis that GRBs have stunted biological development on Earth fails to explain this.

The high variance in the arrival time of intelligent life under the neocatastrophism model also undermines the argument that intelligent life would originate concurrently across the galaxy once GRBs have fallen to a low enough rate. Instead, we would still expect some planets to host intelligent life hundreds of millions of years before others. This is because although most series of vulnerable evolutionary transitions occur only once the catastrophe rate is low, a substantial minority of transition sequences can occur more rapidly than usual, occurring even when the catastrophe rate is higher (see the previous discussion on the odds ratios).

Finally, if the primary bottleneck to intelligent life truly was a high catastrophe rate which declined over time, we should perhaps expect to see more dramatic evidence of this in the fossil record. For example, if vertebrates on land had evolved early in Earth's history, followed by hundreds of cycles of extinction and gradual recolonization of land, this would be strong evidence that the extinctions were preventing life from fully exploring the niches that could beget intelligence. One could imagine an even more dramatic fossil record exhibiting an increase in animal brain size over time that was regularly cut short by each catastrophe, before eventually reaching the size of a human brain. However, we instead see that vertebrate life on land did not emerge until very recently (340 Mya) (Long and Gordon, 2004), and that animal body sizes were large enough to have had human-sized brains for for hundreds of millions of years as well (Benson et al., 2014). Although catastrophes may have played a role in the subsequent evolution and radiation of land vertebrates, it is hard to see how catastrophes would have delayed the emergence or have prevented intelligence from originating over 300 million years ago.

We conclude that the Neocatastrophism Hypothesis fails on two fronts. First, although it can explain a delay for transitions that are vulnerable to

GRBs, it fails to explain the delay for transitions that could occur in environments less susceptible to such catastrophes, such as eukaryotic life or the Cambrian explosion. Second, it fails to explain the absence of intelligent civilisations elsewhere in our galaxy. The high variance in the arrival time of intelligent life under a neocatastrophsim model would instead suggest that a non-trivial fraction of intelligent life will emerge hundreds of millions of years before ourselves, contrary to the evidence we see. Conversely, the Carter Argument resolves both of these problems: the long transition times and the absence of intelligent life in our galaxy can both be explained by a series of exceptionally difficult evolutionary transitions.

Conclusions

Intelligent life emerged late in Earth's history, after a series of evolutionary transitions occurring over billions of years. Some of these transitions, such as the emergence of eukaryotes, seem to have occurred after hundreds of millions of years of relative evolutionary stasis. Other transitions, such as the emergence of intelligence, were preceded by a staggering increase in biological complexity.

The Carter Model, which posits a series of exceptionally rare critical steps required for intelligent life, successfully explains why intelligent life emerged late in Earth's history and why some evolutionary transitions may be spaced far apart from other transitions. The model also explains why we fail to see any evidence for intelligent life beyond Earth. However, the model also predicts that rare critical steps will not occur close to one another, which has led some to conclude that the emergence of prokaryotic life had at most one critical step (Watson, 2008) and that no more than one critical step has occurred within the past 500 Myr (Hanson, 1998*b*).

Here, we show that by modifying the Carter Model to include catastrophes or evolutionary dead ends, the timing of critical steps can become accelerated, and the time between critical steps compressed. This opens up new hypotheses

about whether some critical steps have been overlooked by previous literature. In particular, if we assume that increasing biological complexity also leads to increasing vulnerability to catastrophes, the adjusted model predicts that the final transitions will exhibit an accelerating pattern, with each critical step occurring more rapidly than the previous one. This is a striking result, given how much biological complexity has increased in the most recent 500 Myr.

Chapter 5

Bounding the destruction rate of Earth's biosphere

Andrew E. Snyder-Beattie, Anders Sandberg, Michael B. Bonsall

Abstract

Earth's biosphere could be destroyed by a number of natural catastrophes, ranging from large asteroid impacts to runaway climate feedback. Using the fact that Earth has remained habitable for hundreds of millions to billions of years, we show that the rate of all catastrophes that would sterilise the biosphere is almost certainly below 10^{-8} per year, and very likely below 10^{-9} per year, so long as present rates are broadly consistent with past averages. Previous analysis suggested that these inferences could not be made since intelligent life required billions of years to emerge, leaving any intelligent observers with a long lasting biosphere regardless of the true destruction rate. Here, we show that this is true only if no shortcuts exist that could have accelerated the time for intelligent life to emerge. So long as the process leading to intelligent life could be substantially shortened with chance evolutionary breakthroughs, these observer selection effects have a limited impact on destruction rate estimates. This is because if destruction rates were exceptionally high, intelligent life would tend to find itself in biospheres in which intelligence emerged unusually rapidly. Under these conditions, the long track record of Earth's habitability would still

constitute reliable evidence that the biosphere is resilient, rather than the result of perpetual chance and observer selection bias.

Introduction

Earth's biosphere has been inhabited for over 3.5 billion years (Gyr). Yet its habitability could be terminated by any one of a number of events. Such risks include exceptionally large asteroid impacts (Lowe and Byerly, 2015; Segura et al., 2013; Sloan et al., 2017), runaway climate feedback (Kasting, 1988; Kirchner, 2002), nearby supernovae or gamma ray bursts (Horvath and Galante, 2012; Melott and Thomas, 2011), perturbations by nearby stars or black holes (Malmberg et al., 2011), chaotic diffusion of inner solar system orbits leading to planetary collisions (Laskar and Gastineau, 2009), or a collapse of a metastable vacuum (Turner and Wilczek, 1982). There could also be risks that we have not yet discovered (Bostrom, 2002*b*). It is also estimated that in approximately 1 Gyr increasing solar luminosity will make complex life impossible on Earth due to increased weathering which would remove atmospheric CO_2, preclude photosynthesis, and undermine thermoregulation (Caldeira and Kasting, 1992; Franck et al., 2006). How confident can we be that life on Earth will survive the full billion years up until this point and not have habitability cut short by a natural catastrophe? Our goal is to assess the natural rate at which catastrophes could destroy Earth's habitability.

One approach toward this end would be to document specific risks for which we can estimate frequencies and evaluate the associated probabilities of destroying the biosphere. For example, Sloan et al. (2017) examined the probability of asteroid impacts, supernovae, and gamma ray bursts that would be severe enough to render tardigrades extinct. They found that the rate of sterilising asteroid impacts ($> 10^{18}$ kg) can be estimated to a reassuring rate of less than 10^{-14} per year, while the frequency of sterilising supernovae and gamma ray bursts can be estimated to be below 10^{-18} and 3.2×10^{-19} per year, respec-

tively (Sloan et al., 2017)*. However, it would be hard to claim with confidence that such studies have captured all of the possible risks to the biosphere. In addition to a number of more exotic risks, there could be risks that remain unknown to current science. Furthermore, these studies exclude many more mundane risks that could hypothetically snowball into a full destruction event via climate perturbations and runaway feedback, including impacts, volcanism, and biogeneic instabilities (Ward, 2009).

An alternative approach to this risk evaluation is to use the fact that Earth's biosphere has persisted for over 3.5 Gyr. This long track record of survival can be used to produce an upper bound on the rate of destruction from all sources combined, assuming past rates are sufficiently representative of rates in the future. However, this approach is complicated by the possibility of survivorship bias. If many billions of years of biological evolution are required for the existence of intelligent observers, such observers are guaranteed to find themselves in a biosphere with a multi-billion year track record of survival, regardless of the true destruction rate (Tegmark and Bostrom, 2005*a*). Inferring a rate based on that survival time could be misleading unless one had a mechanism to correct for this bias.

Observer selection bias

Observer selection bias is a particular form of sample bias that occurs as observers can only sample data that are consistent with their own existence. These observer selection effects are commonly invoked as a cause for cosmological fine tuning, explaining why we find ourselves in a universe with physical laws compatible with the existence of intelligent life (Barrow and Tipler, 1986; Bostrom, 2007; Carr and Rees, 1979). Accounting for observer selection effects is also important when estimating the rate of catastrophes that could preclude intelligent life (Ćirković et al., 2010). For example, a naive rate estimate for

*The rate of biosphere-affecting but not terminal radiation events is higher, on the order of 10^{-8} per year (Beech, 2011), and similarly there are far more non-terminal asteroid impacts.

large asteroid impacts could be calculated by examining the number of large craters on Earth. But this neglects the possibility that intelligent life may require a period of evolutionary time uninterrupted by large catastrophes. If scientific observers require 50 million years (Myr) of impact-free time, then any rate estimate will be biased if it relies on the past 50 Myr of crater data on Earth, as intelligent observers would not find large craters in that time period regardless of the true catastrophe rate. Alternative data could be used to obtain an unbiased estimate of asteroid impacts, for example using crater data on the moon and on Mars, where presumably the rate of impacts is highly correlated with those on Earth.

In some instances, unbiased data can be obtained for risks that would destroy the biosphere; a comprehensive overview is given by Tegmark and Bostrom (2005*a*). These are most readily available when considering exotic forms of risk that would destroy a planet entirely, for example processes that could form stable 'strangelets' that convert ordinary matter to strange matter (Jaffe et al., 2000). For these risks, the fact no other planet in our solar system has been destroyed in this way over billions of years allows us to infer that the rate of such a catastrophe, if it exists, must be low. The presence of similarly ancient exoplanets constitutes another reassuring source of data that remains unbiased even if the risk were correlated within a solar system, a risk such as a chain reaction of catastrophes among planets. Another unbiased source of data can be found in the fact that Earth itself emerged late in the universe's history when compared to the majority of potentially habitable exoplanets (Lineweaver et al., 2004; Zackrisson et al., 2016). This enables one to put a bound on the rate of catastrophes such as vacuum decay that would destroy large regions of space, since if the rate were high we would expect find ourselves living on one of the first habitable planets (Tegmark and Bostrom, 2005*a*).

Although these data provide a way to bypass observer selection bias, they are only applicable to a limited set of catastrophes that would destroy a planet entirely or otherwise leave evidence of catastrophe on neighboring planets.

Many risks are excluded from this set, notably including risks to the stability of Earth's climate, simply because other planets or solar systems provide very little information on the fragility or resilience of Earth's biosphere to this larger class of risks. Given that alternative sources of unbiased data are not available for this broader set of risks, our aim is to use the Earth-specific data while accounting for potential observer selection bias. Indeed, it has been hypothesised that the Earth's long track record of habitability is in part due to observation selection effects (Waltham, 2007, 2014; Watson, 2004). Using some simplified models for the emergence of intelligent life, we can quantify the amount of observer selection bias given varying biological assumptions. This in turn enables us to use Earth-specific data to estimate a bound for the rate of biosphere destruction from all natural sources combined, and even allowing us to obtain a central estimate for biosphere destruction rates given certain biological assumptions.

Life's history on Earth

The biosphere has persisted since life arose on Earth over 3.5 billion years ago (Gya) (Awramik, 1992). More complex forms of life have also persisted for long periods of time, with eukaryotes estimated to have been present for over 1.6 Gyr (Parfrey et al., 2011), and vertebrate life on land since 340 Mya (Long and Gordon, 2004). This long track record of life on Earth is suggestive of an inherently stable biosphere and low rates of destruction. However, despite the ancient history of life on Earth, intelligent life has emerged only recently. If intelligent life inherently requires a long period of time to emerge, this track record of survival could be deceptive. Here we briefly examine the pathway to intelligent life on Earth and the time it to took reach milestones along the way.

Intelligent life on Earth required a number of preceding evolutionary transitions (Smith and Szathmary, 1997). These transitions include the origin of life itself, along with the origin of photosynthesis, eukaryotes, sexual reproduc-

tion, multicellular life, and the emergence of intelligence itself. Some of these evolutionary hurdles, such as the transition to eukarytoic life, took hundreds of millions of years to cross and seem to be singular events in Earth's history (Koonin, 2011; Lane, 2011), while others, such as multicellularity, seem to have occurred fairly quickly and emerged independently multiple times (Knoll, 2011). An overview of some of the major evolutionary transitions is provided in Table 5.1.

