Ascertaining the Uncertainty in Astrobiology: Leveraging mathematical and computational models to aid with rational decision making in astrobiology

CATHERINE LUCY GILLEN

Ascertaining the Uncertainty in Astrobiology: Leveraging mathematical and computational models to aid with rational decision making in astrobiology

Catherine Gillen

Uncertainty is a defining feature of astrobiology, shaping our ability to apply mathematical tools and computational models for decision making. This thesis examines how uncertainty affects these methods and explores ways to navigate it within a mathematical and computational framework. While uncertainty over key probabilities — such as abiogenesis or technological advancement — limits tools like maximising expected utility and Bayesian statistics, it does not render decision making impossible. Instead, this thesis highlights both the challenges and potential strategies for addressing uncertainty in astrobiology.

Five primary conclusions emerge:

C1: I propose and defend a new definition of biosignature: *any phenomenon for which biological processes are a known possible explanation and whose potential abiotic causes have been reasonably explored and ruled out* (Gillen et al., 2023, p.1228). This is strong enough to be meaningful but leaves room for uncertainty over the list of possible explanations captured by the problem of unconceived alternatives (Stanford 2001, 2006a). This conclusion is discussed in Chapter Two and Chapter Three of this thesis.

C2: I propose a new corresponding definition of potential biosignature: *any phenomenon for which biological processes are a known possible explanation but whose potential abiotic causes have not yet been reasonably explored and ruled out* (Gillen et al., 2023, p.1238). This is argued for in Chapter Two and Chapter Three.

C3: In light of the vast and unexplored potential research areas constituting the field of astrobiology, theoretical arguments exist for funding high-uncertainty, high-payoff research. These arguments do not apply to high-risk, high-payoff research. This conclusion is discussed in Chapter Four and Chapter Five of this thesis.

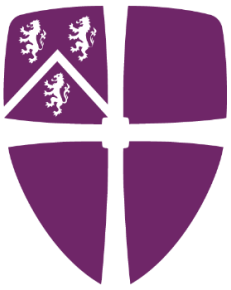
C4: High uncertainty surrounds key fundamental probabilities in astrobiology, such as the probability of abiogenesis. This uncertainty means that astrobiologists disagreeing about fundamental probabilities might be beneficial. This conclusion is found in Chapter Six.

C5: The deployment of computer models of scientific communities, such as astrobiology, requires a close correspondence between the functional form of working real-world features and the functional form of working model features. This is consistent with a structural realist view of computer modelling. This conclusion is discussed in Chapter Seven.

Despite the challenges posed by uncertainty, it need not paralyse scientific progress. Recognising true uncertainty (as opposed to calculatable risk) can guide research toward fascinating and obscured discoveries. Moreover, computational models, while valuable, cannot produce certain conclusions from uncertain inputs. Yet, when designed with structural realism in mind and hence a strong correspondence to real-world systems, they remain useful tools for modelling epistemic communities in astrobiology.

Ascertaining the Uncertainty in Astrobiology: Leveraging mathematical and computational models to aid with rational decision making in astrobiology

Catherine Gillen



PhD Thesis

Department of Philosophy

Durham University

March 2025

Contents

Acknowledgements

I am deeply grateful to my supervisor, Peter Vickers, for his guidance, insight, and kindness throughout this PhD. I have long admired his research, and it has been a delight to work with him on this exciting project. To my fellow EURiCA group members, thank you for the thoughtful discussions, sharp critiques, and occasional existential and cosmic tangents — this work is stronger because of you. And to my fiancé Duncan — my constant amidst the uncertainty. You have always met me with patience and unwavering support; your love is the one variable I never have to question.

Publications and Their Relationship to this Thesis

Published Peer-Reviewed Journal Articles

Gillen, C. (**lead author)**, Jeancolas, C., McMahon, S., & Vickers, P. (2023). The call for a new definition of biosignature. *Astrobiology*, *23*(11), 1228-1237.

Jeancolas, C., Gillen, C., McMahon, S., Ward, M., & Vickers, P. J. (2024a). Breakthrough results in astrobiology: is 'high risk' research needed?. *International Journal of Astrobiology*, *23*, e1.

Jeancolas, C., Gillen, C., McMahon, S., & Vickers, P. (2024b). Is astrobiology serious science?. *Nature Astronomy*, *8*(1), 5-7.

Vickers, P., Cowie, C., Dick, S. J., Gillen, C., Jeancolas, C., Rothschild, L. J., & McMahon, S. (2023). Confidence of life detection: The problem of unconceived alternatives. *Astrobiology*, *23*(11), 1202-1212.

Vickers, P., Gardiner, E., Gillen, C., Hyde, B., Jeancolas, C., Finnigan, S., Novakova, J., Strandin, H., Tasdan, U., Taylor, H., McMahon, S. (2025). Surveys of the scientific community on the existence of extraterrestrial life. *Nature Astronomy*.

This thesis has been informed and moulded by discussions within the EURiCA research group (Exploring Uncertainty and Risk in Contemporary Astrobiology). Several publications have resulted from the research group, and I will outline here how these publications broadly relate to the context covered in this thesis. Specific and more detailed references to these publications will be made throughout the thesis.

Chapter one of this thesis is a slightly modified version of a paper, for which I am the lead author, published in *Astrobiology* and titled "The Call for a New Definition of Biosignature" (Gillen et al., 2023). This paper includes several of the key themes addressed throughout the thesis. These include how the problem of unconceived alternatives (Stanford, 2001, 2006a) is born out of uncertainty over the list of

possible outcomes and how this undermines our ability to make definitive, objective statements about many detections in astrobiology. The problem of unconceived alternatives renders specific strong definitions of biosignature unusable. As such, a new and epistemically modest definition of biosignature is proposed in this published work and Chapter Two of this thesis.

The problem of unconceived alternatives and its impact on astrobiology is further discussed in a paper by Vickers et al., "Confidence of life detection: the problem of unconceived alternatives" (2023). In this paper, we expose how the problem of unconceived alternatives introduces uncertainty in many astrobiology claims. Given the need for honest communication of such uncertainty, we endorse using the uncertainty assessment approach advocated by the Intergovernmental Panel on Climate Change (IPCC), within the field of astrobiology. There is some overlap between the content of Chapter Two and Chapter Three of this thesis and the content covered by Vickers et al. (2023). This is especially regarding why the problem of unconceived alternatives is so troubling in astrobiology and how the IPCC confidence framework can work in tandem with the proposed definition of biosignature.

Chapter Four of this thesis has been partly informed and motivated by Jeancolas et al. (2024a). This paper is titled "Breakthrough results in astrobiology: is 'high risk' research needed?". In this, we investigate the relationship between the impact of astrobiology research projects and their degree of risk. We argued that the most impactful papers in astrobiology disproportionately come from projects considered high-risk at the time. More specifically, projects with highly exploratory approaches disproportionately result in high-impact results.

In Chapter Four, I argue that high-uncertainty (defined as an unknown probability of success) projects should be encouraged rather than explicitly high-risk (a known low probability of success) projects. This is consistent with the findings reported in Jeancolas et al. (2024a), whereby projects with highly exploratory approaches would better fit under the definition of high uncertainty given in this chapter, rather than high risk. Moreover, I argue that the distinction between the two is a useful one within the field of astrobiology and should be adopted by funding bodies.

In the paper "Is astrobiology serious science?" (Jeancolas et al., 2024b), we address and rebut three arguments that astrobiology is not a serious scientific discipline. These arguments are 1) astrobiology does not have a proper subject of study, 2) astrobiology is more science fiction than science, and 3) astrobiology is just a buzzword for researchers to gain funding. The discussions addressing the first argument have been recreated in several places in this thesis. Jeancolas et al. (2024b) acknowledge that, indeed, if the goal of astrobiology is to detect life, then we have not succeeded. However, any individual astrobiology research proposal is rarely designed to have the detection of life as its sole goal. Rather,

astrobiology research proposals often comprise compound objectives with varying degrees of risk and uncertainty. This is recognised in several areas of the present thesis, but especially in Chapter Five, which discusses the motivations behind SETI. Unlike most astrobiology missions, SETI research can fall into this more single-goal approach and hence, the attack that SETI has not yet discovered its subject of study is more persuasive than the same attack on astrobiology generally.

Finally, the paper "Surveys of the scientific community on the existence of extraterrestrial life" (Vickers et al., 2025) has motivated and informed Chapter Six of this thesis. In Vickers et al. (2025), we report and analyse the responses of four surveys of astrobiology researchers. The surveys asked researchers how likely they think it is for basic, complex, and intelligent life to exist somewhere in the universe. There is strong agreement within the astrobiology community over the likely existence of basic life. This agreement drops off for complex life and even more for intelligent life. Interestingly, analysis of the data suggests that the community ascribes a more significant gap between basic and complex life than between complex and intelligent life.

Our findings, reported in Vickers et al. (2025), highlight the differing prior probabilities that researchers within the field hold for fundamental events, such as the probability of intelligent extraterrestrial life. This work has quantified a central assumption in Chapter Six of this thesis. This Chapter analyses the disagreement over the origin of a mysterious object that hurtled through our solar system in 2017. I argue that the disagreement over its origin is likely due to disagreeing priors for the probability of detecting extraterrestrial intelligence. Our paper (Vickers et al., 2023) provides data to support the statement that disagreement over the existence of extraterrestrial intelligence is prevalent in the astrobiology community.

Chapter One

Introduction: Astrobiology, an Uncertain Field

This chapter motivates the need for an in-depth analysis of uncertainty in astrobiology and the limitations that this uncertainty imposes on the use of mathematical and computational models. It aids in situating the current thesis within existing literature.

Astrobiology is a field at the frontier of science. Breakthrough results are frequent, and the field is continually shifting. Two key reasons for the high output of this dynamic field are its relative immaturity and its vast research area. I begin by first making the case for astrobiology as a young and vast field, before turning to how these features lead to astrobiology being an uncertain field.

1.1. A Young Field

Astrobiology is a relatively immature field, and this contributes to its high output and fast-evolving nature. The initiation of a dedicated research program within NASA for what was then termed *exobiology* was only established in 1960 (Hubbard, 2022). As a point of comparison, the beginnings of modern thermodynamics can be drawn at several places: e.g., with the publication of Sadi Carnot's *Reflections on the Motive Power of Fire* in 1824 (Carnot, 1824); with Rudolf Clausius' introduction of the concept of entropy in 1850 (Saslow, 2020); or with Thomson's phrasing of the Second Law in 1853 (Thomson, 1853), as notable examples. Taking any of these events as the starting point puts modern thermodynamics between 170 and 200 years old; astrobiology has existed as a field for about 65 years. Even the modern quantum theory outdates astrobiology, with its origins in Heisenberg's 1925 matrix mechanics (Edwards, 1981).