Transition	Timing on Earth
Abiogenesis	> 3.5 Gya Consensus that life existed > 3.5 Gya (Awramik, 1992), more controversial evidence suggesting life as far back as $3.77 - 4.28$ Gya (Dodd et al., 2017). This suggests life first arose between $200 - 900$ Myr after the formation of the Earths oceans at 4.4 Gya (Peck et al., 2001). Despite the early emergence of life, abiogenesis could still be the result of exceptionally rare fluke events (Spiegel and Turner, 2012)
Oxygenic photosynthesis	$3.8 - 2.32$ Gya Consensus that oxygenic photosynthesis was present during the Great Oxygenation Event (GOE) $2.45 - 2.32$ Gya, and ongoing debate as to whether it evolved substantially earlier. Some biomarker evidence suggests it evolved hundreds of millions of years before GOE (Buick, 2008), although this has been disputed and some argue oxygenic photosynthesis could have evolved immediately before the GOE (Rasmussen et al., 2008).
Oxygen accumulation	$2.45 - 2.32$ Gya It may have taken hundreds of millions of years of photosynthetic activity to fill up all near-surface oxygen sinks and start the process of oxygen accumulation in the atmosphere (Buick, 2008; Canfield, 2005). Other hypotheses suggest long fixed periods of photosynthesis were necessary to create an ozone layer that could allow for substantial oxygen buildup (Goldblatt et al., 2006). Alternatively, photosynthesis may not have been required for long periods of time, and could have evolved right before the GOE (Rasmussen et al., 2008), or perhaps about 50 Myr earlier (Anbar et al., 2007).
Eukaryotic life	$2.7 - 1.68$ Gya

	Although some biomarkers have suggested that eukaryotes may have been present as early as 2.7 Gya (Brocks et al., 1999), this evidence has been called into question (Rasmussen et al., 2008). Molecular clocks suggest timings before $1.87 - 1.68$ Gya (Parfrey et al., 2011), supported potentially by fossil evidence (Knoll, 2014).
Sexual reproduction	> 1.2 Gya Sexual reproduction may have evolved rapidly and concurrently with eukaryotic life (Radzvilavicius and Blackstone, 2015) or perhaps even earlier (Garg and Martin, 2016), but unambiguous evidence exists with fossils of red algae at 1.2 Gya (Butterfield, 2000).
Multicellular life	> 1.2 Gya Multicellular life has evolved independently dozens of times (Grosberg and Strathmann, 2007), including some cell differentiation as early as > 2 Gya ago (Tomitani et al., 2006). Conclusive fossil evidence exists with fossils of red algae at 1.2 Gya (Butterfield, 2000).
Cambrian explosion	541 Mya Although the Cambrian explosion was short, evidence suggests that the major animal clades diverged earlier and obtained their core machinery in the $100 - 300$ Myr preceding the Cambrian explosion (Erwin et al., 2011).
Vertebrates on land	340 Mya It took about 25 Myr for fish to evolve into tetrapods (Long and Gordon, 2004). Although aquatic intelligence is possible, the use of tools may require organisms to operate in dry environments (Lingam and Loeb, 2018; Stern, 2016)
Encephalisation	< 10 Mya Over 30% of total encephalisation of humans occurred in just the past 600 kyr (Ruff et al., 1997).
Hierarchical language	$200 - 100$ kya Unique to humans, hierarchical language is thought to be an important and potentially sudden breakthrough that enabled general intelligence and human society (Berwick and Chomsky, 2016).

Table 5.1: Prerequisites for intelligent life

Many of these hurdles may have taken a long time not due to physical constraints, but rather because they were the result of rare chance occurrences. For example, the first eukaryotic cell was likely the result of an extraordinarily rare merger between a bacterium and an archaea, and likely required a substantial amount of luck to have survived in an intermediate chimeric state long enough to produce viable descendants (Koonin, 2011). The billion years that it took for eukaryotic life to emerge therefore seems to be the result of long expected waiting times for a singular rare event to happen, as opposed to a gradual and deterministic accumulation of complexity. While many of the evolutionary transitions seem to have occurred due to rare chance circumstances, others may have been incremental, gradual, and almost inevitable. The gradual accumulation of oxygen after the emergence of photosynthesis is an example of such a transition. Some evolutionary transitions may lie somewhere in between the two extremes of an unlikely hard transition and a gradual but deterministic process. For example, the increase in primate brain size over the past few million years would have required a series of beneficial mutations, each of which built upon a previous one.

It is not clear to what extent intelligent life required time for unlikely breakthroughs, as opposed to time for gradual processes to unfold. This question in part mirrors the more general discussion on theories of punctuated equilibrium versus gradualism (Gould and Eldredge, 1993) in light of many instances of evolutionary stasis in the fossil record, sometimes lasting well over 100 Myr (Bomfleur et al., 2014; Eldredge et al., 2005). We return to this question later, as it turns out to have significant bearing on the potential for severe observer selection bias in our rate estimates.

Bounding the rate of biosphere destruction

The fact that the biosphere has persisted for over 3.5 Gya enables us to put an upper bound on the total rate of biosphere destruction, assuming past rates of

destruction are sufficiently representative of rates in the future.[†] We initially set aside the issue of observer selection effects and the time it would take for intelligent life to emerge, and instead derive an 'naive' upper bound based only on how long the biosphere has survived so far. This naive model serves as a baseline that we can use as a comparison with estimates that account for observation selection effects.

Let L be the random variable denoting the lifetime of the biosphere, and let $\lambda(t)$ be the rate of destruction over time. The probability that the biosphere expires before time t, denoted $P(L < t)$, can be found from the cumulative distribution function $F_L(t)$:

$$F_L(t) = \int_0^t \lambda(\xi) e^{-\Lambda(\xi)} d\xi \tag{5.1}$$

$$= 1 - e^{-\Lambda(t)} \tag{5.2}$$

where $\Lambda(t)$ is the integral over destruction rates for a period of time from 0, the start of the biosphere, up to time t:

$$\Lambda(t) = \int_0^t \lambda(s) ds. \tag{5.3}$$

Conversely, the probability that the biosphere lasts beyond time t is $P(L \geq t) = 1 - F_L(t)$. We calculate the likelihood of destruction rates conditional on the fact that the biosphere has lasted longer than t, such that $L \geq t$. Using different values of t, we can derive different likelihood bounds (Figure 5.1) with the likelihood function

$$\mathcal{L}(\lambda(t)|L \geq t) = P(L \geq t|\lambda(t)) \tag{5.4}$$

$$= e^{-\Lambda(t)}. \tag{5.5}$$

[†]This method is somewhat similar to the Copernican argument advanced by Gott III (1993) to estimate the future lifespan of humanity; however, we exploit extra information about the maximal future lifespan of the biosphere.

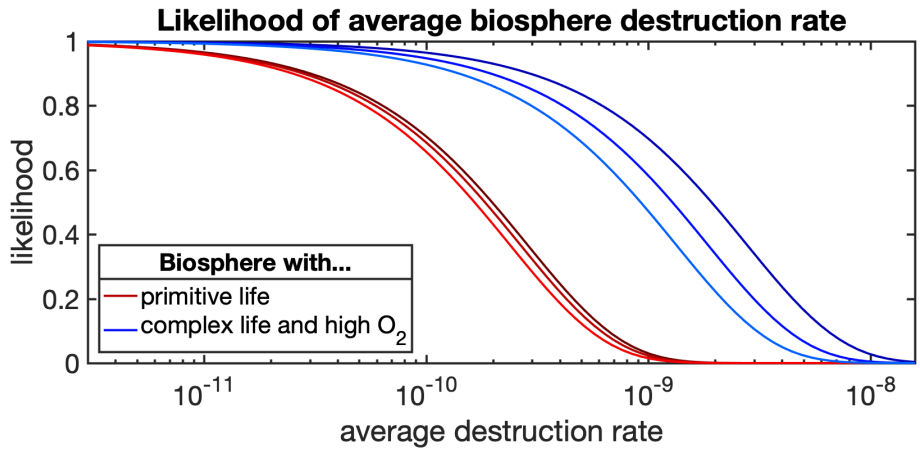


Figure 5.1: Likelihood curves for average biosphere destruction rates, setting aside observation selection effects and using estimated dates for the origin of life (red) and the presence of high oxygen levels and complex animal life (blue). Multiple curves represent different estimates in the literature, darkest curves are most conservative estimates. Rates of events that would totally sterilise the biosphere can be bounded below 10^{-9}, while rates that would eliminate animal life can be bounded below 10^{-8}.

We evaluate the biosphere destruction rate using a likelihood ratio, comparing the likelihood of $\Lambda(t)$ to the likelihood of a baseline $\hat{\Lambda}(t) = 10^{-12}$. We set two bounds, an exceptionally conservative bound with a likelihood ratio of one to a million, and a more moderate bound of one to ten. Since $\mathcal{L}(\hat{\Lambda}) \approx 1$, we can also define this as a relative likelihood, $\mathcal{L}(\Lambda)/\mathcal{L}(\hat{\Lambda})$, with relative likelihood bounds of 10^{-6} and 0.1.

Given that life has lasted continuously for over 3.5 Gyr, this naive model predicts with confidence that any risk that would entirely sterilise the biosphere must have an average rate below about 3.9×10^{-9} per year (corresponding to one expected sterilisation on average every 250 Myr). The relative likelihood of such rates is below one in a million when compared to the baseline average rate of 10^{-12}. Rates above 6.5×10^{-10} per year are also unlikely according to this model (corresponding to one expected sterilisation every 1.5 Gyr), with relative likelihood below one in 10 compared to the baseline average rate of 10^{-12}. We can generalise this result to risks that wouldn't entirely sterilise the biosphere but still destroy all life at a certain level of complexity. For example,

using the fact that the Cambrian explosion happened about 541 Mya and that there has not been a catastrophe since then that has reverted the biosphere to a Precambrian level of complexity, we can obtain bounds for the average rate of destruction: 2.5×10^{-8} per year for likelihood ratios of one in a million and 4.3×10^{-9} per year for likelihood ratios of one in 10 (when compared to a baseline average rate of 10^{-12} per year). These correspond to fewer than one expected event roughly every 40 to 230 Myr on average.

These upper bound estimates could be undermined if past biosphere destruction rates are not indicative of future rates. However, to be meaningful on human timescales, the rate would need to be substantially larger than its past average. For example, in order to have a 1% chance of biosphere destruction over the next 10,000 years, the future destruction rate would need to increase by a factor of over 1,500 when compared to the past average (using the 1 in 10 likelihood bound on the past average). Although we cannot rule out such a possibility with certainty, it seems *a priori* unlikely that a naturally occurring risk would exhibit such steeply increasing rates just around the time that humans emerged (anthropogenic risks are a different matter). The fact that all risks must be consistent with an average destruction rate below 3.9×10^{-9} over the biosphere's history provides a strong constraint on any proposed risk to the habitability of Earth. But another way that our upper bound could be undermined is if our data on the biosphere's track record of survival is not representative due to observation selection effects.

Deriving bounds with observer selection effects

The likelihood function can be adjusted to account for observation selection effects. We do this by conditioning the rate likelihood not only on the data that the biosphere has persisted for a certain period of time, but also on the fact that intelligent life was present in that biosphere. This corrects for the sample bias experienced by the observers.

Let V be the random variable denoting the time that it takes for intelligent life to arise after the start of the biosphere. The probability that intelligent life has occurred before time t is given by the cumulative distribution function $F_V(t)$. We now must condition our estimates on the fact that observers occurred before the biosphere collapsed. The adjusted likelihood function, denoted $\mathcal{L}^\star$, evaluates the likelihood of a destruction rate $\lambda(t)$, given the lifetime of the biosphere *and* the fact that intelligent observers must have emerged before the destruction of the biosphere. The likelihood function is:

$$\mathcal{L}^\star(\lambda(t)|L \geq t) = P(L \geq t|L > V, \lambda(t)) \tag{5.6}$$

$$= P(L \geq t, L > V|\lambda(t))\ P(L > V|\lambda(t))^{-1} \tag{5.7}$$

$$= \int_t^\infty f_L(\xi)F_V(\xi)d\xi \left(\int_0^\infty f_L(t)F_V(t)dt\right)^{-1} \tag{5.8}$$

where the second term is a normalising constant equal to $P(L > V)$. Intuitively, this is the probability that the biosphere survives beyond time t given a rate of destruction, but truncated and renormalised to exclude the scenarios where the biosphere is destroyed before intelligent life arises. Although this method is fully general, we make a simplifying assumption in the following sections that the destruction rate is constant over time; we use in the definition of the density $f_L(t)$ a constant parameter λ instead of $\lambda(t)$. The probability that intelligent life emerges before time t, defined by the function $F_V(t)$, could be described by different processes.

Models for the pathway to intelligent life

Let us assume that intelligent life emerges after a series of n steps (or 'transitions'). These steps could represent key evolutionary transitions, the result of steady evolutionary innovation, or a change in environmental conditions. We assume that the steps must occur in sequence, so that the first step must come before the second, the second before the third, and so on. Let the random

variable S_i be the time it takes between the ith step and the previous one, and let t_i denote the time of the ith step. This includes the time taken to reach previous steps, so that $S_i = t_i - t_{i-1}$. The time it takes to obtain intelligent life is therefore the sum of transition times, $V = \sum_{i=1}^{n} S_i$. We examine several models for the distribution of each transition S_i, allowing us to derive a cumulative distribution function for V, the time it takes for intelligent life to emerge.

Stochastic step model

The first model represents an evolutionary transition that occurs through a single breakthrough (such as the endosymbiont merger resulting in eukaryotic life). If we assume the probability of such a breakthrough is constant over time, each S_i will follow an exponential distribution with rate θ_i. Let V_θ denote the time it takes for intelligent life to emerge under this model, so that $V_\theta = \sum_{i=1}^{n} S_i$. The cumulative distribution for V_θ is the same as the distribution of the final transition and can be found by integrating over the possible times of all previous transitions:

$$F_{V_\theta}(t_n) = \int_0^{t_n} \int_0^{t_n} \cdots \int_0^{t_2} \prod_{i=1}^{n} \theta_i e^{-\theta_i(\xi_i - \xi_{i-1})} d\xi_1 \ldots d\xi_{n-1}\ d\xi_n \tag{5.9}$$

We can also generalise the model of each step to represent transitions that do not occur as a single step, but rather as a gradual accumulation of more likely chance events (for example the process of encephalisation). We can do this by representing a transition not as a single S_i, but rather as a series of S_i consisting of sequential steps. The functional form of V_θ therefore remains unchanged if the appropriate number of steps are added.

Fixed time process

An alternative model would be a 'slow but inevitable' model that represents a transition bound to occur after a certain fixed amount of time. One example

could be the accumulation of oxygen in the atmosphere after the arrival of photosynthesis. For these steps, chance events do not accelerate or slow the time that the step occurs. If the step S_i takes a fixed amount of time τ_i, the cumulative distribution function of S_i will equal $\mathbb{1}_{[t>\tau_i]}$, the characteristic function denoting probability 0 for the step occurring before time τ_i and probability 1 for the step having occurred after time τ_i.

If every transition on the path to intelligent life is the result of a fixed time process, the total amount of time it takes for intelligence to emerge will also be a fixed amount of time. Let V_τ denote this degenerate random variable, so that the cumulative distribution function of V_τ will be $F_{V_\tau}(t) = \mathbb{1}_{[t>\sum_{i=1}^{n}\tau_i]}$, taking on the probability 0 for times below the sum of all fixed transition times, and probability 1 for any time beyond that same sum of transition times.

Hybrid model

In reality, it is likely that intelligent life required time for both gradual processes and unlikely breakthroughs. We can combine these models to calculate the probability that intelligence will have emerged by a certain point in time when n steps occur as a sudden breakthrough and m steps occur as gradual processes. Let $V_\theta = \sum_{i=1}^{n} S_i^{\times}$, the sum of the exponentially distributed steps, $V_\tau = \sum_{j=1}^{m} S_j^{\star}$, the sum of the fixed time steps, and $V = V_\theta + V_\tau$. We can calculate the cumulative distribution function of V under the hybrid model using the convolution:

$$
\begin{aligned}
F_V(t) &= P(V_\theta + V_\tau \leq t) && (5.10)\\
&= \int_0^\infty \left(\int_0^{t-x} f_{V_\theta}(x) f_{V_\tau}(y) dy \right) dx && (5.11)\\
&= \int_0^\infty f_{V_\theta}(x) F_{V_\tau}(t-x) dx && (5.12)\\
&= \int_0^{t-\tau^\star} f_{V_\theta}(x) dx && (5.13)\\
&= F_{V_\theta}(t-\tau^\star) && (5.14)
\end{aligned}
$$

using the fact that $F_{V_\tau}(t-x)$ takes on the value 1 when $t - x > \tau^\star$ and 0 otherwise, where $\tau^\star = \sum_{j=1}^{m} \tau_j$, the sum of all fixed times. Intuitively, this is simply the same distribution as the stochastic step model, but with all values shifted by the total amount of fixed time required for intelligent life.

Results for bound with observer selection effects

We evaluate the likelihood of rates produced for each of these models across a wide range of parameters that create observation selection bias (Figure **??**). The faster the expected emergence time for intelligence, the less biased our destruction rate estimates become. If the expected time for intelligence is instantaneous, there is no bias at all, which can be verified by taking the limits as all rates go to infinity in the stochastic step models, or all fixed times go to zero in the fixed time models. This is equivalent to taking the limit as V goes to 0, resulting in the naive likelihood function from the previous section:

$$\lim_{V \to 0} \mathcal{L}^*(\lambda | L \geq t) = \lim_{V \to 0} P(L \geq t | L > V, \lambda) \tag{5.15}$$

$$= P(L \geq t | \lambda) \tag{5.16}$$

$$= \mathcal{L}(\lambda | L \geq t) \tag{5.17}$$

Conversely, as the expected emergence time for intelligence increases, the bias increases. However, as we increase the amount of expected time it takes for intelligence to emerge, the bias emerges in different ways depending on the model. For the stochastic step model, one can increase the expected time it takes for intelligence to emerge by slowing the rates at which key evolutionary transitions occur, or increasing the number of transitions that need to be traversed. In the fixed time model, it is simply a matter of increasing the amount of fixed time required to reach intelligence. All three of these adjustments can be done in the hybrid model.

We first examine the effect of slowing transition rates for a model with three transitions (Figure **??**.a). As transition rates slow to the point where intelligence is expected to take almost as long as our current track record of survival, destruction rates that were previously ruled out by the biosphere's track record of survival become more plausible. For example, without accounting for selection bias the likelihood ratio of rates of 6.5×10^{-10} compared with 10^{-12} were 1 to 10, but with three transitions each expected to take 1 Gyr, this likelihood ratio goes to about 4 in 10. Interestingly, this effect is limited.

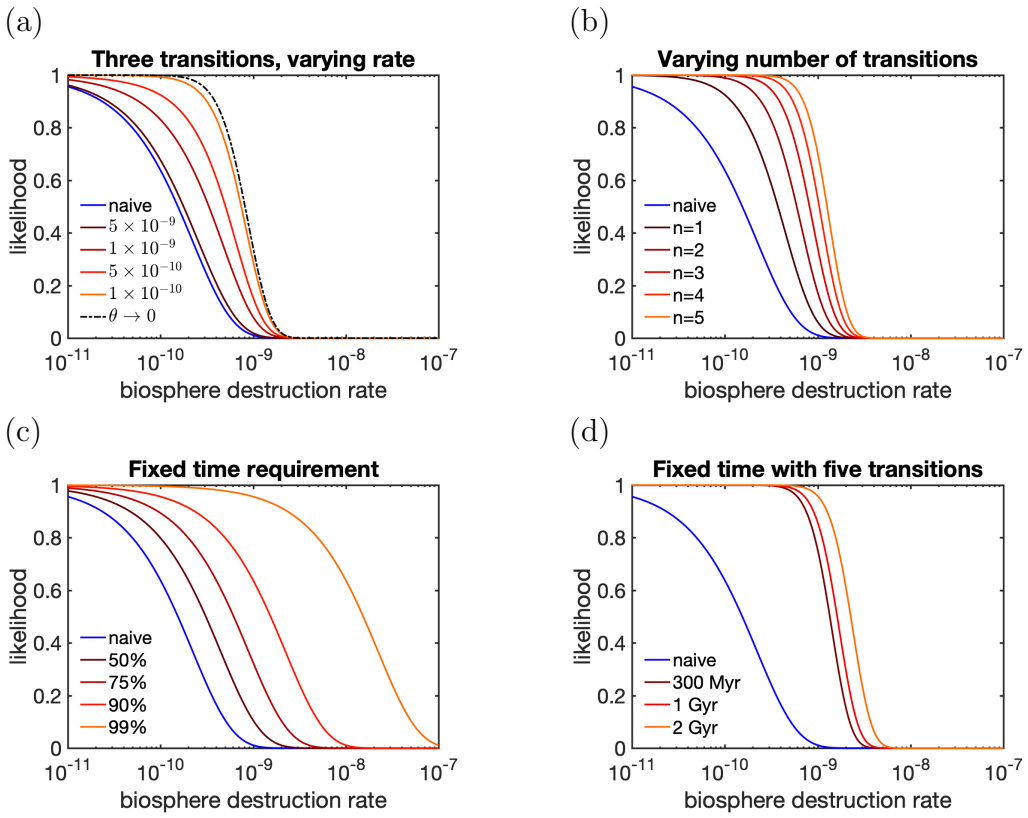


Figure 5.2: Biosphere extinction rates with observer selection effects. (a) as the rate decreases, the bias increases, but only up to a certain point (b) as number of hard transitions increases, bias increases, (c) fixed time produces significant bias (d) a hybrid model with five transitions and three amounts of fixed time: 300 yr, 1 Gyr, and 2 Gyr.

Even in the limit as the rates go to 0 (so that each evolutionary transition

would be expected to take an infinite amount of time), the bias stays bounded. For example, the 1 in a million likelihood bound for such a model is a rate of 6.1×10^{-9}, differing by less than a factor of two from the naive model bound of 3.9×10^{-9}. Although there is bias introduced by a handful of exceptionally unlikely evolutionary transitions, it does not eliminate most of the information provided by a long track record of biosphere survival.

The reason for this is that if destruction rates were exceptionally high, intelligent life would tend to find itself within biospheres in which intelligence emerged exceptionally early. In general, if biospheres are short lived and the expected time it takes for intelligence to emerge is long, intelligent life is an uncommon occurrence. When intelligence does arise under these models, it is because either 1) the biosphere lasted an unusually long period of time, or because 2) intelligent life emerged unusually rapidly. The relative strength of these two factors determines the extent to which observer selection effects bias estimates of biosphere destruction rates. If intelligence required a handful of stochastic breakthroughs that could have occurred much earlier in the biosphere's lifetime, then we can take the long survival time of the biosphere close to face value.

Conversely, if no 'lucky shortcuts' exist and intelligent life fundamentally requires many billions of years to evolve, then the biosphere's track record of survival becomes a less reliable indicator of a low destruction rate. This is best illustrated with the fixed time model. If intelligence requires an fixed 3.4 Gyr minimum, then a 3.5 Gyr biosphere lifetime can only be treated as a 100 Myr track record of survival (Figure **??**.c). One could not even confidently rule out the possibility that biosphere sterilising events happen with an average frequency of once every 50 Myr, as such a model would have a likelihood of about 0.14, with a likelihood ratio of about one in seven when compared with a rate with an average catastrophe frequency of 10 Gyr. Although not as extreme as the fixed time model, this effect can be seen as one increases the number of

required sequential transitions to get intelligence (Figure **??**.b).[‡]

However, even a substantial amount of fixed time is insufficient to create enough bias to be meaningful on human timescales. If we assume that intelligent life required a fixed time equal to half of the biosphere's track record of survival, our bounds on the risks will only be impacted by a factor of two, leaving us with a 1 in a million likelihood bound of 7.9×10^{-9} rather than 3.9×10^{-9}. If we assume that the evolution of intelligent life required 2 Gyr fixed time and 5 extremely difficult transitions (with transition rates going to 0), the adjusted 1 in a million likelihood bound is within a factor of 4 of the naive bound that ignores sample bias (Figure **??**.d). In order to have a higher than 1 in 10 likelihood ratio for rates high enough to put a 1% chance of destruction in the next 100 Myr, the amount of fixed time would need to be about 3 Gyr. How plausible is such a large amount of fixed time? Our view is that such a large amount of fixed time is unrealistic, a position we discuss before concluding. First, we examine another interesting result.

Predicting rates based on time to intelligence

If we assume that a handful of evolutionary transitions were exceptionally difficult, so that their expected transition time greatly exceeds the lifetime of Earth, then an interesting predictive relationship emerges between the number of such 'hard' evolutionary transitions and the remaining lifetime of the biosphere. On biospheres where all evolutionary transitions successfully occur within the lifetime of the biosphere (e.g., on biospheres where intelligence successfully emerges), these evolutionary transitions will tend to be distributed uniformly over that interval (Carter, 2008). Moreover, the expected amount of time left between the final transition and the destruction of the biosphere will be $L/(n+1)$, where L is the lifetime of the biosphere and n is the number

[‡]It is worth noting that the fixed time model is a limiting case of the stochastic step model, as the number of transitions goes to infinity and the time between transitions goes to zero.

of hard evolutionary transitions (Carter, 1983, 2008; Hanson, 1998*b*; Watson, 2008).