The result of astrobiology's immaturity as a field is simply that it has much yet to discover, and debate over how to interpret results is to be expected. The development of the new quantum model can be taken as an example of this. The old quantum theory existed in the transition between classical Newtonian mechanics and modern, fully relativistic quantum mechanics. This transitionary theory was initiated in 1900 with Max Planck's quantisation of energy to explain black-body radiation (Planck, 1900). Following this, a hybrid theory of classical notions (such as electron trajectories within Bohr's (1913) and

Sommerfeld's (1916) atomic models) and quantum ideas (such as quantised energy levels within both Bohr's and Sommerfeld's models) resulted. Ideas were changing and evolving quickly until the old model was scrapped entirely in 1925 in place of the new, fully relativistic quantum model.

Even since the adoption of the new quantum model, huge discoveries have been made, and debates have ensued. Schrödinger unified Heisenberg's matrix mechanics with his wave function in 1926 (Schrödinger, 1926); Heisenberg first formulated his uncertainty principle in 1927 (Heisenberg, 1927); quantum chromodynamics was developed in the 1960s (Gell-Mann, 2018); the groundwork of quantum computing can be attributed to Yuri Manin's book *Computable and uncomputable* (1980); and the debate over preferred interpretations of quantum mechanics remains rife (e.g., as discussed in Jammer, 1974). These are just a small selection of the continual discoveries and debates within quantum mechanics. Nevertheless, the selection shows that quantum mechanics is very much a fertile field, with constant new discoveries and applications, even 100 years later. `

Astrobiology is a field even younger than quantum mechanics. It is unsurprising, then, that breakthrough and field-changing results continue to fill astrobiology journals. Compiling a brief list of these breakthrough results would reduce a rich field to an arbitrarily short highlights reel – see do Nascimento-Dias and Martinez-Frais (2023) for an entire paper dedicated to summarising the history of astrobiology results. Having said this, any summary of breakthrough results in astrobiology would undoubtedly include the detection of the first exoplanet around a sun-like star (Mayor & Queloz, 1995), the continual discovery of the environmental tolerance of extremophiles (e.g., Rothschild & Mancinelli, 2001), and the flurry of since-rescinded biosignature claims (e.g., McKay et al., 1996; Lepland et al., 2005; Nutman et al., 2016, 2019, 2021; Greaves et al., 2021).

The youth of astrobiology as a field suggests that there is still much to discover. We should expect breakthrough results to continue to come in at pace. And, of significance to this thesis, we should expect some of these results to upend aspects of our current understanding. The detection of the first exoplanet around a sun-like star, 51 Pegasi b, is a good example of a discovery radically shifting our beliefs. Mayor and Queloz's (1995) detection used the radial velocity method, which is only effective on high-mass planets in small-radius orbits (so as to noticeably wobble the host star). Such planets were considered rare on account of the existing theory of planet formation. However, the detection of 51 Pegasi b led to a reworking of planet formation theory that has large planets forming far from their star and migrating in over millennia (Lin et al., 1996).

So, we should expect astrobiology to have much still to show us. We have only recently been able to send probes to other celestial bodies, and the distances at which we can see are only increasing — especially given the recent launch of the James Webb Space Telescope (JWST) with its ability to see into mid-

infrared (Gardner et al., 2006). This is compounded by considerations of the vast size of astrobiology's remit, both in terms of physical space and content. It is to this second contributing factor to the high output and changing nature of the field that we now turn.

1.2. A Broad Field

Astrobiology is the study of the origin, evolution, distribution, and future of life in the universe (NASEM, 2019). Such a research remit is vast — galactic, even; when deciding where to direct, for example, the JWST, researchers have a 360-degree span and a 13.7 billion light-year range (Yan et al., 2022). The result of this is that there is still much to discover. The proportion of potentially habitable planets in the observable universe that we have even glanced at will be vanishingly small. With astrobiology, there is much still to discover.

Moreover, astrobiology is somewhat of an umbrella field. It encompasses a range of sciences, including, but not limited to, biology, geology, chemistry, physics, and astronomy. The most cited astrobiology journal, *Astrobiology,* lists on its homepage 17 subject areas that it widely publishes on. These span from geomicrobiology to space exploration technology. As such, not only is astrobiology's physical research area vast, but so are the number of sub-disciplines within the field. It is the broadness of astrobiology, alongside its relative newness as a field, that should leave us unsurprised by the constantly evolving and high-output nature of astrobiology. What, though, might this have to do with uncertainty?

1.3. An Uncertain Field

So far, I have presented astrobiology as an especially young and broad scientific field. The primary goal of this thesis is to evaluate the impact that uncertainty has on the utilisation of mathematical and computational tools within astrobiology. To establish the need for such a thesis, it is necessary first to make the case that uncertainty is indeed a prevalent feature of astrobiology (in a way above and beyond other scientific disciplines), and I will argue that this is due to the young and broad nature of the field. Throughout the thesis, I will delve into more detail regarding specific aspects and consequences of uncertainty in astrobiology. However, for the purposes of this introduction, an overview of three main categories of uncertainty within astrobiology will suffice.

Uncertainty, as defined in rational decision making, economics, and related disciplines, refers to making a choice while working with incomplete information. This incomplete information can be with regard to 1) what the set of potential outcomes is, 2) the probabilities of each outcome resulting, or 3) the potential

payoffs of each outcome (Binmore, 2008, Chapter 3). This is in contrast to decision making under risk, whereby the set of potential outcomes, the probabilities, and the payoffs are all known. The newness and vastness of astrobiology contribute to high uncertainty in all three of these forms.

1.3.1. Uncertainty in the Set of Possible Outcomes

The first type of uncertainty associated with decision making is uncertainty over the list of potential outcomes. This means you do not have access to the complete set of possible choices. For example, imagine you are ordering at a restaurant, but the ink has run on the menu such that half of the options are unreadable. Due to a lack of information regarding what options are available, you will not be in a position to maximise your choice. This would be an example of uncertainty over the set of possible options.

Astrobiology is especially burdened with this form of uncertainty, and the field's youth and breadth fuel this. As discussed above, we should expect our current understanding to be far from complete, and as such, we should expect our possible explanations for any particular phenomenon to be incomplete.

For a historical example, consider the discovery of the Martian meteorite Allen Hills 84001 (ALH84001) in Antarctica in 1984. This particular case study will be explored in more depth in Chapter Two of this thesis, but for now, a summary is sufficient. ALH8001 had some unusual properties; it appeared to possess fossil structures. The researchers at the time considered their list of possible abiotic explanations alongside the conclusion that life was the cause. Considering three key features of the fossil-like structures, they concluded that "although there are alternative explanations for each of these phenomena taken individually, when they are considered collectively, particularly in view of their spatial association, we conclude that they are evidence for primitive life on Mars" (McKay et al., 1996, p.929).

However, McKay and his team were not working with the complete set of possible explanations for the structures within ALH84001 – they were making a choice under uncertainty over the list of possible outcomes. This became apparent when new, previously unconceived alternative explanations for the structures arose over the following decades (see, for example, Golden et al., 2001, 2004, 2006 and Bell, 2007). Most recently, Steele et al. (2022) showed that mineral carbonation and serpentinization reactions on early Mars explain the occurrence of organic matter in the meteorite.

The example of ALH84001 shows how researchers in astrobiology often make choices between candidate explanations without having access to the entire list of possible explanations. They must instead choose the best explanation from an incomplete list.

This decision making under uncertainty of possible outcomes has high relevance to Kyle Stanford's problem of unconceived alternatives (2001, 2006a, 2006b), which will be discussed at greater length in Chapter Two of this thesis. In short, the problem of unconceived alternatives challenges how a scientific community can be confident in their best scientific inferences or theories when the history of science has continually shown such inferences and theories to be false. Scientists choose the best theory from their list of possible theories, but they are working with an incomplete list. Take, for example, the luminiferous ether of the 19th century. This false theory of light as the propagation through a physical medium fooled even Maxwell (Maxwell, 1861). However, the now-accepted theory of quantum electrodynamics was simply not conceived of in the 19th century, so scientists could not factor it into their theory choice.

It is true that the problem of unconceived alternatives is a problem shared across all disciplines of science, not just astrobiology; we are often working with incomplete information. One lesson from Stanford's problem of unconceived alternatives is that we simply do not know how exhaustively we have explored the list of possible explanations. We might have studied a particular field extensively for a long time, and our list of known possible explanations for any phenomena is, say, 95% as long as the actual list. Maybe there is a good chance that the real explanation is on our list of known possible explanations. Or we might only have access to a tiny fraction of the list of possible explanations, so our best guess will not likely be the correct one. As such, this has been framed as a problem for eliminative inference (Ruhmkorff, 2011, p.876).

However, this does not leave us powerless to make predictions about which of these scenarios we are in. If the scientific discipline in question is especially large, we should expect the relevant knowledge set to be similarly large. Moreover, if we have only been exploring a scientific discipline for a short time, we should expect only to have discovered a small part of that knowledge set (Vickers et al., 2023). I have argued that both of these antecedents are true in the case of astrobiology. Meadows et al. (2022, p.26) capture this sentiment well in the case of astrobiology: "If the scope of possible abiotic explanations is known to be poorly explored, it suggests we cannot adequately reject abiotic mechanisms". As such, we should expect the problem of unconceived alternatives to be especially prevalent in the young, broad field of astrobiology, and hence, uncertainty over the set of possible outcomes for any choice to be significant.

1.3.2. Uncertainty in the Probabilities of Outcomes

The second category of uncertainty concerns a lack of information regarding the probabilities of each outcome. A paradigmatic example of such uncertainty is the choice of whether to bring an umbrella out

for the day, given that you have not heard the weather forecast. You know the list of possible options: 1) you bring an umbrella and it rains, 2) you bring an umbrella and it does not rain, 3) you do not bring an umbrella and it rains, 4) you do not bring an umbrella and it does not rain. Moreover, you can determine the payoffs of each outcome by quantifying the utilities of carrying/not carrying an umbrella and getting/not getting wet. However, if you do not know the probability of it raining, you cannot calculate the expected utilities of carrying an umbrella and not carrying an umbrella.

The notion of maximising expected utility is the centrepiece of orthodox decision theory, the formalism of which was derived by Von Neumann and Morgenstern in the appendix of their revolutionary *Theory of Games and Economic Behaviour* (1947). Briefly (though this will be explored in depth in the thesis), orthodox decision theory states that the most rational decision is the one that maximises expected utility, where expected utility for any particular action is given by the product of the probability of the corresponding event occurring and the payoff of the event. Hence, to select the action that results in the event with the highest expected utility, the list of possible events, as well as their probabilities and payoffs, must be known.

Let us work through an example to help clarify how this type of uncertainty impacts the use of expected utility. To return to the umbrella example, we would calculate the expected utility of carrying an umbrella by doing the following:

$$\textit{Expected Utility(bringing umbrella)} = \textit{Utility(having to carry the umbrella)} + \textit{Utility(raining with an umbrella)Probability(raining)} + \textit{Utility(not raining with an umbrella)Probability(not raining)}$$

Now say that you did watch the weather forecast, and you heard that there is a 40% chance of rain. Inputting sensible values for the utilities (though this would be a subjective question for the individual to answer), we would get the following expected utility for bringing an umbrella:

$$\textit{Expected Utility(bringing umbrella)} = (-5) + (-5)(0.4) + (15)(0.6) = 2$$

Therefore, the expected utility of bringing an umbrella, with the above utilities and a 40% chance of rain, is 2. Similarly, the expected utility for not bringing an umbrella can be calculated as so:

$$\textit{Expected Utility(not bringing umbrella)} = \textit{Utility(not carrying the umbrella)} + \textit{Utility(raining without an umbrella)Probability(raining)} + \textit{Utility(not raining without an umbrella)Probability(not raining)}$$

Again, inputting sensible utilities into this toy calculation gives:

$$\textit{Expected Utility(not bringing umbrella)} = (0) + (-30)(0.4) + (15)(0.6) = -3$$

Comparing the expected utilities of bringing an umbrella and not bringing an umbrella shows the former to be the better choice. However, such a calculation simply cannot be done if the probability of rain is unknown — if we had not inputted our 40% chance of rain, then the expected utilities would be undefined.

We should expect this type of uncertainty in the probability of outcomes to be prevalent in astrobiology on account of our relatively small exploration of all there is to know within the field. In a recent paper surveying astrobiologists, we quantify the degree of agreement regarding the probability of extraterrestrial life (Vickers et al., 2025). Respondents were asked about the probability of simple life, complex life, and intelligent life existing elsewhere in the universe. They positioned themselves on a five-point Likert scale ranging from strongly agree to strongly disagree. The results showed strong agreement among the astrobiology community regarding the existence of simple life, with an agreement score of 86.6% (Vickers et al., 2025, p. 16). However, this agreement score dropped to 58.2% when considering the existence of intelligent life (ibid., p.16). The limitations of using a five-point Likert scale when determining agreement (especially concerning how to interpret "neutral" votes) are discussed (ibid.).

Disagreement over prior probabilities for fundamental events like the existence of intelligent life elsewhere in the universe will be addressed in depth in this thesis (especially in Chapters Five and Six). However, for present purposes, the disagreement reported by Vickers et al. (2025) is a motivating example of how difficult it can be to define key probabilities within astrobiology. If we cannot input a probability for detecting intelligent life, we cannot calculate the expected utility of, for example, funding a SETI mission over a mission that searches for simple life. Astrobiology is, therefore, also subject to this second category of uncertainty: uncertainty over the probabilities of outcomes.

1.3.4. Uncertainty in the Payoff of Outcomes

The final category of uncertainty is uncertainty in the payoffs of the possible outcomes (Binmore, 2008, Chapter 3). To get a handle on this type of uncertainty, let us use again the example of deciding whether to bring an umbrella out for the day. This time, let us say that you did watch the weather forecast and, indeed, there is a 40% chance of rain. However, let us imagine that you have never experienced rain! You do not know how unpleasant you will find the experience of being rained on, with or without an umbrella.

In such a scenario of limited information, you will again find yourself unable to calculate the expected utility of either carrying an umbrella or not. This is because of a lack of utilities to substitute into the calculation.

Such uncertainty over the payoffs of outcomes should be expected to be a prevalent form of uncertainty in astrobiology. As with the two other types of uncertainty discussed in §1.3.2. and §1.3.3., the present thesis will delve into a deeper analysis of why this is the case. However, for introductory purposes, consider the following motivating example.

In a paper titled "Breakthrough results in astrobiology: is 'high risk' research needed?" (Jeancolas et al., 2024a), we report the findings of interviews carried out with the authors of some of the most impactful papers in astrobiology over the past 20 years. The impact here is taken to track with citation count. The authors of ten high-impact projects were interviewed, and it was found that eight of these ten projects were regarded as exploratory in nature. The relationship between a project being high-impact and being highly exploratory has interesting consequences that will be especially addressed in Chapter Four. However, for present purposes, the prevalence of such highly exploratory research in astrobiology hints towards the prevalence of high uncertainty over the payoff of outcomes.

Highly exploratory research is defined as research for which the specific goal of what will be found is unknown (Jeancolas et al., 2023, p.8). An example of this, discussed in Jeancolas et al. (2024a, pp.8-9), is the result of the first mass spectrometry analysis performed in situ on the surface of a comet (Goesmann et al., 2015). The novelty of this project invited a significant degree of uncertainty over what would be found; such an analysis had not yet been performed. Another such example of highly exploratory research, cited in Jeancolas et al. (2024a, p.9), comes from analysing ice grains from Saturn's rings (Postberg et al., 2009). The researchers were unsure of what they might find. Indeed, the author interviewed remarked that "I was investigating the composition of these ice grains from Enceladus (…) just because it was fascinating to look at the composition of material that originates from the subsurface of an ice moon" (cited in Jeaoncolas et al., 2024a, p.9).

The researchers had a serendipitous breakthrough when they detected an unexpected peak in the spectroscopy data. Such a peak corresponded to molecules that necessarily originated from a liquid water ocean (Postberg et al., 2009; Jeancolas et al., 2024a). In this way, the researchers did not anticipate the payoffs of this highly exploratory project prior to their decision to pursue this particular project over any other. One of the authors commented that he thought it fascinating to look at the subsurface material of an icy moon, but he did not anticipate just how impactful it would end up being.

It is unsurprising that highly exploratory research, where high uncertainty surrounds the potential payoffs, is prevalent in astrobiology. The newness and vastness of astrobiology leave us with much we do not know. Serendipitous discoveries are common – the haphazard discovery of abiotic mimics on Mars is another example (McMahon & Cosmidis, 2022, p.29).

The decisions made by researchers as to which projects they should pursue, and the decisions made by funding bodies as to which of these projects should receive funding, are obscured by uncertainty over the potential payoffs of outcomes. In this way, this third and final form of uncertainty is a significant feature of astrobiology.

1.4. Existing Literature on the Philosophy of Astrobiology

At present, I hope to have sufficiently convinced the reader of the existence of the three types of uncertainty within astrobiology. The field's newness and vastness lend well to high uncertainty over the set of possible outcomes, their probabilities, and their payoffs.

The primary goal of this thesis is to explore the limitations imposed by high uncertainty on the utilisation of mathematical and computational models in astrobiology. Prior to this though, it is helpful to establish the existing related literature within which to situate the current work.

1.4.1. Notable Literature on the Philosophy of Astrobiology

As stated, this thesis considers the impact of high uncertainty on the utilisation of mathematical and computational models in astrobiology. However, this should be taken to fall under the category of the philosophy of astrobiology on account of the methods and arguments employed throughout the thesis. I will appeal to seminal arguments within the philosophy of science literature, such as Stanford's problem of unconceived alternatives (Stanford, 2001; 2006a; 2006b), the scientific realism debate (see, for example, Putnam 1975; Laudan, 1985; Psillos, 1999, 1996; Poincare, [1905] 1952; Worrall, 1994), and applications of formal epistemology, e.g., in community modelling (Weisburg & Muldoon, 2009; Avin, 2019) and orthodox decision theory (Koch, 1990; Von Neumann & Morgenstern, 1947).

To better establish where this thesis fits into the existing literature, it will be helpful to provide an overview of current work on the philosophy of astrobiology more generally before exploring literature in the narrower context of uncertainty in astrobiology.

Influential writers in the philosophy of astrobiology include historians of science such as Steven Dick (2020, 2015, 2013, 2005, 2001, 1984) and David Dunér (2013, with Holmberg & Persson); philosophers of science, including Christophe Malaterre (2024, 2024 (with Lareau), 2022 (with Jeancolas & Nghe)) and Carol Cleland (2019); and an ever-increasing number of astrobiologists publishing on themes in the

philosophy of astrobiology, e.g., Charley Lineweaver (2007, 2016 (with Chopra), 2022a, 2022b) and Chris McKay (2020).

For as long as there has been astrobiology, there have been individuals studying the practice of astrobiology itself, and hence, there has been the philosophy of astrobiology. Having said this, the dedication of books (e.g., Dick, 2020, 2015; Dunér et al., 2013; Vakoch, 2013; Ćirković, 2012), conferences, research clusters (e.g., the EURiCA project at Durham University and Edinburgh University, and the Leverhulme Centre for Life in the Universe at Cambridge University), and even a term for the specific philosophical study of astrobiology, "astrophilosophy" has recently boomed.

1.4.2. Literature on astrophilosophy

"Astrophilosophy" has been used by Kristina Šekrst (2024) to capture the broader philosophical framework that she argues is required to study the origins and distribution of life in the universe. In her paper "Astrobiology in philosophy or philosophy in astrobiology?" (2024), Šekrst argues that astrobiology research inherently raises philosophical questions about our place in the universe and discusses how anthropocentric approaches can lead us astray. A two-way interplay between philosophy and astrobiology exists, and hence a need for a term that captures both the philosophy within astrobiology and astrobiology within philosophy is defended: astrophilosophy being just this term.

Šekrst was not the first to propose this term. Von Hegner wrote on the role of astrophilosophy in his paper "Astrobiology and astrophilosophy: subsuming or bifurcating disciplines?" (2019). In this, he stresses the intertwined yet separate roles that astrobiology and astrophilosophy play in the search for life in the universe. Each discipline comes with different methodologies, and von Hegner summarises that "Astrobiology is needed to address life beyond this planet, but this life, and its interaction with *Homo sapiens*, will lead to many questions better addressed by astrophilosophy" (von Hegner, 2019, p.77).

So, establishing the philosophy of astrobiology as its own research area has attracted discussion. But, more than this, the content of this research is rich. The first of these I will raise in this literature review is the philosophical study of life itself.

1.4.3. Literature on the Definition and Ontic Status of Life

Common themes in the philosophy of astrobiology include questions about the definition, and even ontic status, of life itself. Popular definitions of life include those based on thermodynamics (Schrödinger,

1944; Schneider & Kay, 1994; Lineweaver, 2006; Macklem & Seely, 2010), evolution (Joyce, 1994), and metabolism (Sagan, 1970).

Each definition has its proponents and practical advantages over the next. For example, NASA endorses an evolutionary definition of life as "a self-sustaining chemical system capable of undergoing Darwinian evolution" (Joyce, 1994). This definition is succinct and intuitive. However, it does not lend itself well to the search for life beyond Earth. Even on Earth, the discovery of dinosaur fossils would not, in isolation, meet the criteria of an evolutionary definition of life — we cannot see evolutionary history with a single specimen. Individuals do not evolve; populations evolve. Instead, the millions of years of dinosaur fossil records build the case for life, as defined evolutionarily.