This relationship was first discovered by Carter (1983), and has subsequently been used to infer the number of hard evolutionary transitions based on the assumption that the biosphere would end in a fixed amount of time due to the increasing luminosity of the Sun (Carter, 2008; McCabe and Lucas, 2010; Watson, 2008). Here, we investigate the reverse problem suggested by Barrow and Tipler (1986): predicting the destruction rate of the biosphere if we assume the biosphere could end at random and a certain fixed number of hard evolutionary transitions were required for intelligent life.

We do this by evaluating the likelihood of a destruction rate given that 1) $L > t$, the lifetime of the biosphere has been greater than observed time t, and 2) $V = t$, that the final transition to intelligent life occurred at time t. We must also condition on $L > V$, such that the biosphere is not destroyed before intelligent life emerges.

Intelligence arose at a specific time, and the biosphere has lasted beyond that point in time, conditioned on the fact that intelligence must have arisen before the end of the biosphere:

$$\begin{aligned}
\mathcal{L}^{\star}(\lambda|V = t, L > t) &= P(V = t, L > t|L > V, \lambda) && (5.18)\\
&= P(L > t|V = t, L > V, \lambda)P(V = t|L > V, \lambda) && (5.19)\\
&= P(V = t|L > V, \lambda) && (5.20)\\
&= f_V(t)\Big(1 - F_L(t)\Big)\left(\int_0^{\infty} f_V(t)(1 - F_L(t))dt\right)^{-1} && (5.21)
\end{aligned}$$

where f_V and F_V denote the probability and cumulative distribution functions for V, respectively, and f_L and F_L denote the probability and cumulative distribution functions for L, respectively. As before, we assume F_L follows an exponential distribution, and that each evolutionary transition is also exponentially distributed, resulting in F_V following an Erlang distribution with rate θ approaching 0 for the hard evolutionary transitions. these components become:

$$P(V = t|L > V, \lambda) = \lim_{\theta \to 0} \left(\frac{\theta^n}{(n-1)!} t^{n-1} e^{-\theta t} \right) e^{-\lambda t} \left(\frac{\theta + \lambda}{\theta} \right)^n \tag{5.22}$$

$$= \frac{\lambda^n t^{n-1} e^{-\lambda t}}{(n-1)!} \tag{5.23}$$

which interestingly is an Erlang distribution with the destruction rate parameter λ taking the place of the transition rate parameter θ. We can also adjust this rate estimate to account for the fact that another known process (the increasing luminosity of the Sun) will terminate the habitability of Earth in a fixed amount of time, which we represent with the parameter γ. Let $(L \vee \gamma)$ denotes the minimum value of L and γ. This results in the following likelihood function:

$$\mathcal{L}^\star(\lambda|V = t, L > t) = P(V = t, L > t|(L \vee \gamma) > V) \tag{5.24}$$

$$= P(V = t, L > t|L > V, \gamma > V) \tag{5.25}$$

$$= P(L > t|V = t, L > V, \gamma > V) P(V = t|L > V, \gamma > V) \tag{5.26}$$

$$= P(V = t|L > V, \gamma > V) \tag{5.27}$$

$$= f_V(t) \Big(1 - F_L(t) \Big) \mathbb{1}_{[t<\gamma]} \left(\int_0^\gamma f_V(t)(1 - F_L(t))dt \right)^{-1}. \tag{5.28}$$

We can also introduce a fixed amount of time that must transpire between hard evolutionary transitions, adjusting $f_V(t)$ to include the total amount of fixed time represented with the parameter τ. We run six variations of the model: with and without an assumption that the increasing luminosity of the Sun will make Earth uninhabitable for intelligent life in 1 Gyr, and with three differing amounts of fixed time in between evolutionary transitions 0 Gyr, 2 Gyr, and with 95% of the biosphere's track record of survival being evolutionary fixed time (Figure 5.3).

If we assume that the only way the biosphere can end is through a catastrophe with rate λ, we can directly estimate λ using only knowledge of the number

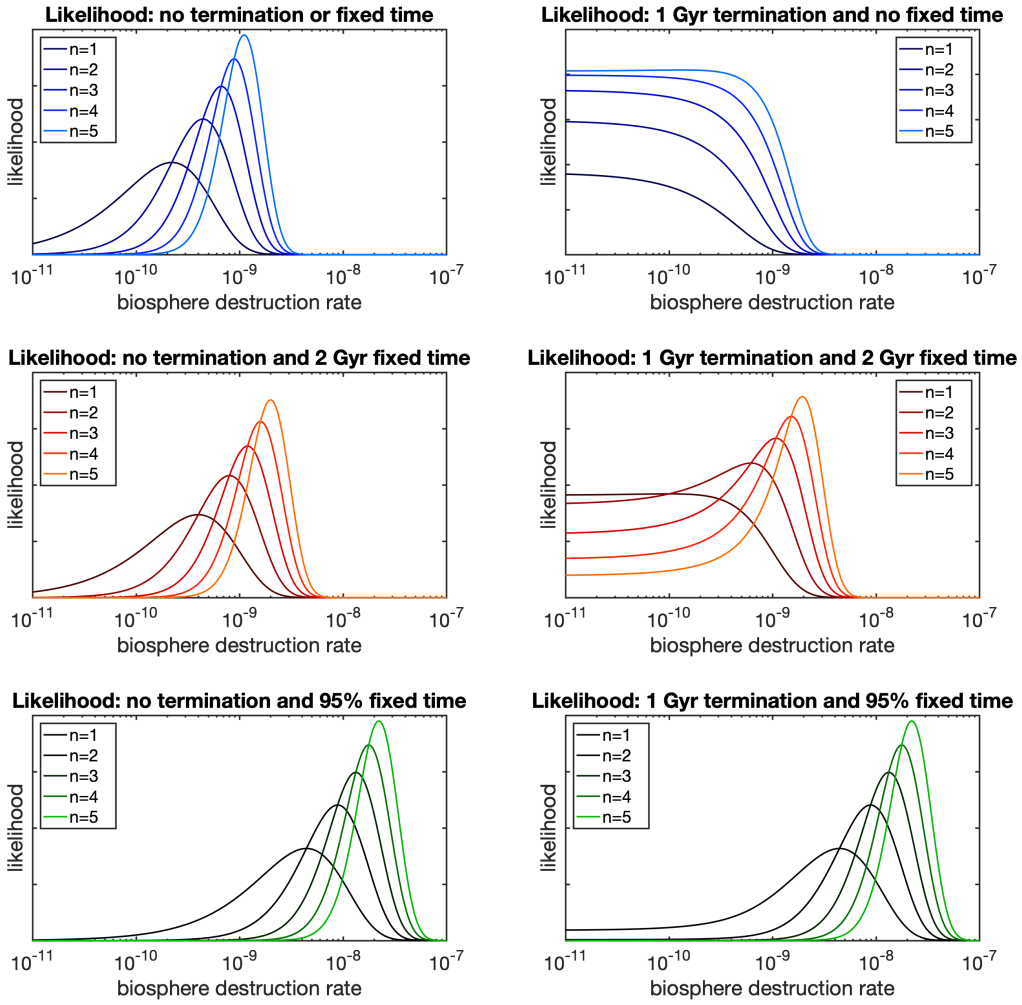


Figure 5.3: Figures on the left show destruction rate estimates given n hard transitions and an infinitely long possible biosphere. Figures on the right indicate rate estimates given n transitions and the fact that the biosphere will end in 1 Gyr regardless. Blue indicates an additional 2 Gyr of fixed time on top of the n transitions, green indicates extreme example with 95% of our evolutionary history being fixed time in addition to n transitions.

of difficult evolutionary transitions in our past and the fact that it took 4.5 Gyr for all evolutionary transitions to occur (Figure 5.3.a). If we assume between two and six difficult transitions, the estimates center around a median value of about $\lambda = 10^{-9}$. This is also the case if we assume 2 Gyr of fixed time in addition to the evolutionary transitions (Figure 5.3.c).

Rates exceeding 5×10^{-8} are ruled out because if the rates were this high, intelligent life would have emerged much more rapidly on Earth. However,

exceptionally low rates are ruled out for the opposite reason. As the hard evolutionary transitions have rates that are exceptionally low, intelligence will be unlikely to emerge in the first half of any biosphere's lifetime, and extremely unlikely to emerge in the first one percent or so. If the rate of biosphere destruction were exceptionally low (e.g. on the order of 10^{-12}), then we should also expect intelligent life to emerge much later. The fact that intelligent life took 4.5 Gyr to emerge is therefore indicative of a biosphere that will likely expire on roughly the same timescale, assuming intelligence required a handful of evolutionary transitions that were exceptionally unlikely.

However, the biosphere need not expire due to a chance catastrophe. We know that the increasing luminosity of the Sun will make complex life impossible on Earth in about 1 Gyr. After accounting for this, the likelihood of extremely low destruction rates increases (Figure 5.3.b). Accounting for 2 Gyr of fixed time does not materially change this result, as all destruction rates below 10^{-10} have nonzero likelihoods (Figure 5.3.d). The fact that intelligent life did not take substantially longer than 4.5 Gyr to emerge could be due either to a destruction rate on the order of 10^{-9}, or simply because the biosphere is approaching the end of Earth's natural lifetime and any intelligent life would need to emerge within that window.

However, if we assume that 95% of the time the biosphere existed was a fixed time required for intelligent life, and that a certain number of extremely difficult evolutionary transitions still needed to have occurred in addition to this fixed time, then we would have reason to believe that the biosphere would be destroyed long before the increased luminosity of the Sun does so (Figure 5.3.e, Figure 5.3.f). The reason for this is that these hard transitions would have occurred within an extremely narrow time window, suggesting that the total window available for them to occur could not be substantially longer. In this scenario, we would have good reason for thinking not only that high rates of biosphere destruction are compatible with a long track record of survival

due to observation selection effects, but also that high rates were substantially more likely than low rates.

We suggest in the next section that such an extreme amount of fixed time seems unlikely. But even with 3.5 Gya of fixed time, extremely low rates of biosphere destruction are still assigned more likelihood than rates between 10^{-9} and 10^{-8}. For less extreme amounts of fixed time and number of hard evolutionary transitions, the fact that it took 4.5 Gyr for intelligence to emerge suggests that the biosphere will likely end due to the increasing luminosity of the Sun in about 1 Gyr, and not due to an earlier catastrophe.

Is the Earth unusually long-lived?

We have earlier emphasised that estimates of biosphere destruction are unlikely to be biased severely enough to be meaningful on human timescales (e.g. off by more than two orders of magnitude). However, our results are consistent with the hypothesis that Earth's biosphere is unusually long lasting due to observation selection effects (Waltham, 2014; Watson, 2004). Specifically, given a destruction rate λ and a number of hard evolutionary transitions n, we found that the time it takes to reach intelligent life (conditional on intelligent life emerging within the lifetime of the biosphere) is Erlang distributed with parameters λ and n. The expected value of this distribution is n/λ, so although a typical biosphere with destruction rate λ will have an expected lifetime of $1/\lambda$, the expected lifetime of a biosphere that happens to harbour intelligent life will be $(n+1)/\lambda$. Given a series of hard evolutionary transitions, we should therefore expect our biosphere to be longer lasting by a factor of roughly $n+1$ when compared to a 'typical' biosphere that does not end up hosting intelligent life.

Could intelligence have emerged earlier?

Our primary result is that there is a strong link between the estimated rate of biosphere destruction and the extent to which intelligent life could have emerged earlier in Earth's history. If the pathway to intelligent life was immutable, such that the steps leading to intelligent life strictly required well over 3 Gyr, then we can not rule out the possibility of high biosphere destruction rates (and when hard evolutionary transitions are needed in addition to exceptionally long fixed time, we should actually expect high destruction rates!). Conversely, if the pathway to intelligent life involved a substantial amount of contingency, such that intelligent life could have occurred earlier in the biosphere's history with chance breakthroughs, then the observer selection bias is limited and we can reliably rule out high rates of biosphere destruction. This is because if the destruction rate were high, intelligent life would tend to find itself in biospheres in which it emerged unusually early, rather than on biospheres that lasted for an unusually long period of time. A long track record of biosphere survival could then be taken at close to face value.

Could intelligent life have emerged substantially earlier in Earth's history? Obtaining a conclusive answer is difficult, but our tentative view is that at least 10% of the biosphere's history, and possibly much more, was not strictly required for intelligent life to emerge. Although some geological processes and prerequisites for intelligent life could be immutable (perhaps the most notable example being the buildup of oxygen in Earth's atmosphere), biological evolution is subject to an exceptional amount of contingency and likely allows for many possible shortcuts to intelligent life. Adding up these shortcuts could substantially reduce the amount of time it takes for intelligent life to emerge.

Certainly in more recent history, it is hard to imagine any fundamental reason preventing intelligence from emerging soon after the arrival of the first large vertebrate animals on land ($\approx$ 340 Mya) (Clack, 2006) or the dinosaurs (Russell et al., 1982) ($\approx$ 233 Mya) (Langer et al., 2018). Over 30% of human

brain size increase occurred during a span of only 600 kyr (Ruff et al., 1997), suggesting a rapid pathway to intelligent life that comprises a minuscule fraction of the total evolutionary history on Earth. The stem amniote ($\approx$ 314 Mya), the ancestor of mammals, dinosaurs, and birds, likely had a cerebral cortex/pallium (Goffinet, 2017), and one can imagine a counterfactual evolutionary history where an intelligent cynodont lineage evolved in the Permian. If intelligent life was possible 300 Mya, then at a minimum this seems to provide a modest track record of survival even if we pessimistically assume there were no possible shortcuts for all previous evolutionary or geological processes leading to land vertebrates or the stem amniote.

More generally, it seems likely that there were possible shortcuts in life's early history. Many of the key evolutionary transitions that preceded the Cambrian explosion, such as the origin of eukaryotic life, seem to be the result of a chance coincidence rather than the result of an incremental process. But even if incremental evolutionary processes were the primary rate limiting factor, the pathway leading to intelligent life could still have contained shortcuts. As demonstrated by adaptive radiations, the pace of evolution can change dramatically due to environmental changes (Schluter, 2000), and even small variations in evolution rates could induce significant time variability over the long run. Evolutionary rates vary over at least a factor of 10 within lineages (Kinnison and Hendry, 2001), and even a 10% change in evolution rate during the "boring billion" 1.8-0.8 Gya would imply a possible shortcut of 100 million years. If the pathway to intelligent life were as fast as possible, we might also expect to see unusually rapid rates of evolution in the lineage leading to intelligent life. The fact that primates exhibit rates of DNA sequence evolution that are less than or equal to other lineages suggests that this hypothesis is not supported (Britten, 1986; Kumar and Subramanian, 2002).

One hypothesised source of time delay for intelligent life is the time it took for animal body size to increase 16 orders of magnitude (Simpson, 2017). However, even this step seems like it could occur fairly rapidly. While there appear

to exist rate limits to the speed with which mammals can evolve large body sizes, (Evans et al., 2012), rates appear to increase in substantial jumps (Mattila and Bokma, 2008) and are highly variable (Benson et al., 2014), dependent on highly developed ecological infrastructure (Vermeij, 2016) and fast lifestyles (Okie et al., 2013). We also know experimentally from agriculture that strong consistent selection can produce substantial size increases in animals in just a few decades (Zuidhof et al., 2014).

Oxygen accumulation as rate limiting step?

Although many key evolutionary transitions seem to have been the result of contingency and could have much happened earlier, there are other processes that may have inherently taken a long time to unfold. Foremost among possible sources of fixed time is the gradual buildup of oxygen in the atmosphere. High levels of oxygen seem to be a universal prerequisite for complex life with high energy demands (Catling et al., 2005), but primordial Earth likely had negligible oxygen (Anbar et al., 2007; Kasting, 1987; Lyons et al., 2014). And although photosynthesis may have evolved early in Earth's history, oxygen levels remained at 10^{-5} of the present atmospheric level until the Great Oxidation Event (GOE) 2.4 Gya when it jumped to more than 0.01 of the present atmospheric level. It may have taken hundreds of millions of years for photosynthetic activity to fill up all near-surface oxygen sinks and start the process of oxygen accumulation in the atmosphere (Buick, 2008; Canfield, 2005). Other hypotheses also suggest that oxygen was removed quickly in a high UV environment until enough reductant had been removed, allowing ultraviolet shielding due to accumulated O_3 in the stratosphere that quickly shifted the equilibrium to a higher oxygen state (Goldblatt et al., 2006). Although some question the evidence that oxygenic photosynthesis evolved long before the GOE (Rasmussen et al., 2008), the gradual accumulation of oxygen is one plausible rate-limiting step for intelligent life that does not permit shortcuts (Catling et al., 2005;

McKay, 1996). Moreover, oxygen levels appear to have remained fairly constant from the GOE until 0.8-0.5 Gya when they began rising toward present levels due to as yet unclear reasons (Fike et al., 2006; Lyons et al., 2014).

Could oxygenation on Earth have occurred in fewer than 3.9 Gyr? Some factors impacting oxygenation are inexorably linked to contingent biological innovations, such as the emergence of eukaryotic algae (Lenton et al., 2014) or the presence of land plants (Lenton et al., 2016). If these innovations had emerged earlier, one could plausibly imagine a faster oxygenation time. However, even if we pessimistically assume that the 3.9 Gyr planetary oxygenation time could not have been compressed, we can still derive bounds with the remaining 360 Myr track record of survival, producing risk bounds that are still reassuringly small on human timescales: 6.4×10^{-9} and 3.8×10^{-8} for relative likelihoods of 0.1 and 10^{-6} when compared to a rate of 10^{-12}, respectively.

Conclusions

Despite the manifold ways that our biosphere could be destroyed, the probability of such a catastrophe is low. The risk of catastrophes that would sterilise the biosphere is almost certainly below 10^{-8} per year. Our estimates are unlikely to be substantially distorted by observation selection effects, since if the catastrophe rate were much higher, we would expect intelligence to emerge earlier in Earth's history. Observation selection bias becomes a significant factor only if the evolutionary processes leading to intelligence were immutable, such that intelligence could not have emerged much earlier. But aside from the gradual buildup of oxygen in the atmosphere, it seems that most of the key evolutionary transitions leading to intelligence required a substantial amount of contingency rather than being the result of a gradual but inevitable process.

Our upper bounds on biosphere destruction rates are compatible with previous estimates found in the literature, including rates for sterilising asteroid impacts (Sloan et al., 2017), gamma ray bursts (Beech, 2011) and vacuum de-

cay (Tegmark and Bostrom, 2005*a*). Our approach has two advantages. The first is that the bound is extremely wide in scope, applying to any possible risk for which the past rate of destruction is indicative of future rates, even extending to risks as yet unknown to science. The second is that the methods are different than others used in the literature, enabling an independent form of validation. In situations where the probability of a risk is exceptionally small, finding alternative methods for cross validation of a model becomes especially important, as a small risk of a model error can still dominate otherwise low risk estimates (Ord et al., 2010). Finding an independent way of verifying previous estimates allows more confidence that the risk is truly at an acceptably low level.

It is important to note that these bounds do not in general apply to risks caused by human activity. Nuclear warfare, anthropogenic climate change, or future technologies in the decades and centuries to come could inflict damage on the biosphere in novel ways. It has also been hypothesised that high energy physics experiments could create novel risks that would destroy the Earth outright, suggesting another possible way in which human activity could prematurely end the biosphere. Ultimately, it has been shown that risks from particle collider experiments should be no higher than risks from naturally occurring cosmic rays, which occur frequently (Hut and Rees, 1983). In this way, obtaining a bound on naturally occurring risk enables a direct bound of a particular form of anthropogenic risk. So although the bounds we have found here cannot in general be extended to risks caused by human activity, there could be specific instances in which a bound on natural risk will provide information on anthropogenic risks. For example, the fact that Earth's climate has remained stable for so long could (perhaps weakly) support the notion that a runaway greenhouse effect is unlikely to be triggered by current human activity. In general, the long track record of survival undermines any hypothesis that suggests that the biosphere is inherently fragile and only persists due to observation selection effects.

Chapter 6

Discussion

Each chapter has contained a discussion section, so here I discuss some general principles and reflections. Using highly simplified models, this thesis has examined a few potential sources of observation selection bias in our estimates of existential risk and the emergence of intelligent life. I have found that in some instances, observation selection effects were unlikely to play a large role (e.g. estimates of existential risk from natural sources), whereas in other instances observation selection effects were highly significant (e.g. the probability that intelligent life emerged on Earth). Are there general principles or techniques that can predict whether observation selection effects will be a major issue for a given phenomena?

Perhaps the first general principle worth discussing is the role that a prior distribution plays in identifying observation selection effects. Candidates for observer selection bias occur when our location or evolutionary history is atypical compared to those without observers. Generating a notion of 'atypical' then requires an estimate or guess of what is 'typical,' and this can be represented by one's prior distribution. For example, claims that the size of our moon is atypical and the result of observer selection effects can rely on a prior created from simulations of the impact that created the moon (Waltham, 2019). In Chapter 3, I provided a few reasons for adopting broad priors over the possible rates of evolutionary transitions, and these priors then produced the conclusion that intelligent life was rare. In the other chapters, the priors have been more

implicit. For example, I concluded in Chapter 2 that observer selection effects did not play a big role. Although I did not place an explicit prior over rates of extinction and observerhood, I used fossil record data and intuitions about human development to support the arguments and cross check the results.

This cross-checking leads us to a second (informal) technique, which is to simply imagine a world in which the parameter in question was set to an extreme value, and determine if there are any testable predictions that would follow from such a scenario. In Chapter 2, I argued that if the natural extinction rate was actually once every 100 years or so, and our survival was just due to luck, then we should still expect to see a much shorter track record of survival and perhaps much more frequent extinction in nearby animal relatives. We may also expect to see an incredibly optimised 'shortest path' in the evolution leading up to intelligent life. Conversely in Chapter 3, we can set the probability of an evolutionary transition to an exceptionally extreme value (e.g. odds of happening in Earth's lifetime of 1 to 10^{100}) and it still seems broadly consistent with the timing and nature of the transitions.

Reflections

I began this research by searching for possible observation selection bias across a wide range of evolutionary and ecological phenomenon. One hypothesis was that the complex ecosystems we see around us were far less stable than one might naively expect (especially after the seminal work by May (1972)), and that perhaps they persist only due to observation selection effects. I examined ecological 'paradoxes,' for example the question of why there are so many species despite a seemingly small number of niches and the theory of competitive exclusion, which states that two competing species will not coexist in the same niche (this is known as the 'paradox of the plankton' from Hutchinson (1961)). These were relatively short-lived projects, as the standard ecological explanations were substantially more compelling than a hypothesis of observer selection bias. The notion that the world around us could collapse at any

moment was further undermined by the insight that observers in such worlds would find a very different evolutionary history than our own, an evolutionary path without much meandering but rather one with a straight line towards intelligent observers.

The models used in this thesis were exceptionally simple. While I believe that simple models are best for clarifying general principles and generating new hypotheses, a number of research avenues could be pursued with more sophisticated models. Climate models, phylogenetic trees, models of fitness landscapes, etc, could potentially enable us to produce priors that we could compare to data, potentially to detect more subtle patterns of observer selection bias. However, my intuition is still that most accessible insights will be in identifying instances in which observation selection effects have a large enough impact to be detectable with simple toy models.

In closing, this thesis has aimed to make progress on a set of questions related to existential risk and the possibility of intelligent life elsewhere in the universe. Beyond having great practical importance to humanity's future, these questions are linked by common methodological challenges associated with observation selection effects. The ultimate conclusions of the thesis are good news. First, existential risks from natural sources are unlikely. Second, our failure to detect flourishing civilisations elsewhere in the universe should not lead us to believe our own civilisation will expire soon, since the Great Filter likely lies in humanity's past, rather than humanity's future.

Bibliography

Abramov, O. and Mojzsis, S. J. (2011), 'Abodes for life in carbonaceous asteroids?', *Icarus* **213**(1), 273–279.

Alroy, J. (1996), 'Constant extinction, constrained diversification, and uncoordinated stasis in North American mammals', *Palaeogeography, Palaeoclimatology, Palaeoecology* **127**(1-4), 285–311.

Amari, S. V. and Misra, R. B. (1997), 'Closed-form expressions for distribution of sum of exponential random variables', *IEEE Transactions on reliability* **46**(4), 519–522.

Anbar, A. D., Duan, Y., Lyons, T. W., Arnold, G. L., Kendall, B., Creaser, R. A., Kaufman, A. J., Gordon, G. W., Scott, C., Garvin, J. et al. (2007), 'A whiff of oxygen before the great oxidation event?', *Science* **317**(5846), 1903–1906.