In contrast, the referents of metabolic and thermodynamic definitions of life can be readily searched for in astrobiology missions. Sagan (2010) defines life metabolically as "an object with a definite boundary, continually exchanging some of its materials with its surroundings, but without altering its general properties at least over some period of time" (p.303). A prevalent example of the search for metabolism was the infamous Labeled Release experiment on the Viking lander mission. Nutrients were added to Martian soil, and the equipment looked for signs of metabolism (Navarro-González et al., 2006; Bianciardi et al., 2012). Similarly, the search for chemical disequilibria in planetary atmospheres (that is, a combination of volatiles that will react with each other to become more stable gases) is a common method in the search for extraterrestrial life (e.g., Young et al., 2024; Wogan & Catling, 2020; Krissansen-Totton et al., 2016).

However, metabolic and thermodynamic definitions of life come with unintuitive consequences. It has been argued that mineral crystals with high internal structure, such as diamonds and quartz, meet thermodynamic definitions of life (Cleland, 2019, p.36). As is the case with hurricanes, which maintain the highly ordered motion of air molecules in part by the increase of environmental entropy as heat is transported from the warm ocean surface to the atmosphere (Emanuel, 1991; see especially pp. 179-186). Regarding metabolic definitions of life, unintuitive admissions include flames (Cleland, 2019, p.37); unintuitive omissions include seeds and spores (Cleland, 2019, p.37).

Despite much attention, the community has not reached a consensus regarding a single, unified definition of life. This prompted Edouard Machery's (2012) seminal paper, "Why is stopped worrying about the definition of life… and why you should as well". In this, Machery defends the position that life is a folk concept and ultimately concludes that this renders life impossible to define.

A brief summary of Machery's argument is as follows. Folk concepts have tangible meaning within peoples' minds but lack objective conditions to describe them. Examples of folk concepts include good,

justice, and knowledge (Machery, 2012, p.153). Hold any of these concepts in your mind and try to define it – you have a sense of what it means to be good, but the term evades definitive pinning down. Moreover, the attempts of the person next to you to pin down what it is to be good may look different to yours, even though you are both broadly referring to the same concept. The consequence of this, Machery argues, is that defining folk concepts is pointless. When evaluating whether something is good, or whether justice has been served, most people do not refer to a fixed definition of each. By holding the view that life is a folk concept, Machery denies that life has a single and universal list of criteria to define it.

A final significant debate I will raise here surrounding the ontology of life concerns its vague or precise nature. Malaterre (2024) writes that "mounting evidence indicates a gray area between what is intuitively clearly alive and what is distinctly not alive. This prompts consideration of a gradualist perspective, depicting life as a spectrum with varying degrees of 'lifeness'" (Malaterre, 2024, p.1). A primary motivation behind this position is the existence of what Malaterre and others have called proto-living systems (Mann, 2012; Fox, 1991). These borderline cases indicate that, not only is evolution a gradual process, but the transition from chemical systems to biological ones is gradual too.

1.4.4. Literature on the Ethics of Astrobiology

Beyond philosophical discussions over the nature of life, a great deal of literature on the philosophy of astrobiology is concerned with the ethics of astrobiology. Questions include: if we find extraterrestrial life, do we have moral obligations to it? (Milligan, 2021; Smith, 2014; Persson, 2012); how can we terraform a planet ethically? (Stoner, 2021; Schwartz, 2013, 2019; Haqq-Misra, 2012; Sparrow, 1999); and is it possible to balance the commercial use of space with planetary protection? (Persson, 2017; Persson et al., 2018). Tangential to the ethics of astrobiology is the politics of astrobiology. Questions including who has ownership over celestial bodies such as the moon (Butler, 2017; Pop, 2008) or who has the right to space travel (Campion, 2016; Langston, 2016) also abound in the literature.

The present thesis will touch on the themes discussed thus far in the philosophy of astrobiology to varying degrees. More narrowly, though, the focus of this thesis is on statistical uncertainty within astrobiology. As such, a brief review of existing literature on Bayesian statistics and astrobiology will further aid in situating this work.

1.4.5. Bayesian Statistics and Astrobiology

Bayesian statistics offers an exciting and potentially illuminating mechanism by which to update our probabilities rationally. The centerpiece of Bayesian statistics is Bayes' theorem. This is widely given as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

where *A* and *B* are events. *P(A)* and *P(B)* are the probabilities of *A* and *B* occurring respectively, and $P(A|B)$ is a conditional probability representing the probability of *A given* that *B* has already occurred. Similarly, $P(B|A)$ is the probability of *B given* that *A* has already occurred (Bayes, 1764). The probabilities on the right-hand side represent the priors; the conditional probability on the left-hand side is our posterior and is the probability to be determined.

The benefit of utilising Bayes' theorem to update your beliefs, given new evidence, is that you end up with posterior probabilities that do not violate Kolmogorov's axioms of probability. What this means is that, by ensuring you update in compliance with Bayes' theorem, you will not violate the axioms of non-negativity (the probability of any event must not be strictly less than zero), normalisation (the probability of the events in the sample space must sum to one), and additivity (If *A* and *B* are mutually exclusive events such that $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$) (Kolmogorov & Bharucha-Reid, 2018).

Bayes' theorem has broad applicability, from predicting animal behaviour (McNamara et al., 2006) to aiding in cardiovascular nursing (Thompson & Martin, 2017). Moreover, its use in astrobiology has been increasing, especially within the context of biosignature research (see, for example, Bains & Petkowski, 2021; Catling et al., 2018; Walker et al., 2018; Spiegel & Turner, 2012). I will now provide a brief overview of these notable pieces of literature that utilise Bayesian statistics in astrobiology.

In a paper titled "Astrobiologists are rational but not Bayesian" (2021), Bains and Petkowski analyse the results of two surveys of the astrobiology community regarding their prior and posterior beliefs about the probability of life on bodies within our solar system. In the time between the two surveys, a divisive paper claiming that phosphine had been detected in the cloud decks of Venus (Greaves et al., 2021) was published. Bains and Petkowski (2021) wanted to ascertain whether the surveyed astrobiology community could be modelled as updating their prior beliefs about life on Venus, given the 2020 paper, in a Bayesian way. They concluded that no Bayesian model fitted the survey data and hence the astrobiology community is not Bayesian (Bains & Petkowski, 2021).

The conclusions of Bains and Potkowski's paper might appear discouraging for those advocating the use of Bayesian statistics in astrobiology. One particularly prominent paper endorsing Bayesian statistics in

astrobiology is Catling et al.'s "Exoplanet Biosignatures: A Framework for Their Assessment" (2018). In this, the authors propose a system by which to categorise confidence in any particular biosignature claim. The system they advocate centers on carrying out a Bayesian update on prior beliefs. A posterior probability in the 90-100% range would then correspond to "very likely", whereas a posterior probability of <10% corresponds to "very unlikely". This work is complemented by Walker et al.'s paper "Exoplanet Biosignatures: Future Directions" (2018). In this, the authors argue that Bayesian statistics can be used to guide future search strategies, as well as to constrain the probability of a detection of life.

Beyond applications to biosignature research, some have argued for using Bayesian statistics to determine the probability of abiogenesis. Spiegel and Turner's "Bayesian Analysis of the astrobiological implications of life's early emergence on Earth" (2012) explores the use of Bayes' theorem and the early emergence of life on Earth to constrain the probability of life emerging from non-life. They conclude that, although life's early emergence on Earth provides weak evidence for abiogenesis not being exceedingly rare, our inability to constrain our priors undermines the enterprise. Spiegel and Turner's paper does well to emphasise how highly sensitive Bayesian posteriors are on the priors. This roadblock that Spiegel and Turner (2012) hit up against is known as the "problem of the priors" in Bayesian statistics (Titelbaum, 2022), and it is something that I will discuss at length in this thesis (see, especially Chapter Five, §5.2.2 and Chapter Six).

The papers from Bains and Petkowski (2021), Catling et al. (2018), Walker et al. (2018), and Spiegel and Turner (2012) are a small selection of work using Bayesian statistics in astrobiology. However, the present thesis will go on to disagree with how many of these authors have employed Bayes' theorem in this context. It is at this point that I will carve out a space for my thesis within this existing literature.

1.5. A Gap in the Literature for this Thesis

In this thesis, I will argue that existing research into the application of Bayesian statistics in astrobiology has stumbled because of a key limitation in the use of the theorem. The limitation is that Bayesian statistics is built to handle decision making under risk, not under uncertainty. As I have outlined in §1.3 of this chapter, decision making under uncertainty pertains to a lack of knowledge regarding either the set of possible outcomes, the probabilities, or the payoffs involved (Binmore, 2008, Chapter 3). This is in contrast to decision making under risk, whereby the set of possible outcomes, the probabilities, and the payoffs are all known (Binmore, 2008, Chapter 3).

To clarify this point, let us take again Bayes' theorem, but substituting in the events L = there is life on Venus, and P = there is phosphine on Venus:

$$P(L|P) = \frac{P(P|L)P(L)}{P(P)}.$$

$P(L|P)$ is the probability of there being life on Venus, given that there is phosphine on Venus. $P(P|L)$ is the probability of there being phosphine on Venus, given that there is life on Venus. $P(L)$ and $P(P)$ are the prior probabilities of there being life on Venus and there being phosphine on Venus respectively.

Note that this is a hypothetical toy example in light of the widely undermined results of the Greaves et al.'s (2021) paper claiming to have detected phosphine in the cloud decks of Venus. See, for example, rebuttals to Greaves et al. (2021): Villanueva et al. (2021), Trompet et al. (2021), Akins et al. (2021), Lincowski et al. (2021), Encrenaz et al. (2020), and Snellen et al. (2020).

Nonetheless, in such an illustrative example of using Bayes' theorem to update our prior probability of there being life on Venus, given that we have found phosphine on Venus, we need defined probabilities for $P(P|L)$, $P(L)$, and $P(P)$. In other words, we need to know our priors in order to calculate our posteriors. If, instead, we were working under uncertainty, we would not know what value to input for, as per the example, the probability of detecting phosphine on Venus in the first place. In short, Bayes' theorem aids with decision making under risk, not uncertainty.

Thus far in this introduction, I have laid the groundwork for why we should view astrobiology as a highly uncertain field. The result of this is that the utilising of central mathematical tools in rational decision making, such as maximising expected utility and Bayes' theorem, is undermined. There is a need, therefore, for a detailed analysis of when we can and cannot leverage mathematical and computational models to aid in rational decision making in astrobiology. This thesis aims to respond to this need.

Indeed, when considering the decision making process within astrobiology, the stakes are significant. The costs associated with the field are vast — NASA's 2022 budget alone was $24 billion (The Planetary Society, n.d.). Yet, demand for funding in astrobiology continues to far outstrip supply (Bitten et al., 2019). Competition for money, laboratory time, or telescope time is fierce, and many fruitful projects will never gain funding. Mechanisms by which the best research projects are selected are therefore paramount to maximising the outputs of astrobiology. However, despite how useful a mathematical or computational model would be in aiding decision making in astrobiology, many of those models are undermined by uncertainty within the field.