Anderson, W. J. (2012), *Continuous-time Markov chains: An applications-oriented approach*, Springer Science & Business Media.

Annis, J. (1999), 'An astrophysical explanation for the "great silence".', *Journal of the British Interplanetary Society* **52**, 19–22.

Arbour, J. H. and Santana, S. E. (2017), 'A major shift in diversification rate helps explain macroevolutionary patterns in primate species diversity', *Evolution* **71**(6), 1600–1613.

Awramik, S. M. (1992), 'The oldest records of photosynthesis', *Photosynthesis Research* **33**(2), 75–89.

Bambach, R. K. (2006), 'Phanerozoic biodiversity mass extinctions', *Annu. Rev. Earth Planet. Sci.* **34**, 127–155.

Banks, W. E., d'Errico, F., Peterson, A. T., Kageyama, M., Sima, A. and Sánchez-Goñi, M.-F. (2008), 'Neanderthal extinction by competitive exclusion', *PLoS One* **3**(12), e3972.

Barnosky, A. D., Matzke, N., Tomiya, S., Wogan, G. O., Swartz, B., Quental, T. B., Marshall, C., McGuire, J. L., Lindsey, E. L., Maguire, K. C. et al. (2011), 'Has the Earth's sixth mass extinction already arrived?', *Nature* **471**(7336), 51.

Barrow, J. and Tipler, F. (1986), *The cosmological anthropic principle*, Oxford: Oxford University Press.

Baum, S. D. (2018), 'Uncertain human consequences in asteroid risk analysis and the global catastrophe threshold', *Natural Hazards* **94**(2), 759–775.

Beckstead, N. (2013), *On the overwhelming importance of shaping the far future*, Rutgers The State University of New Jersey-New Brunswick.

Beech, M. (2011), 'The past, present and future supernova threat to Earths biosphere', *Astrophysics and Space Science* **336**(2), 287–302.

Bell, E. A., Boehnke, P., Harrison, T. M. and Mao, W. L. (2015), 'Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon', *Proceedings of the National Academy of Sciences* **112**(47), 14518–14521.

Benson, R. B., Campione, N. E., Carrano, M. T., Mannion, P. D., Sullivan, C., Upchurch, P. and Evans, D. C. (2014), 'Rates of dinosaur body mass evolution indicate 170 million years of sustained ecological innovation on the avian stem lineage', *PLoS Biology* **12**(5), e1001853.

Benton, M. J. (2009), 'The Red Queen and the Court Jester: species diversity and the role of biotic and abiotic factors through time', *Science* **323**(5915), 728–732.

Berwick, R. C. and Chomsky, N. (2016), *Why only us: Language and evolution*, MIT press.

Betts, H. C., Puttick, M. N., Clark, J. W., Williams, T. A., Donoghue, P. C. and Pisani, D. (2018), 'Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin', *Nature ecology & evolution* **2**(10), 1556.

Bolhuis, J. J., Tattersall, I., Chomsky, N. and Berwick, R. C. (2014), 'How could language have evolved?', *PLoS biology* **12**(8), e1001934.

Bomfleur, B., McLoughlin, S. and Vajda, V. (2014), 'Fossilized nuclei and chromosomes reveal 180 million years of genomic stasis in royal ferns', *Science* **343**(6177), 1376–1377.

Bond, D. P. and Wignall, P. B. (2014), 'Large igneous provinces and mass extinctions: an update', *Volcanism, Impacts, and Mass Extinctions: Causes and Effects* **505**, 29–55.

Bostrom, N. (2001), 'The Doomsday Argument, Adam & Eve, UN++, and Quantum Joe', *Synthese* **127**(3), 359–387.

Bostrom, N. (2002*a*), *Anthropic bias: Observation selection effects in science and philosophy*, Routledge.

Bostrom, N. (2002*b*), 'Existential risks: Analyzing human extinction scenarios and related hazards', *Journal of Evolution and Technology* **9**.

Bostrom, N. (2007), 'Observation selection theory and cosmological fine-tuning', *Universe Or Multiverse?* p. 431.

Bostrom, N. (2008), 'Where are they?', *Technology Review* **111**(3).

Bostrom, N. (2013*a*), *Anthropic bias: Observation selection effects in science and philosophy*, Routledge.

Bostrom, N. (2013*b*), 'Existential risk prevention as global priority', *Global Policy* **4**(1), 15–31.

Bostrom, N. (2014), *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press.

Bostrom, N. and Ćirković, M. M. (2003), 'The doomsday argument and the self-indication assumption: reply to Olum', *The Philosophical Quarterly* **53**(210), 83–91.

Bostrom, N. and Cirkovic, M. M. (2011), *Global catastrophic risks*, Oxford University Press.

Bradbury, R. J., Cirkoivc, M. M. and Dvorsky, G. (2011), 'Dysonian approach to seti: a fruitful middle ground?', *Journal of the British Interplanetary Society* **64**(5), 156.

Britten, R. J. (1986), 'Rates of DNA sequence evolution differ between taxonomic groups', *Science* **231**(4744), 1393–1398.

Brocks, J. J., Logan, G. A., Buick, R. and Summons, R. E. (1999), 'Archean molecular fossils and the early rise of eukaryotes', *science* **285**(5430), 1033–1036.

Buick, R. (2008), 'When did oxygenic photosynthesis evolve?', *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **363**(1504), 2731–2743.

Butterfield, N. J. (2000), 'Bangiomorpha pubescens n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the mesoproterozoic/neoproterozoic radiation of eukaryotes', *Paleobiology* **26**(3), 386–404.

Caldeira, K. and Kasting, J. F. (1992), 'The life span of the biosphere revisited', *Nature* **360**(6406), 721.

Canfield, D. E. (2005), 'The early history of atmospheric oxygen: homage to Robert M. Garrels', *Annu. Rev. Earth Planet. Sci.* **33**, 1–36.

Carlson, J. M. and Doyle, J. (2002), 'Complexity and robustness', *Proceedings of the national academy of sciences* **99**(suppl 1), 2538–2545.

Carr, B. J. and Rees, M. J. (1979), 'The anthropic principle and the structure of the physical world', *Nature* **278**(5705), 605.

Carroll, S. B. (2001), 'Chance and necessity: the evolution of morphological complexity and diversity', *Nature* **409**(6823), 1102.

Carter, B. (1974), Large number coincidences and the anthropic principle in cosmology, *in* 'Symposium-international astronomical union', Vol. 63, Cambridge University Press, pp. 291–298.

Carter, B. (1983), 'The anthropic principle and its implications for biological evolution', *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* **310**(1512), 347–363.

Carter, B. (2008), 'Five-or six-step scenario for evolution?', *International Journal of Astrobiology* **7**(2), 177–182.

Catling, D. C., Glein, C. R., Zahnle, K. J. and McKay, C. P. (2005), 'Why o2 is required by complex life on habitable planets and the concept of planetary" oxygenation time"', *Astrobiology* **5**(3), 415–438.

Cavosie, A., Valley, J., Wilde, S. et al. (2005), 'Magmatic δ18o in 4400–3900 Ma detrital zircons: A record of the alteration and recycling of crust in the early Archean', *Earth and Planetary Science Letters* **235**(3-4), 663–681.

Chapman, C. R. and Morrison, D. (1994), 'Impacts on the Earth by asteroids and comets: assessing the hazard', *Nature* **367**(6458), 33.

Chen, J. and Kipping, D. (2018), 'On the rate of abiogenesis from a Bayesian informatics perspective', *arXiv preprint arXiv:1806.08033* .

Ćirković, M. M. (2002), 'On the first anthropic argument in astrobiology', *Earth, Moon, and Planets* **91**(4), 243–254.

Ćirković, M. M. (2008), 'Observation selection effects and global catastrophic risks', *Global catastrophic risks* pp. 120–41.

Ćirković, M. M. (2016), 'Anthropic arguments outside of cosmology and string theory', http://philsci-archive.pitt.edu/12553/1/Anthropic_MMC_v4.pdf.

Ćirković, M. M., Sandberg, A. and Bostrom, N. (2010), 'Anthropic shadow: Observation selection effects and human extinction risks', *Risk Analysis: An International Journal* **30**(10), 1495–1506.

Ćirković, M. M., Vukotić, B. and Dragićević, I. (2009), 'Galactic punctuated equilibrium: how to undermine Carter's anthropic argument in astrobiology', *Astrobiology* **9**(5), 491–501.

Clack, J. A. (2006), 'The emergence of early tetrapods', *Palaeogeography, Palaeoclimatology, Palaeoecology* **232**(2-4), 167–189.

Cockell, C. (2003), *Impossible extinction: natural catastrophes and the supremacy of the microbial world*, Cambridge University Press.

Cockell, C. S. and Bland, P. A. (2005), 'The evolutionary and ecological benefits of asteroid and comet impacts', *Trends in ecology & evolution* **20**(4), 175–179.

Davies, P. C. W. (2003), 'Does life's rapid appearance imply a martian origin?', *Astrobiology* **3**(4), 673–679.

De Vos, J. M., Joppa, L. N., Gittleman, J. L., Stephens, P. R. and Pimm, S. L. (2015), 'Estimating the normal background rate of species extinction', *Conservation Biology* **29**(2), 452–462.

Deamer, D. W. (2018), *Assembling Life: How Can Life Begin on Earth and Other Habitable Planets?*, Oxford University Press.

Dill, K. A. and Chan, H. S. (1997), 'From Levinthal to pathways to funnels', *Nature Structural and Molecular Biology* **4**(1), 10.

Dodd, M. S., Papineau, D., Grenne, T., Slack, J. F., Rittner, M., Pirajno, F., O'Neil, J. and Little, C. T. (2017), 'Evidence for early life in Earth's oldest hydrothermal vent precipitates', *Nature* **543**(7643), 60–64.

Eldredge, N., Thompson, J. N., Brakefield, P. M., Gavrilets, S., Jablonski, D., Jackson, J. B., Lenski, R. E., Lieberman, B. S., McPeek, M. A. and Miller, W. (2005), 'The dynamics of evolutionary stasis', *Paleobiology* **31**(S2), 133–145.

Erwin, D. H., Laflamme, M., Tweedt, S. M., Sperling, E. A., Pisani, D. and Peterson, K. J. (2011), 'The Cambrian conundrum: early divergence and later ecological success in the early history of animals', *science* **334**(6059), 1091–1097.

Evans, A. R., Jones, D., Boyer, A. G., Brown, J. H., Costa, D. P., Ernest, S. M., Fitzgerald, E. M., Fortelius, M., Gittleman, J. L., Hamilton, M. J. et al. (2012), 'The maximum rate of mammal evolution', *Proceedings of the National Academy of Sciences* **109**(11), 4187–4190.

Fike, D., Grotzinger, J., Pratt, L. and Summons, R. (2006), 'Oxidation of the Ediacaran ocean', *nature* **444**(7120), 744.

Flambaum, V. V. (2003), 'Comment on" does the rapid appearance of life on Earth suggest that life is common in the universe?"', *Astrobiology* **3**(2), 237–239.

Foote, M. and Raup, D. M. (1996), 'Fossil preservation and the stratigraphic ranges of taxa', *Paleobiology* **22**(2), 121–140.

Franck, S., Bounama, C. and Von Bloh, W. (2006), 'Causes and timing of future biosphere extinctions', *Biogeosciences* **3**(1), 85–92.

Frank, A. and Sullivan III, W. T. (2016), 'A new empirical constraint on the prevalence of technological species in the universe', *Astrobiology* **16**(5), 359–362.

Fritz, S. A., Bininda-Emonds, O. R. and Purvis, A. (2009), 'Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics', *Ecology letters* **12**(6), 538–549.

Garg, S. G. and Martin, W. F. (2016), 'Mitochondria, the cell cycle, and the origin of sex via a syncytial eukaryote common ancestor', *Genome biology and evolution* **8**(6), 1950–1970.

Goffinet, A. M. (2017), 'The evolution of cortical development: the synapsid-diapsid divergence', *Development* **144**(22), 4061–4077.

Goldblatt, C., Lenton, T. M. and Watson, A. J. (2006), 'Bistability of atmospheric oxygen and the Great Oxidation', *Nature* **443**(7112), 683.

Gott III, J. R. (1993), 'Implications of the Copernican principle for our future prospects', *Nature* **363**, 315–319.

Gould, S. J. (1990), *Wonderful life: the Burgess Shale and the nature of history*, WW Norton & Company.

Gould, S. J. and Eldredge, N. (1993), 'Punctuated equilibrium comes of age', *Nature* **366**(6452), 223.

Gray, R. H. (2015), 'The Fermi paradox is neither Fermi's nor a paradox', *Astrobiology* **15**(3), 195–199.

Grosberg, R. K. and Strathmann, R. R. (2007), 'The evolution of multicellularity: a minor major transition?', *Annu. Rev. Ecol. Evol. Syst.* **38**, 621–654.

Hanson, R. (1998*a*), 'The great filter-are we almost past it?'.

Hanson, R. (1998*b*), 'Must early life be easy? The rhythm of major evolutionary transitions', *Unpublished manuscript, September* **23**, 168–189.

Haqq-Misra, J., Kopparapu, R. K. and Wolf, E. T. (2017), 'Why do we find ourselves around a yellow star instead of a red star?', *International Journal of Astrobiology* pp. 1–10.

Hart, M. H. (1975), 'Explanation for the absence of extraterrestrials on Earth', *Quarterly Journal of the Royal Astronomical Society* **16**, 128.

Henshilwood, C. S., d'Errico, F., Yates, R., Jacobs, Z., Tribolo, C., Duller, G. A., Mercier, N., Sealy, J. C., Valladas, H., Watts, I. et al. (2002), 'Emergence of modern human behavior: Middle Stone Age engravings from South Africa', *Science* **295**(5558), 1278–1280.

Hoffman, P. F., Kaufman, A. J., Halverson, G. P. and Schrag, D. P. (1998), 'A Neoproterozoic snowball Earth', *science* **281**(5381), 1342–1346.

Holland, H. D. (2006), 'The oxygenation of the atmosphere and oceans', *Philosophical Transactions of the Royal Society B: Biological Sciences* **361**(1470), 903–915.

Horvath, J. E. and Galante, D. (2012), 'Effects of high-energy astrophysical events on habitable planets', *International Journal of Astrobiology* **11**(4), 279–286.

Hublin, J.-J., Ben-Ncer, A., Bailey, S. E., Freidline, S. E., Neubauer, S., Skinner, M. M., Bergmann, I., Le Cabec, A., Benazzi, S., Harvati, K. et al. (2017), 'New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens', *Nature* **546**(7657), 289.

Hut, P. and Rees, M. J. (1983), 'How stable is our vacuum?', *Nature* **302**(5908), 508.

Hutchinson, G. E. (1961), 'The paradox of the plankton', *The American Naturalist* **95**(882), 137–145.

Jablonski, D. (1994), 'Extinctions in the fossil record', *Phil. Trans. R. Soc. Lond. B* **344**(1307), 11–17.

Jaffe, R., Busza, W., Wilczek, F. and Sandweiss, J. (2000), 'Review of speculative disaster scenarios at RHIC', *Reviews of Modern Physics* **72**(4), 1125.

Kasting, J. F. (1987), 'Theoretical constraints on oxygen and carbon dioxide concentrations in the precambrian atmosphere', *Precambrian research* **34**(3-4), 205–229.

Kasting, J. F. (1988), 'Runaway and moist greenhouse atmospheres and the evolution of Earth and Venus', *Icarus* **74**(3), 472–494.

Kennedy, R. F. (2011), *Thirteen days: A memoir of the Cuban missile crisis*, WW Norton & Company.

Kinnison, M. T. and Hendry, A. P. (2001), The pace of modern life ii: from rates of contemporary microevolution to pattern and process, *in* 'Microevolution rate, pattern, process', Springer, pp. 145–164.

Kirchner, J. W. (2002), 'The Gaia hypothesis: fact, theory, and wishful thinking', *Climatic Change* **52**(4), 391–408.

Knoll, A. H. (2011), 'The multiple origins of complex multicellularity', *Annual Review of Earth and Planetary Sciences* **39**, 217–239.

Knoll, A. H. (2014), 'Paleobiological perspectives on early eukaryotic evolution', *Cold Spring Harbor Perspectives in Biology* **6**(1), a016121.

Koonin, E. V. (2011), *The logic of chance: the nature and origin of biological evolution*, FT press.

Kumar, S. and Subramanian, S. (2002), 'Mutation rates in mammalian genomes', *Proceedings of the National Academy of Sciences* **99**(2), 803–808.

Lane, N. (2011), 'Energetics and genetics across the prokaryote-eukaryote divide', *Biology direct* **6**(1), 35.

Langer, M. C., Ramezani, J. and Da Rosa, Á. A. (2018), 'U-Pb age constraints on dinosaur rise from south Brazil', *Gondwana Research* **57**, 133–140.

Laskar, J. and Gastineau, M. (2009), 'Existence of collisional trajectories of Mercury, Mars and Venus with the Earth', *Nature* **459**(7248), 817.

Lenton, T. M., Boyle, R. A., Poulton, S. W., Shields-Zhou, G. A. and Butterfield, N. J. (2014), 'Co-evolution of eukaryotes and ocean oxygenation in the neoproterozoic era', *Nature Geoscience* **7**(4), 257.

Lenton, T. M., Dahl, T. W., Daines, S. J., Mills, B. J., Ozaki, K., Saltzman, M. R. and Porada, P. (2016), 'Earliest land plants created modern levels of atmospheric oxygen', *Proceedings of the National Academy of Sciences* **113**(35), 9704–9709.

Lenton, T. M. and von Bloh, W. (2001), 'Biotic feedback extends the life span of the biosphere', *Geophysical research letters* **28**(9), 1715–1718.

Lenton, T., Watson, A. and Watson, A. J. (2011), *Revolutions that made the Earth*, Oxford University Press.

Leslie, J. (1990), 'Risking the world's end', *Interchange* **21**(1), 49–58.

Leslie, J. (2002), *The End of the World: the science and ethics of human extinction*, Routledge.

Levin, S. R., Scott, T. W., Cooper, H. S. and West, S. A. (2017), 'Darwin's aliens', *International Journal of Astrobiology* pp. 1–9.

Lineweaver, C. H. (2009), Paleontological tests: human-like intelligence is not a convergent feature of evolution, *in* 'From fossils to astrobiology', Springer, pp. 353–368.

Lineweaver, C. H. and Davis, T. M. (2002), 'Does the rapid appearance of life on Earth suggest that life is common in the universe?', *Astrobiology* **2**(3), 293–304.

Lineweaver, C. H., Fenner, Y. and Gibson, B. K. (2004), 'The galactic habitable zone and the age distribution of complex life in the Milky Way', *Science* **303**(5654), 59–62.

Lingam, M. and Loeb, A. (2018), 'Subsurface exolife', *International Journal of Astrobiology* pp. 1–30.

Livio, M. (1999), 'How rare are extraterrestrial civilizations, and when did they emerge?', *The Astrophysical Journal* **511**(1), 429.

Livio, M. (2018), 'How special is the solar system?', *arXiv preprint arXiv:1801.05061* .

Livio, M. and Kopelman, A. (1990), 'Life and the Sun's lifetime', *Nature* **343**, 25.

Long, J. A. and Gordon, M. S. (2004), 'The greatest step in vertebrate history: a paleobiological review of the fish-tetrapod transition', *Physiological and Biochemical Zoology* **77**(5), 700–719.

Lowe, D. R. and Byerly, G. R. (2015), 'Geologic record of partial ocean evaporation triggered by giant asteroid impacts, 3.29–3.23 billion years ago', *Geology* **43**(6), 535–538.

Lsor, A. D. A. and Shaviv, N. J. (1997), 'Life extinctions by neutron star mergers', *Phénomènes À Très Hautes Énergies Dans L'univers* **32**, 379.

Lyons, T. W., Reinhard, C. T. and Planavsky, N. J. (2014), 'The rise of oxygen in Earth's early ocean and atmosphere', *Nature* **506**(7488), 307.

Malmberg, D., Davies, M. B. and Heggie, D. C. (2011), 'The effects of fly-bys on planetary systems', *Monthly Notices of the Royal Astronomical Society* **411**(2), 859–877.

Manheim, D. (2018), 'Questioning estimates of natural pandemic risk', *Health security* .

Marean, C. W., Bar-Matthews, M., Bernatchez, J., Fisher, E., Goldberg, P., Herries, A. I., Jacobs, Z., Jerardino, A., Karkanas, P., Minichillo, T. et al. (2007), 'Early human use of marine resources and pigment in South Africa during the Middle Pleistocene', *Nature* **449**(7164), 905.

Mason, B. G., Pyle, D. M. and Oppenheimer, C. (2004), 'The size and frequency of the largest explosive eruptions on Earth', *Bulletin of Volcanology* **66**(8), 735–748.

Matheny, J. G. (2007), 'Reducing the risk of human extinction', *Risk Analysis: An International Journal* **27**(5), 1335–1344.

Mattila, T. M. and Bokma, F. (2008), 'Extant mammal body masses suggest punctuated equilibrium', *Proceedings of the Royal Society B: Biological Sciences* **275**(1648), 2195–2199.

May, R. M. (1972), 'Will a large complex system be stable?', *Nature* **238**(5364), 413.

Mayr, E. (1994), 'Does it pay to acquire high intelligence?', *Perspectives in biology and medicine* **37**(3), 337–338.

McBrearty, S. and Brooks, A. S. (2000), 'The revolution that wasn't: a new interpretation of the origin of modern human behavior', *Journal of human evolution* **39**(5), 453–563.

McCabe, M. and Lucas, H. (2010), 'On the origin and evolution of life in the galaxy', *International Journal of Astrobiology* **9**(4), 217–226.

McDougall, I., Brown, F. H. and Fleagle, J. G. (2005), 'Stratigraphic placement and age of modern humans from Kibish, Ethiopia', *Nature* **433**(7027), 733.

McKay, C. P. (1996), Time for intelligence on other planets, *in* 'Circumstellar habitable zones', p. 405.

Melott, A. L., Lieberman, B. S., Laird, C. M., Martin, L. D., Medvedev, M. V., Thomas, B. C., Cannizzo, J., Gehrels, N. and Jackman, C. (2004), 'Did a gamma-ray burst initiate the late Ordovician mass extinction?', *International Journal of Astrobiology* **3**(1), 55–61.

Melott, A. L. and Thomas, B. C. (2011), 'Astrophysical ionizing radiation and Earth: a brief review and census of intermittent intense sources', *Astrobiology* **11**(4), 343–361.

Mendez, F. L., Poznik, G. D., Castellano, S. and Bustamante, C. D. (2016), 'The divergence of Neandertal and modern human Y chromosomes', *The American Journal of Human Genetics* **98**(4), 728–734.

Meyer, M., Arsuaga, J.-L., de Filippo, C., Nagel, S., Aximu-Petri, A., Nickel, B., Martínez, I., Gracia, A., de Castro, J. M. B., Carbonell, E. et al. (2016), 'Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins', *Nature* **531**(7595), 504.

Millett, P. and Snyder-Beattie, A. (2017), 'Existential risk and cost-effective biosecurity', *Health security* **15**(4), 373–383.

Mojzsis, S. J., Arrhenius, G., McKeegan, K., Harrison, T., Nutman, A. and Friend, C. (1996), 'Evidence for life on Earth before 3,800 million years ago', *Nature* **384**(6604), 55.

Morris, J. L., Puttick, M. N., Clark, J. W., Edwards, D., Kenrick, P., Pressel, S., Wellman, C. H., Yang, Z., Schneider, H. and Donoghue, P. C. (2018), 'The timescale of early land plant evolution', *Proceedings of the National Academy of Sciences* **115**(10), E2274–E2283.

Napier, W. (2008), *Hazards from comets and asteroids*, Oxford University Press: Oxford, UK.

Okie, J. G., Boyer, A. G., Brown, J. H., Costa, D. P., Ernest, S. M., Evans, A. R., Fortelius, M., Gittleman, J. L., Hamilton, M. J., Harding, L. E. et al. (2013), 'Effects of allometry, productivity and lifestyle on rates and limits of body size evolution', *Proceedings of the Royal Society of London B: Biological Sciences* **280**(1764), 20131007.

Olson, S. J. (2016), 'On the visible size and geometry of aggressively expanding civilizations at cosmological distances', *Journal of Cosmology and Astroparticle Physics* **2016**(04), 021.

Olson, S. J. (2017), 'Estimates for the number of visible galaxy-spanning civilizations and the cosmological expansion of life', *International Journal of Astrobiology* **16**(2), 176–184.

Olum, K. D. (2002), 'The doomsday argument and the number of possible observers', *The Philosophical Quarterly* **52**(207), 164–184.

Ord, T., Hillerbrand, R. and Sandberg, A. (2010), 'Probing the improbable: methodological challenges for risks with low probabilities and high stakes', *Journal of Risk Research* **13**(2), 191–205.