With these motivations in mind, this thesis will arrive at five primary conclusions:

> **C1:** I propose and defend a new definition of biosignature: *any phenomenon for which biological processes are a known possible explanation and whose potential abiotic causes have been reasonably explored and ruled out* (Gillen et al., 2023, p.1228). This is strong enough to be meaningful but leaves room for uncertainty

over the list of possible explanations captured by the problem of unconceived alternatives (Stanford 2001, 2006a). This conclusion is discussed in Chapter Two and Chapter Three of this thesis.

C2: I propose a new corresponding definition of potential biosignature: *any phenomenon for which biological processes are a known possible explanation but whose potential abiotic causes have not yet been reasonably explored and ruled out* (Gillen et al., 2023, p.1238). This is argued for in Chapter Two and Chapter Three.

C3: In light of the vast and unexplored potential research areas constituting the field of astrobiology, theoretical arguments exist for funding high-uncertainty, high-payoff research. These arguments do not apply to high-risk, high-payoff research. This conclusion is discussed in Chapter Four and Chapter Five of this thesis.

C4: High uncertainty surrounds key fundamental probabilities in astrobiology, such as the probability of abiogenesis. This uncertainty means that astrobiologists disagreeing about fundamental probabilities might be beneficial. This conclusion is found in Chapter Six.

C5: The deployment of computer models of scientific communities, such as astrobiology, requires a close correspondence between the functional form of working real-world features and the functional form of working model features. This is consistent with a structural realist view of computer modelling. This conclusion is discussed in Chapter Seven.

The structure of the thesis is as follows. Firstly, before delving into the limits that uncertainty imposes on mathematical and computational modelling in astrobiology, there is a central term within astrobiology that is especially affected by uncertainty. I address this first in Chapters Two and Three.

Chapter Two, titled "The Call for a New Definition of Biosignature" explores the damaging effect that the problem of unconceived alternatives has on a working definition of biosignature. Existing definitions of biosignature are critiqued and found to be either so strong in their requirements as to be unusable, or so weak as to lack meaning. A new, pragmatic definition of biosignature is proposed that circumvents these attacks and accounts for the problem of unconceived alternatives. This novel definition of biosignature is: *any phenomenon for which biological processes are a known possible explanation and whose potential abiotic causes have been reasonably explored and ruled out* (Gillen et al., 2023, p.1228). A complementary definition of potential biosignature is also proposed in Chapter Two: *any phenomenon for which biological processes are a known possible explanation but whose potential abiotic causes have not yet been reasonably explored and ruled out* (Gillen et al., 2023, p.1238).

Chapter Three, titled "A Defence of the Proposed Definition of Biosignature: Epistemic Vs Ontic Definitions", defends the proposed definition of biosignature against recent attacks in the literature. At the heart of the primary attack on the proposed definition of biosignature is the desire that it be ontic. On the contrary, my proposed definition of biosignature is an epistemic one in that the conditions must be *known* to be met for something to qualify as a biosignature. In this way, and unlike ontic definitions, a biosignature does not exist in the absence of our knowledge of it. This unusual position is defended.

The remainder of this thesis then turns to the ill effects of uncertainty on the quantitative modelling of decision making in astrobiology. Resultingly, Chapter Four, titled "Risk vs Uncertainty in Astrobiology: More than Semantic and Not Pedantic", makes the case for a clear distinction between these terms in the lexicon of astrobiology. I argue that such a distinction is not merely of semantic importance but is needed to ensure rational selection between candidate research proposals. The recent call for high-*risk* research is misleading. Clarifying that high-*uncertainty* research is what is needed might provide an important step in boosting the output of astrobiology.

Having defended the funding of high-uncertainty, high-payoff research, Chapter Five, "The Search for a Justification of SETI", evaluates whether the search for extraterrestrial intelligence (SETI) might fall within this category. If it indeed falls within this category, the conclusions of Chapter Four might be used to advocate for the renewed public funding of SETI. I evaluate this by delving into the seven parameters of the famous Drake equation to quantify the uncertainty associated with the search for extraterrestrial intelligence. The high degree of uncertainty associated with at least four of the seven parameters seals SETI as high-uncertainty research. Having said this, I argue in this chapter that SETI research is not unproblematically high-payoff, and it is this that invalidates it from being classified as high-uncertainty, high-payoff.

Remaining within the realm of the extraterrestrial, Chapter Six delves into a recent and controversial case study of disagreement in astrobiology. Titled "Revisiting the Oumuamua Debate: The 'Problem' of the Priors and its Value for Astrobiology", Chapter Six discusses the strange case of Oumuamua. Since its sighting in 2017, impassioned debates have ensued concerning just what this object is, with some claiming it is simply a planetesimal or comet (Bergner & Seligman, 2023; The 'Oumuamua ISSI Team, 2019), whilst others insist it is an alien light sail (Loeb, 2021, 2018; Bialy & Loeb, 2018). Working within a Bayesian framework, I argue in this chapter that the crux of this disagreement is not either side's understanding of the evidence but rather differing priors. Moreover, I argue that these differing priors may be an advantage to the highly speculative and uncertain field of astrobiology.

Following this, Chapter Seven, titled "The Promise and Pitfalls of Community Modelling in Astrobiology", takes the suggestion that disagreement is beneficial to astrobiology and puts it to the test

via an epistemic community model. I present interpreted results of a novel model to show how easy and tempting it is to draw real-world conclusions from such models. I then further argue how the over-interpretation of the computer modelling of scientific communities is widespread. This is done by delving into three well-cited community models (Avin, 2019; Hong & Page, 2004; Weisburg & Muldoon, 2009). It is found that the conclusions of each of these models are both baked into the models and are overinterpreted. They hence do not map onto their real-world target systems. I present an argument that appeals to structural realism over computer modelling and conclude that computer models in astrobiology (and other sciences) can be informative, so long as the features responsible for deriving the conclusions are functionally equivalent to the working features in their target system.

Concluding remarks of the thesis will then follow in Chapter Eight, the final chapter of this thesis. I will summarise the findings of each chapter and reiterate the five key conclusions of the thesis. To give these again:

C1: I propose and defend a new definition of biosignature: *any phenomenon for which biological processes are a known possible explanation and whose potential abiotic causes have been reasonably explored and ruled out* (Gillen et al., 2023, p.1228). This is strong enough to be meaningful but leaves room for uncertainty over the list of possible explanations captured by the problem of unconceived alternatives (Stanford 2001, 2006a). This conclusion is discussed in Chapter Two and Chapter Three of this thesis.

C2: I propose a new corresponding definition of potential biosignature: *any phenomenon for which biological processes are a known possible explanation but whose potential abiotic causes have not yet been reasonably explored and ruled out* (Gillen et al., 2023, p.1238). This is argued for in Chapter Two.

C3: In light of the vast and unexplored potential research areas constituting the field of astrobiology, theoretical arguments exist for funding high-uncertainty, high-payoff research. These arguments do not apply to high-risk, high-payoff research. This conclusion is discussed in Chapter Four and Chapter Five of this thesis.

C4: High uncertainty surrounds key fundamental probabilities in astrobiology, such as the probability of abiogenesis. This uncertainty means that astrobiologists disagreeing about fundamental probabilities might be beneficial. This conclusion is found in Chapter Six.

C5: The deployment of computer models of scientific communities, such as astrobiology, requires a close correspondence between the functional form of working real-world features and the functional form of working model features. This is consistent with a structural realist view of computer modelling. This conclusion is discussed in Chapter Seven.

The set of conclusions within this thesis might be captured with an overarching sentiment:

Uncertainty is indeed a pervasive feature of astrobiology, and this does often act to undermine the use of mathematical tools such as maximising expected utility, Bayes' theorem, and even the Drake equation. However, the uncertainty in astrobiology need not paralyse us. The recognition of true uncertainty (as opposed to risk), that arises from a rich and unexplored research area, can point us to fascinating and presently obscured discoveries. Concerning computational models, their analytic nature precludes them from telling us anything synthetically new about the world. And, in this way, we cannot get certain data out of these models by inputting uncertain data. Computational models also run the risk of being over-interpreted. Nonetheless, so long as the functional form of the working parts of our models corresponds to the functional form of the working parts of our target system, computer modelling of epistemic communities can be a valuable tool in astrobiology, as in other sciences.

With this final conclusion in mind, let us now turn to Chapter Two of this thesis. Chapter Two, alongside Chapter Three, does not directly consider the impact that uncertainty has on utilising mathematical and computational models in astrobiology. However, these chapters take a central term within the field, *biosignature,* and evaluate how uncertainty over the list of possible options (captured by Stanford's problem of unconceived alternatives (Stanford, 2001, 2006a, 2006b) problematises its definition. Establishing this type of uncertainty within astrobiology, and subsequently offering a definition that acknowledges this, is an important building block to the later chapters of this thesis. Hence, it is to an analysis of the term *biosignature* that we now turn.

Chapter Two

The Call for a New Definition of Biosignature

Before delving into the impact that uncertainty has on mathematical and computational models in astrobiology, there is a central term within the field that suffers from uncertainty over the list of possible options. This term, biosignature, has become increasingly prevalent in astrobiology literature as our ability to search for life advances. Although this term has been helpful to the community, its definition is not settled. Existing definitions conflict sharply over the balance of evidence needed to establish a biosignature, which leads to misunderstanding and confusion about what is being claimed when biosignatures are purportedly detected.

To resolve this, I offer a new definition of a biosignature as any phenomenon for which biological processes are a known possible explanation and whose potential abiotic causes have been reasonably explored and ruled out. This definition is strong enough to do the work required of it in multiple contexts — from the search for life on Mars to exoplanet spectroscopy — where the quality and indeed quantity of obtainable evidence is markedly different. Moreover, it addresses the pernicious problem of unconceived abiotic mimics that is central to biosignature research and arises from uncertainty over the list of possible options. I show that the new definition yields intuitively satisfying judgments when applied to historical biosignature claims. I also reaffirm the importance of multidisciplinary work on abiotic mimics to narrow the gap between the detection of a biosignature and a confirmed discovery of life.

2.1. Introduction

Scattered throughout the literature in astrobiology is a term central to the search for extraterrestrial life, that is, *biosignature*. This term, however, has been defined inconsistently by different individuals at different times. Taken in a very general sense, a biosignature is a phenomenon which signals the presence of life (extraterrestrial or terrestrial). Although many in the scientific community and elsewhere have a general conception of what the term biosignature refers to, its widespread adoption as a technical term belies a fundamental ambiguity. Is a biosignature supposed to represent tentative evidence, strong evidence, or overwhelming evidence (a true "signature") for life? In other words, is the detection of a biosignature only the beginning of the process of discovering extraterrestrial life, or is it the successful conclusion of it? Without some consensus on this basic question, the term may be positively unhelpful, obscuring more than it reveals about the quality and finality of the evidence to which it is applied. Given recent calls for clearer communication about life detection (Green et al., 2021; Malaterre et al., 2023), now seems an opportune moment to revisit and reunify the disparate definitions of a biosignature.