Pääbo, S. (2014), *Neanderthal man: In search of lost genomes*, Hachette UK.

Parfit, D. (1984), *Reasons and persons*, Oxford University Press.

Parfrey, L. W., Lahr, D. J., Knoll, A. H. and Katz, L. A. (2011), 'Estimating the timing of early eukaryotic diversification with multigene molecular clocks', *Proceedings of the National Academy of Sciences* **108**(33), 13624–13629.

Pearce, B. K., Tupper, A. S., Pudritz, R. E. and Higgs, P. G. (2018), 'Constraining the time interval for the origin of life on Earth', *Astrobiology* **18**(3), 343–364.

Peck, W. H., Valley, J. W., Wilde, S. A. and Graham, C. M. (2001), 'Oxygen isotope ratios and rare earth elements in 3.3 to 4.4 Ga zircons: Ion microprobe evidence for high δ18o continental crust and oceans in the early Archean', *Geochimica et Cosmochimica Acta* **65**(22), 4215–4229.

Pinker, S. and Bloom, P. (1990), 'Natural language and natural selection', *Behavioral and brain sciences* **13**(4), 707–727.

Piran, T. and Jimenez, R. (2014), 'Possible role of gamma ray bursts on life extinction in the universe', *Physical review letters* **113**(23), 231102.

Pisaturo, R. (2009), 'Past longevity as evidence for the future', *Philosophy of Science* **76**(1), 73–100.

Powell, A., Shennan, S. and Thomas, M. G. (2009), 'Late Pleistocene demography and the appearance of modern human behavior', *Science* **324**(5932), 1298–1301.

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O'connor, T. D., Santpere, G. et al. (2013), 'Great ape genetic diversity and population history', *Nature* **499**(7459), 471.

Purvis, A., Gittleman, J. L., Cowlishaw, G. and Mace, G. M. (2000), 'Predicting extinction risk in declining species', *Proceedings of the Royal Society of London B: Biological Sciences* **267**(1456), 1947–1952.

Queller, D. C. (2000), 'Relatedness and the fraternal major transitions', *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **355**(1403), 1647–1655.

Rabosky, D. L. (2010), 'Extinction rates should not be estimated from molecular phylogenies', *Evolution* **64**(6), 1816–1824.

Radzvilavicius, A. L. and Blackstone, N. W. (2015), 'Conflict and cooperation in eukaryogenesis: implications for the timing of endosymbiosis and the evolution of sex', *Journal of the Royal Society Interface* **12**(111), 20150584.

Rampino, M. R. and Stothers, R. B. (1988), 'Flood basalt volcanism during the past 250 million years', *Science* **241**(4866), 663–668.

Rasmussen, B., Fletcher, I. R., Brocks, J. J. and Kilburn, M. R. (2008), 'Reassessing the first appearance of eukaryotes and cyanobacteria', *Nature* **455**(7216), 1101.

Raup, D. M. (1986), 'Biological extinction in Earth history', *Science* **231**(4745), 1528–1533.

Raup, D. M. (1992), *Extinction: bad genes or bad luck?*, WW Norton & Company.

Raup, D. M. and Sepkoski, J. J. (1982), 'Mass extinctions in the marine fossil record', *Science* **215**(4539), 1501–1503.

Richter, D., Grün, R., Joannes-Boyau, R., Steele, T. E., Amani, F., Rué, M., Fernandes, P., Raynal, J.-P., Geraads, D., Ben-Ncer, A. et al. (2017), 'The age of the hominin fossils from Jebel Irhoud, Morocco, and the origins of the Middle Stone Age', *Nature* **546**(7657), 293–296.

Robinson, C., Campbell, T. L., Cote, S. and de Ruiter, D. J. (2018), 'Temporal ranges and ancestry in the hominin fossil record: The case of Australopithecus sediba', *South African Journal of Science* **114**(3-4), 1–7.

Rosenzweig, M. L. and McCord, R. D. (1991), 'Incumbent replacement: evidence for long-term evolutionary progress', *Paleobiology* **17**(3), 202–213.

Rougier, J., Sparks, R. S. J., Cashman, K. V. and Brown, S. K. (2018), 'The global magnitude–frequency relationship for large explosive volcanic eruptions', *Earth and Planetary Science Letters* **482**, 621–629.

Ruff, C. B., Trinkaus, E. and Holliday, T. W. (1997), 'Body mass and encephalization in Pleistocene Homo', *Nature* **387**(6629), 173.

Rushby, A. J., Claire, M. W., Osborn, H. and Watson, A. J. (2013), 'Habitable zone lifetimes of exoplanets around main sequence stars', *Astrobiology* **13**(9), 833–849.

Russell, D. A., Seguin, R. et al. (1982), 'Reconstructions of the small cretaceous theropod, stenonychosaurus inequalis, and a hypothetical dinosauroid'.

Sandberg, A. (2018), Human extinction from natural hazard events, *in* 'Oxford Research Encyclopedia of Natural Hazard Science', Oxford Research Encyclopedias.

Sandberg, A., Drexler, E. and Ord, T. (2018), 'Dissolving the Fermi paradox', *arXiv preprint arXiv:1806.02404* .

Schluter, D. (2000), *The ecology of adaptive radiation*, Oxford University Press.

Schulte, P., Alegret, L., Arenillas, I., Arz, J. A., Barton, P. J., Bown, P. R., Bralower, T. J., Christeson, G. L., Claeys, P., Cockell, C. S. et al. (2010), 'The Chicxulub asteroid impact and mass extinction at the Cretaceous-Paleogene boundary', *science* **327**(5970), 1214–1218.

Segura, T. L., Zahnle, K., Toon, O. B. and McKay, C. P. (2013), The effects of impacts on the climates of terrestrial planets, *in* S. J. Mackwell, A. A. Simon-Miller, J. W. Harder and M. A. Bullock, eds, 'Comparative Climatology of Terrestrial Planets', University of Arizona Press Tucson, pp. 417–438.

Sepkoski, J. (1986), Phanerozoic overview of mass extinction, *in* 'Patterns and Processes in the History of Life', Springer, pp. 277–295.

Shampine, L. F. and Reichelt, M. W. (1997), 'The Matlab ODE suite', *SIAM journal on scientific computing* **18**(1), 1–22.

Sherwood, S. C. and Huber, M. (2010), 'An adaptability limit to climate change due to heat stress', *Proceedings of the National Academy of Sciences* **107**(21), 9552–9555.

Simpson, F. (2017), 'The longevity of habitable planets and the development of intelligent life', *International Journal of Astrobiology* **16**(3), 266–270.

Sleep, N. H. and Zahnle, K. (1998), 'Refugia from asteroid impacts on early mars and the early Earth', *Journal of Geophysical Research: Planets* **103**(E12), 28529–28544.

Sloan, D., Batista, R. A. and Loeb, A. (2017), 'The resilience of life to astrophysical events', *Scientific reports* **7**(1), 5419.

Smith, E. I., Jacobs, Z., Johnsen, R., Ren, M., Fisher, E. C., Oestmo, S., Wilkins, J., Harris, J. A., Karkanas, P., Fitch, S. et al. (2018), 'Humans thrived in South Africa through the Toba eruption about 74,000 years ago', *Nature* **555**(7697), 511.

Smith, J. M. (1989), 'The causes of extinction', *Phil. Trans. R. Soc. Lond. B* **325**(1228), 241–252.

Smith, J. M. and Szathmary, E. (1997), *The major transitions in evolution*, Oxford University Press.

Spiegel, D. S. and Turner, E. L. (2012), 'Bayesian analysis of the astrobiological implications of life's early emergence on Earth', *Proceedings of the National Academy of Sciences* **109**(2), 395–400.

Stern, R. J. (2016), 'Is plate tectonics needed to evolve technological species on exoplanets?', *Geoscience Frontiers* **7**(4), 573–580.

Szathmáry, E. (2015), 'Toward major evolutionary transitions theory 2.0', *Proceedings of the National Academy of Sciences* **112**(33), 10104–10111.

Szathmáry, E. and Maynard Smith, J. (1995), 'The major evolutionary transitions', *Nature* **374**(6519), 227–232.

Tattersall, I. (2008), 'An evolutionary framework for the acquisition of symbolic cognition by Homo sapiens', *Comparative cognition & behavior reviews* **3**.

Tegmark, M. and Bostrom, N. (2005*a*), 'Astrophysics: Is a doomsday catastrophe likely?', *Nature* **438**(7069), 754.

Tegmark, M. and Bostrom, N. (2005*b*), 'How unlikely is a doomsday catastrophe?', *arXiv preprint astro-ph/0512204* .

Tipler, F. J. (1980), 'Extraterrestrial intelligent beings do not exist', *Quarterly Journal of the Royal Astronomical Society* **21**, 267–281.

Tomitani, A., Knoll, A. H., Cavanaugh, C. M. and Ohno, T. (2006), 'The evolutionary diversification of cyanobacteria: molecular–phylogenetic and paleontological perspectives', *Proceedings of the National Academy of Sciences* **103**(14), 5442–5447.

Toon, O. B., Zahnle, K., Morrison, D., Turco, R. P. and Covey, C. (1997), 'Environmental perturbations caused by the impacts of asteroids and comets', *Reviews of Geophysics* **35**(1), 41–78.

Turner, M. S. and Wilczek, F. (1982), 'Is our vacuum metastable?', *Nature* **298**(5875), 633–634.

Van Valen, L. (1973), 'A new evolutionary law', *Evol Theory* **1**, 1–30.

Vermeij, G. J. (2016), 'Gigantism and its implications for the history of life', *PLoS One* **11**(1), e0146092.

Vukotic, B. and Cirkovic, M. (2008), 'Neocatastrophism and the Milky Way astrobiological landscape', *Serbian Astronomical Journal* **176**, 71–79.

Waltham, D. (2007), 'Half a billion years of good weather: Gaia or good luck?', *Astronomy & Geophysics* **48**(3), 3–22.

Waltham, D. (2011), 'Anthropic selection and the habitability of planets orbiting M and K dwarfs', *Icarus* **215**(2), 518–521.

Waltham, D. (2014), *Lucky Planet: Why Earth is Exceptional and What That Means for Life in the Universe*, Basic Books a Member of Perseus Books Group.

Waltham, D. (2017), 'Star masses and star-planet distances for Earth-like habitability', *Astrobiology* **17**(1), 61–77.

Waltham, D. (2019), 'Is Earth special?', *Earth-Science Reviews* .

Ward, P. (2009), *The Medea hypothesis: Is life on Earth ultimately self-destructive?*, Princeton University Press.

Ward, P. and Brownlee, D. (1999), 'Rare earth', *Copernicus, Springer-Verlag, New York, NY* .

Watson, A. J. (2004), 'Gaia and observer self-selection', *Scientists debate Gaia: The next century* pp. 201–210.

Watson, A. J. (2008), 'Implications of an anthropic model of evolution for emergence of complex life and intelligence', *Astrobiology* **8**(1), 175–185.

Weinberg, S. (1987), 'Anthropic bound on the cosmological constant', *Physical Review Letters* **59**(22), 2607.

Weitzman, M. L. (2009), 'On modeling and interpreting the economics of catastrophic climate change', *The Review of Economics and Statistics* **91**(1), 1–19.

West, S. A., Fisher, R. M., Gardner, A. and Kiers, E. T. (2015), 'Major evolutionary transitions in individuality', *Proceedings of the National Academy of Sciences* **112**(33), 10112–10119.

Wignall, P. B. (2001), 'Large igneous provinces and mass extinctions', *Earth-science reviews* **53**(1), 1–33.

Wilson, P. A. (1994), 'Carter on anthropic principle predictions', *The British journal for the philosophy of science* **45**(1), 241–253.

Wood, B. and K. Boyle, E. (2016), 'Hominin taxic diversity: Fact or fantasy?', *American journal of physical anthropology* **159**, 37–78.

Zackrisson, E., Calissendorff, P., González, J., Benson, A., Johansen, A. and Janson, M. (2016), 'Terrestrial planets across space and time', *The Astrophysical Journal* **833**(2), 214.

Zendejas, J., Segura, A. and Raga, A. (2010), 'Atmospheric mass loss by stellar wind from planets around main sequence M stars', *Icarus* **210**(2), 539–544.

Zuidhof, M., Schneider, B., Carney, V., Korver, D. and Robinson, F. (2014), 'Growth, efficiency, and yield of commercial broilers from 1957, 1978, and 2005', *Poultry Science* **93**(12), 2970–2982.