To provide a motivating example of this disparity, the often-cited definition from the NASA Astrobiology Roadmap (Des Marais et al., 2003) states that: "A biosignature is an object, substance, and/or pattern whose origin *specifically requires* a biological agent" (p.234, emphasis added). This definition implies a very strong condition for any candidate biosignature: something is a biosignature if and only if life has created it. Explicitly rejecting this in favor of a weaker definition with a wider domain of applicability, Catling et al. (2018) preferred to stipulate that "a biosignature is any substance, group of substances, or phenomenon that provides evidence of life" (p.710). Similarly, Pohorille and Sokolowska (2020) defined biosignatures as "chemical species, features or processes that provide evidence for the presence of life" (p.1236). Any observation that merely hints at the presence of life would fall within the scope of these modest definitions. This ambiguity over the strength of a biosignature is problematic because it leads to misunderstanding and confusion about what is being claimed when biosignatures are purportedly detected.

Hence, this chapter provides a critical account of the term *biosignature* from a philosophical and methodological point of view and ultimately proposes a new definition that circumvents the shortcomings of existing definitions. Presently, §2 surveys the numerous definitions of biosignature in the relevant literature. This comprises both weak and strong definitions. Their various shortcomings will be discussed in turn. In response to the multitude of definitions, §3 argues that a pluralist approach to biosignature definitions cannot be sustained. Hence, §4 both makes and responds to the call for a single definition of biosignature that avoids some of the limitations of previous definitions. The resulting

definition is a pragmatic one that recognises the epistemic limitations of current biosignature research that result from uncertainty over the list of possible explanations. §5 then applies the new definition to three significant historical biosignature claims; it is found that the new definition produces sensible results. Some broader implications of the new definition are then discussed in §6 with a focus on multidisciplinary research and, finally, §7 summarises the key arguments of the chapter.

2.2. Working Definitions of Biosignature

The term "biosignature" appears increasingly in academic journals and popular science magazines alike. With this growing popularity, however, comes a growing catalogue of discrepant definitions (e.g., Thomas-Keprta et al., 2002; Des Marais et al., 2003, 2008; Catling et al., 2018; Pohorille & Sokolowska, 2020). All definitions involve an inferential relationship from the detection of a biosignature to the existence of life. The strength of that inference is, however, ambiguous. In the present study, I classify the various definitions in the relevant literature according to the strength of this inferential link, as follows.

2.1.1. Definitions with Weak Biosignature-to-Life Inferences

A modest definition of biosignature was defended by Catling et al. (2018). These authors proposed that "a biosignature is any substance, group of substances, or phenomenon that provides evidence of life" (p.710). Similarly, Pohorille and Sokolowska (2020) defined biosignatures as "chemical species, features or processes that provide evidence for the presence of life" (p.1236). The broad requirement that a biosignature *provide evidence* of life is significant in allowing any number of observations, with sliding scales of life-confirmation, to enter the ranks of biosignature. Now, this weakening of the term is deliberate and is done so that the term can be confidently used in practice, given that the science of life detection necessarily relies on data of limited quality and completeness. By allowing biosignatures to include observations that provide less than complete confidence in life, the term is made useful and escapes redundancy.

This pragmatic approach to biosignatures is seen also in the work of Schwieterman et al. (2018). They discussed the validity of considering a signal to be a biosignature when there is a non-zero probability that it had a non-biological origin. Correspondingly, Schwieterman et al. recognised that a definition of biosignature that allows for no leeway when regarding possible abiotic causes would always need to be prefaced with *potential*. No known or immediately foreseeable signal in exoplanet spectroscopy could be

regarded as a biosignature on such a definition, only a *potential* biosignature. Schwieterman et al., therefore, provided an alternative definition of biosignature as "the presence of a gas or other feature that is indicative of a biological agent" and hence "a gas may be a biosignature gas, even if the gas may have nonbiological sources (2018, p.666)." The authors pointed out that such a relaxed definition of biosignature guards against overconfidence by acknowledging the uncertainty inherent in life detection science. But do we really want even *very weak* evidence to count as a biosignature? This seems to stray from the intuitive conception of the term.

Consider, for example, Percival Lowell's sensational interpretation of apparent markings on Mars' surface. Lowell attributed these markings to canals built by an ancient Martian civilisation (Lowell, 1908, 1906, 1895). In his popular books, Lowell illustrated hundreds of these alleged canals, and he advanced the theory that they were created to transport water from Mars' polar caps to its equator. Lowell would use his canal theory to argue for the existence of life on Mars, concluding (somewhat meanderingly) "That Mars is inhabited by beings of some sort or other we may consider as certain as it is uncertain what those beings may be" (Lowell, 1906, p.376).

On the definitions offered by Catling et al., Pohorille and Sokolowska, and Schwieterman et al., Lowell could have justifiably insisted that his canals were biosignatures, since his observations did provide some (ultimately refuted) evidence for the existence of life. But most of Lowell's scientific contemporaries dismissed his views because more mundane alternative explanations involving geological features or optical illusions had not been ruled out (e.g., Evans and Maunder, 1903; Wallace, 1907); such mundane alternatives were ultimately vindicated. It is intuitively unacceptable for any observation whatsoever to count as a biosignature if it is merely consistent with the presence of life when alternative (plausible) explanations have not been ruled out.

Another study suggesting only a weak inference from biosignature to life is that of Neveu et al. (2018). These authors characterised biosignatures as follows: "Biosignature measurements directly seek features characteristic of life (such as complex organic matter not known to be formed through only chemical reactions or concentrations of biologically necessary or useful elements) as evidence of ongoing or past biological processes." (Neveu et. al., 2018, p.1). The two examples after the "such as" are, however, very different. The first suggests a definition stronger than the Catling et al., Pohorille and Sokolowska, and Schwieterman et al. definitions with the requirement that the measurement not be known to form through non-biological reactions alone. But the second, by permitting observations of concentrations of biologically necessary or useful elements to rise to the ranks of biosignature, suggests a considerably more liberal definition. Although Neveu et al. (ibid.) stressed the importance of context in evaluating whether

the observation may have an abiotic explanation, it is hard to see how concentrations of oxygen, or phosphorus, or any biologically useful element could be considered a biosignature in themselves (rather than just evidence of habitability as we might prefer to say) except on a very broad definition that would encompass a multitude of possible observations only weakly suggestive of life. Overall, then, the Neveu et al. definition faces the same limitations as the Catling et al., Pohorille and Sokolowska, and Schwieterman et al. definitions. A definition for biosignature that is more restricted in its conditions is required to exclude observations that should not be considered even passably good evidence for life.

2.1.2. Definitions with Strong Biosignature-to-Life Inferences

The widely cited NASA Astrobiology Roadmap (Des Marais et al., 2003) definition of biosignature states that "A biosignature is an object, substance, and/or pattern whose origin specifically requires a biological agent" (p.234)." Hence, the substance is a biosignature if and only if life has caused it: the discovery of a biosignature inescapably implies the existence of life itself in the environment under scrutiny. Such a strict definition is not vulnerable to the charge of having too low a bar for what constitutes a biosignature. Certainly, there is no risk of false positives sneaking into the list of NASA Astrobiology Roadmap (Des Marais et al., 2003) biosignatures. But might the bar actually be too high?

This strong definition, in practice, could almost never be used on account of the uncertainty that shrouds prospective biosignatures. The stipulation that the observation specifically requires a biological agent necessitates that all possible abiotic origins for the observation have been ruled out. It might seem at least possible to rule out all *known* abiotic alternative explanations, but how does the community go about ruling out the unknown ones? This "problem of unconceived alternatives" — explored more generally by Stanford (2001, 2006a) — surfaces continually within philosophy of science literature, e.g., from within the contexts of general relativity (Kashyap, 2023) to legal trials (Jellema, 2024) to Darwin's theory of pangenesis (Stanford, 2006b). Briefly, the problem concerns how a community can be certain of the truth of their current hypotheses if unforeseen alternative hypotheses may well be equally or more successful. Such underdetermination is worrisome as it throws into question the validity of any particular theory.

The problem of unconceived alternatives can readily be couched in terms of uncertainty as the problem foundationally arises from uncertainty over the list of possible hypotheses. When deciding which hypothesis is the best out of a set, we are often limited by an incomplete set. We are often unable, then, to select the best hypothesis, simply because this hypothesis might not be visible to us.

The problem of unconceived alternatives is particularly salient in the discovery and interpretation of signs

of life (Vickers et al., 2023). Astrobiology is a relatively immature field; breakthrough discoveries (in terms of observations, technology, and theory) are frequent. The field is growing, and potential research areas are rich. The implication is that there exists a vast pool of alternative candidate hypotheses that have yet to be explored. In particular, there may be innumerable abiotic processes that mimic evidence of life of which we cannot yet conceive. These abiotic processes might include mechanisms that create organised and complex pseudofossils (McMahon and Cosmidis, 2022), or a stable and significant disequilibrium of oxygen and methane — as long as a suitable energetic process is able to sustain such a disequilibrium, it would not violate the laws of physics (Simoncini et al., 2013; Thompson et al., 2022).

The NASA Astrobiology Roadmap (Des Marais et al., 2003) definition fails to account for the intellectual modesty required by the problem of unconceived alternatives. Unless, of course, advocates of this strong definition are content only to use the term with a prefacing qualifier such as *weak*, *modest*, *strong*, or *potential*. Indeed, the 2003 Astrobiology Roadmap offers a separate definition for a *potential* biosignature: "A *potential biosignature* is a feature that is consistent with biological processes and that, when it is encountered, challenges the researcher to attribute it either to inanimate or biological processes. Such detection might compel investigators to gather more data before reaching a conclusion as to the presence or absence of life" (Des Marais et al., 2003, p.234). This definition for potential biosignature captures the possibility of life as the cause but acknowledges that more data are needed for a conclusive determination of biogenicity. The requirement for the feature to "challenge" the researcher excludes trivial observations that are merely consistent with life without being seriously suggestive of it, assuming that "the researcher" is reasonably well informed about both "inanimate [and] biological processes." Hence, this definition is practicable and — an additional strength — seems implicitly to encourage the driving forward of science to increase confidence in a potential biosignature.

However, the question remains as to how — or whether — a potential biosignature can graduate to an *actual* biosignature as investigators gather more data. Under the two definitions provided by the 2003 Astrobiology Roadmap, it is apparent that this transition occurs only when there is complete confidence that the feature in question "specifically requires a biological agent," a threshold that may be extremely difficult to meet in practice regardless of how much data are gathered (not least because of the problem of unconceived alternatives). The term "potential biosignature" seems to be applicable to candidate life-detections with all degrees of confidence in biogenicity up until the limiting case of complete confidence (in which case the word "biosignature" is arguably redundant since what has been found is just *life*). A more lenient definition of biosignature that can work more effectively in tandem with "potential biosignature" is still needed.

Gargaud et al. (2009) also criticised definitions of biosignature that require the ruling out of all possible abiotic explanations. They recognised that this cannot, in practice, be done. As a result, Gargaud et al. remarked that "the word bio-signature is frequently used in an abusive way" (2009, p.595), and therefore they advocated for the term "bio-indices" instead. Bio-indices would have a lower threshold for confirmation than the NASA Astrobiology Roadmap (Des Marais et al., 2003) definition of biosignature. However, this term would fall within the class of weaker definitions subject to the same issues raised in §2.1. Moreover, introducing a new term to do the work of an existing term is cumbersome; perhaps all we need is a considered and unified repackaging of what it is to be a biosignature.

2.3. Should we embrace a plurality of definitions?

The above discussion has critiqued the popular definitions of biosignature to be found in the literature. It has been shown that these definitions exist along a scale of increasing exclusivity to observations that can be explained solely by the presence of life. The weaker of these definitions might deem a biosignature an observation that could be caused by life but could also be caused by abiotic processes, such that the observation in question is only weakly indicative of life. Conversely, stronger definitions require the total ruling out of abiotic causes, leaving the presence of life as the sole explanation. One might respond to this multitude of definitions by taking a pluralist view, on which different definitions of biosignature are appropriate to different measurement contexts and can thus coexist unproblematically. The consequence of this would be that no individual definition of biosignature should be universally preferred over any other. Moreover, apparent flaws in each definition might be neutralised given the specific context of applicability.

A parallel example is the pluralist approach to the definition of life. Even within science, the term has a multiplicity of definitions, but a popular one emerged from a committee gathered by NASA to discuss the possibility of life in the cosmos: "Life is a self-sustaining chemical system capable of undergoing Darwinian evolution" (Joyce, 1994). This definition is intuitive, precise, and widely cited. However, alternative conceptions of life in the disciplines of evolutionary biology, molecular biology, synthetic life, and the origins of life have been widely discussed (e.g., Machery, 2012; Mix, 2015).

Each discipline calls for a different focus on what constitutes life, and each discipline uses the term for its own ends. Any attempt at a unified definition would fail to do the work that each discipline needs (Machery, 2012). To argue that there is or should be an absolute and single definition of life would betray how the word is used, and therefore a pluralist definition of life is motivated on these grounds. The differing definitions need not be competing; rather each is suitable within its own context.

These differing definitions of life are justified insofar as the differences exist between disciplines, not within them. To prevent miscommunication, there needs to exist a shared language within research fields and thus a shared understanding of which definition of life is being employed. There indeed exists lively and fruitful debate about the most suitable definition of life within the field of astrobiology (e.g., Benner, 2010) — definitions that center on metabolism, for example, are popular. Such disagreement might be attributed to the varied fields that contribute to astrobiology, but many within the discipline clearly feel the need for a shared definition. So, to summarise, pluralism about the definition of life is justified, but with the general maxim of one definition per field.

Pluralism about the definition of a "biosignature" within astrobiology fails for the same reason. All discussed definitions of biosignature are applied within the astrobiology community, where those seeking evidence of life in different contexts communicate frequently with each other through the same meetings and journals (and are often simply the same people). Hence, a pluralist account of the term is likely to create more problems than it solves. To have a multitude of definitions for a single term in the same field would incite confusion and misinterpretation. One subset of astrobiologists may proclaim that the conditions for deeming an observation to be a biosignature have been met, only to be challenged by scientists from an adjacent area of astrobiology with stricter conditions. It is easy to imagine interpretation of evidence being blamed for the confusion when something as simple as misaligned definitions — an unnecessary semantic disagreement — is at fault.

2.4. The New Definition

The unity of astrobiology calls for a single definition of biosignature. Previous definitions discussed in this paper, however, have either allowed the admission of *any* observation that could result from life or been found impractical on the grounds of the problem of unconceived alternatives. With these considerations in mind, I propose the following definition of biosignature:

A <u>biosignature</u> is any phenomenon for which biological processes are a known possible explanation and whose potential abiotic causes have been reasonably explored and ruled out.

The first half of the definition requires that life be a possible explanation for the observation. This includes the requirement that the environment in which the phenomenon originated was habitable at the time. Regarding the second half, the definition does not require absolute confidence in a biotic origin, but it does capture the requirement that the accessible (given the scientific understanding of the time) abiotic alternative explanations be uncovered and disqualified. What is more, unconceived alternative

abiotic explanations that are not readily in reach need not be ruled out (which, of course, they cannot be). As a result, the above definition is stronger than the widely cited Catling et al. definition but not as strong as that of Des Marais et al. (2003).

There are two significant features of this definition that should be explored: the first concerns the phrase "reasonably explored and ruled out" and the other addresses the gap between a biosignature and life. To better pin down and understand the implications of the proposed definition, these features will be discussed in turn.

2.4.1. "Reasonable" Exploration and Ruling Out

The requirement that the potential alternatives be *reasonably* explored and ruled out must be explicated. Much discussion has been devoted to the inability to rule out all the unknown abiotic explanations (§2.2 of this paper, and, e.g., Vickers et al., 2023). This problem is avoided when focusing just on the known ones, but it is still not possible to rule out all known alternative explanations *with absolute certainty*. Such certainty is rare in any science, let alone in the rapidly developing field of astrobiology. Nonetheless, a high bar must be set to ensure that the term *biosignature* carries weight. For the alternative non-biological explanations to be reasonably explored and ruled out, three considerations should be made; these concern time, attention, and consensus. Concerning time and attention, these are necessary to attempt to lessen, though unlikely close, the gap between the list of known alternative explanations and the list of unknown ones. Giving time and attention to a biosignature claim will ideally reveal additional hypotheses about possible non-biological origins that were not initially considered, and these further lines of enquiry can then be explored and closed (Green et al., 2021).

On the path to a phenomenon being categorised as a biosignature, it may enter an intermediate phase where 1) life has been identified as a possible cause of a phenomenon, 2) there does not currently exist a complete abiotic explanation, but 3) the community has not yet given the time nor attention needed to reasonably explore and rule out abiotic alternatives and consequently a consensus has not been met. This is where the term *potential* biosignature would be most appropriate. Resultingly, I define a potential biosignature as follows:

A potential biosignature is any phenomenon for which biological processes are a known possible explanation but whose potential abiotic causes have not yet been reasonably explored and ruled out.

This proposal is in good alignment with the NASA Astrobiology Roadmap (Des Marais et al., 2003)

definition of potential biosignature in that it encourages the active search for abiotic alternative explanations. The definition of potential biosignature proposed here similarly "challenges the researcher to attribute it either to inanimate or biological processes" (Des Marais et al., 2003, p.234), which helps to drive the science forward.

The adoption of our definition of biosignature would enable debates about whether something is a biosignature to focus on the science rather than the semantics. To justify classifying a given phenomenon as a biosignature, a researcher need only conclude that the accessible alternative explanations have been reasonably explored and ruled out. Those who think they have not been ruled out will disagree that the phenomenon is a biosignature. This is a substantive disagreement that will drive the science forward, not a fruitless semantic one. Nevertheless, if there is significant disagreement in the community about whether alternative explanations have been reasonably ruled out, this in and of itself would tend to suggest that they have not. Thus, particularly for non-experts or journalists hoping to evaluate whether something is or is not justly called a biosignature, the state of consensus in the field is significant.

An appeal to the level of consensus in the scientific community as a gauge of confidence in a scientific statement is not novel. The Intergovernmental Panel on Climate Change (IPCC) employs a metric where broad scientific agreement increases confidence in statements about climate change (Mastrandrea et al., 2010). The theoretical backdrop to the framework is that increasing consensus tracks with decreasing uncertainty. This idea has been explored extensively in Vickers' book *Identifying Future-Proof Science* (2022). A central claim defended is that, for any scientific hypothesis, once a 95% consensus is achieved amongst the scientific community, that hypothesis should be considered *future proof science.* This is argued to be the case so long as the scientific community is large, international, and diverse (Vickers, 2022, p.21)

In Vickers et al. (2023), we propose and motivate the transfer of the IPCC framework to confidence in biosignature claims. This metric is shown in Figure 2.1. For something, therefore, to be classified as a biosignature, we should require that there exists robust evidence in favour of this and that there is high agreement amongst the scientific community. This would require that biosignatures sit in the top right box of Figure 2.1.

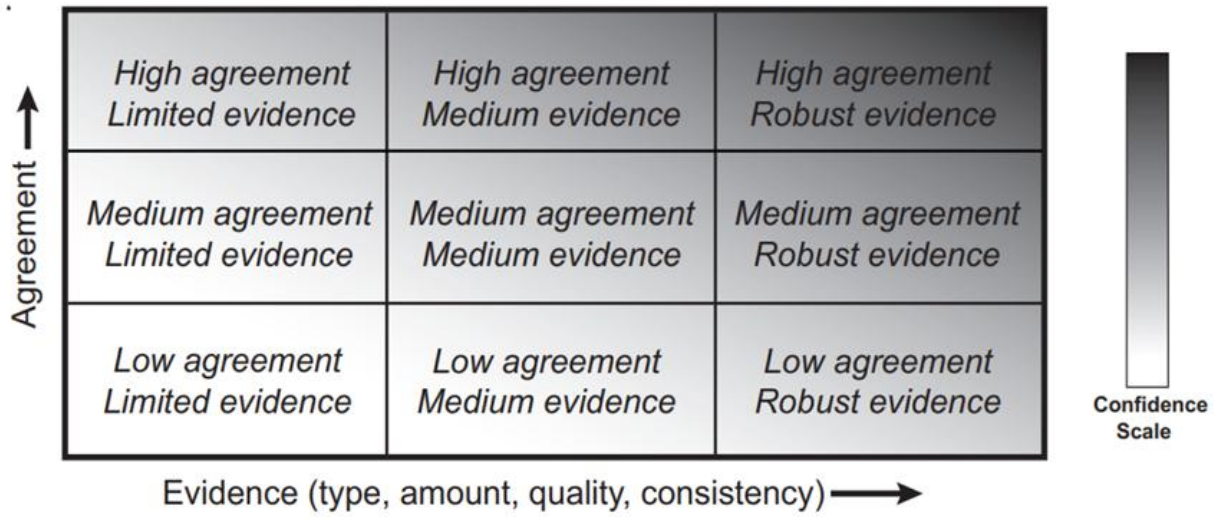


Figure 2.1. The IPCC framework for evaluating confidence in statements about climate change. The x-axis captures the robustness of available evidence; the y-axis captures the agreement amongst the community over the evidence (Mastrandrea et al., 2010; also repurposed for confidence in biosignature claims in Vickers et al., 2023).

It is expected that the level of consensus will grow over time and the rate of this will depend on various relevant factors. The amount and quality of evidence would be influential: as would the availability of relevant theory and technology. Some scientists would be willing to call an observation a biosignature before others, but it is when there is significant consensus in the community that the possible abiotic explanations have been reasonably explored and ruled out, that an observation should be called a biosignature.

2.4.2. The Gap Between Biosignatures and Life

Let us imagine that an observation meets the requirements for a biosignature, under this definition. This is to say that the abiotic alternative explanations have been reasonably explored and ruled out. The question will naturally arise of how certain one can be of its biotic origin. Could the community attempt to quantify their confidence in a biotic cause; might they ascribe 30%, 50%, or 70% confidence? Unfortunately, the choice of quantifier is itself unlikely to be robustly underpinned by available evidence. To assert that we are, say, 70% certain of the biotic origin of a biosignature is to suggest that we have exhaustively explored the causal possibility space for the biosignature and have determined that the abiotic causes have a 30% chance of providing the cause, and a biotic origin has a 70% chance. Yet, as noted previously, there is an acute problem of unconceived alternatives in astrobiology: we are unaware as to what proportion of the causal possibility space we have explored in any of the contexts where biosignatures are sought. To use an analogy: we have scratched part of our scratchcard, but we have no idea how large that card is.

NASA-affiliated researchers have proposed a confidence ranking with a seven-point scale – the "confidence of life detection" (CoLD) scale (Green et al., 2021). The purpose of the CoLD scale is to standardise the stages that a potential biosignature moves through as its biotic origin is tested. The certainty that any observation derives from life increases as the observation moves up the scale, with level seven corresponding to "confirmation of the presence of biology" (Green et al., 2021, p.577). Significantly, level four requires that "all known non-biological sources of signal shown to be implausible in that environment" (p.577). This is indeed a realistic and achievable goal and the new definition of biosignature is in line with this; the definition of biosignature proposed in this chapter would sit at level four on the CoLD scale.

However, the CoLD scale quietly closes the gap between having ruled out all the *known* alternatives and having ruled out all the *possible* alternatives. To arrive at level seven, with the confirmation of biology, an assumption is made that the known alternatives at this point comprise the complete set of alternatives. Clearly, the gap between level four and level seven may be very large or very small depending on the extent to which the alternatives have been explored, and this ambiguity undermines the usefulness of the scale (see the discussion in §3 of Vickers et al. 2023).

The limitations of our knowledge of the relevant possibility space are made salient in McMahon and Cosmidis' "False biosignatures on Mars: anticipating ambiguity" (2021). Speaking specifically about geological biosignatures on Mars, they recognise that, to rule out abiotic mimics, not only do the mimics need to be discovered in the first place (which typically happens by serendipity rather than forethought), but the mechanisms of these mimics need to be understood in enough detail to predict how and when such mimics will arise (p.28). It is simply not the case that we sufficiently understand all possible abiotic mimics: "given the haphazard and unsystematic way in which varieties of false biosignature have so far been identified, we can only assume that many others remain undiscovered" (McMahon and Cosmidis, 2021, p.29).

With levels of confidence dependent on the evaluation of a possibility space that is simply not wholly accessible to us, quantifying the certainty of an observation's biotic origin becomes tricky. What is instead achievable is assessing how thoroughly we have ruled out the known abiotic mimics. By evaluating the how robust the evidence is, and how much agreement exists amongst the astrobiology community, we may *qualitatively* comment on our confidence, as per the IPCC framework (Mastrandrea et al., 2010; Vickers et al., 2023). However, there is always the possibility that numerous undiscovered abiotic processes that mimic biotic ones exist, even if we find ourselves in the high-robustness, high-agreement box of Figure 2.1.

This discussion on the gap between a biosignature and a confirmed consequence of life strikes parallels with the philosophical literature on what constitutes *knowledge*. The traditional view of knowledge equates it to justified true belief, with this view having its origins as far back as Plato, but more recently revived by Lewis (1946) and Ayer (1956, pp. 9–11, 254–259). The threshold by which a true belief is justified, though, has been debated widely. A high bar for this has been defended in Plantinga's warrant theory (Plantinga, 1993, p.48) and reliabilist theories of knowledge (Goldman, 1979). However, in her "The Inescapability of Gettier Problems", Zagzebski (1994) argues that, regardless of how high the threshold is for justification, Gettier problems (Gettier, 1963) will always be problematic for a view of knowledge as justified true belief. Hence Zagzebski concludes that, regardless of how high the threshold is for justification, there will always remain a gap between justified true belief and knowledge.

This epistemic gap is also present when concerning biosignatures. I have argued then that, for any given phenomenon to qualify as a biosignature, we should require that the known abiotic causes have been reasonably explored and ruled out. The jump from there to the conclusion that the phenomenon is most likely caused by life remains just that – a jump. The most that can be done is to assess whether the conditions for a biosignature have been met and to acknowledge that this is simply our best guess at the validity of any life-detection claim. There is therefore an epistemic gap between something being classified as a biosignature and that thing being caused by life. As our understanding of abiotic mimics and the conditions of habitable worlds progresses, this gap might be narrowed, but it should nonetheless be made explicit.

2.5. Applying the New Definition

Having motivated a new definition, it remains to apply the definition to past claims of biosignature detection to assess how the term might be used in practice and how it might have helped mitigate the bumpy and revisionist history of biosignatures. Despite the lack of consensus in the evidentiary criteria of a biosignature, biosignature claims have been abundant throughout the history of astrobiology; with the hasty retraction of such claims being almost equal in number.

2.5.1. ALH84001

A prominent case study to test the proposed definition of biosignature comes from the Martian meteorite discovered in Antarctica 1984: Allan Hills 84001 (ALH84001). In a paper in *Science* in 1996, McKay et al. argued that the evidence supports a conclusion that Martian microorganisms lived on the meteorite. The

authors highlight three key observations regarding apparent fossil structures on the meteorite and conclude that "although there are alternative explanations for each of these phenomena taken individually, when they are considered collectively, particularly in view of their spatial association, we conclude that they are evidence for primitive life on Mars" (McKay et al., 1996, p.929). The likelihood of an abiotic cause that neatly explained all three lines of evidence seemed, to the researchers, much smaller than explaining one of them alone, whereas the presence of life on Mars appeared to account for the three observations well.

Having said this, the Martian microorganism hypothesis was not uncontroversial even in 1996. Alternative explanations including volcanic activity, impact events, and hydrological exposure were suggested but were unable to be convincingly demonstrated as the cause at the time (e.g., Harvey & McSween, 1996). However, a series of progressive refutations over the following 25 years provided abiotic explanations for each of the three key observations (see, for example: Golden et al., 2001, 2004, 2006 and Bell, 2007). Most recently, Steele et al. (2022) showed that mineral carbonation and serpentinisation reactions on an early Mars explain the occurrence of organic matter in the meteorite. Importantly, it is recognised that advancements in relevant technology made this discovery possible. The claim that ALH84001 harboured fossilised Martian microorganisms was largely debunked and its status as a biosignature invalidated. How, though, would this event have played out if the new definition of biosignature were employed? Should the structures on ALH84001 have been classified as a biosignature?

The answer to this depends on whether the accessible (*at the time)* non-biological alternative explanations were reasonably explored and ruled out. In this example potential abiotic causes were suggested before they could be fully tested (in part because methodological advancements were needed). Nonetheless, such limitations in technology should not validate a total bypassing of the exploration of any suggested abiotic explanation. In the case of the structures on ALH84001, fruitful lines of enquiry remained open. Potential abiotic causes were not reasonably ruled out and consequently, the structures on ALH84001 should not have been considered a biosignature, either in 1996 or subsequently.

2.5.2. Stromatolites 3.7 Billion Years Old

Consider now a terrestrial biosignature claim. Astrobiology, broadly construed, also encompasses the study of early life on Earth and the term *biosignature* is also correspondingly used for terrestrial life. Among the oldest accepted biosignatures are stromatolites in the East Pilbara Terrane, dating back to around 3,500 – 3,400 million years ago (e.g., Walter et al., 1980; Van Kranendonk et al., 2008; Baumgartner et al., 2019). However, Nutman et al. (2016) argued for the existence of yet older stromatolites in 3,700 million-year-

old rocks in the Isua supercrustal belt of southwestern Greenland. Nutman et al. offered both a positive account for how life could have produced these structures, and a negative account of the competing abiotic explanations: after considering four features which particularly evaded non-biotic explanation, they concluded that "we rule out an abiogenic origin for Isua stromatolites" (2016, p.3).

However, only two years later, Allwood et al. (2018) challenged the biotic explanation and instead offered an abiotic cause for the proposed stromatolites. Allwood et al. attributed the apparent stromatolites to structural deformation and geochemical alteration of the rock. The debate continues: Nutman et al. (2019, 2021) responded to Allwood et al. (2018), and the discussion remains lively (e.g., see also Zawaski, et al., 2020). So how might these potential stromatolites be best categorised? A biotic account has been provided. Abiotic explanations, however, are still actively under investigation, and there is a lack of consensus. The term "potential biosignature," as defined in §2.4.1, is therefore most appropriate in this case.

2.5.3. Petroleum

The story of Earth's major crustal petroleum accumulations offers an example of a terrestrial biosignature claim whose biogenicity has historically been controversial. Towards the end of the 19th century, it was concluded by various North American geologists that petroleum originated from the remains of marine animals and plants; they began to recognise organic-rich, fossiliferous shales as the source rocks (e.g., Hunt, 1963; Lesquereux, 1866; Newberry, 1873). By the middle of the 20th century, it was clear that many of the molecular constituents of petroleum could be traced back to specific biomolecules such as porphyrins and bacterial membrane lipids (Treibs, 1936).

Nevertheless, the hypothesis of a non-biological origin of petroleum was maintained by a vocal minority of Soviet and American scientists in the late 20th century who argued that oil is produced by non-biological chemical reactions at high temperatures and pressures in the deep crust or mantle (e.g., Kudryavtsev, 1951). They argued that biological molecules in the oil were simply contaminants. These dissenting voices were prominent particularly in the Soviet Union; in the West, they notably included the astrophysicists Fred Hoyle and Thomas Gold (Hoyle, 1955; Gold, 1985). The theory of abiotic petroleum seemed to be supported by evidence for hydrocarbons on other planetary bodies in our solar system, such as asteroids and comets (Cronin et al., 1988).

The origin of oil is now exceptionally well understood. Each stage of the production of petroleum from organic biomass via natural processes of decay, diagenesis, and catagenesis has been explored in the field and in the laboratory (e.g., He, 2018). We know that sedimentary rocks rich in biological organic matter

are effective source rocks for oil, and we can explain the origin of all the world's major economic petroleum reserves very well on the biological model. Abiotic hydrocarbons, especially methane, do occur naturally but do not contribute significantly to the world's known petroleum resources (Sherwood Lollar et al., 2002; Walters, 2017). Petroleum contains a suite of biomarker molecules that only biological processes could generate in such quantities and distributions, and these molecules derive overwhelmingly from the same source as the oil itself, that is, ancient biomass (e.g., Grantham & Wakefield, 1988). There is no longer any doubt that the vast majority of the world's oil derives from this. This understanding has been extensively tested and put into practice by oil companies and geologists.

From the suggestion of petroleum as biotic in origin, up to its wide acceptance as such, petroleum could aptly have been described as a *potential* biosignature. Life was recognised as a coherent explanation, but time was needed to investigate the abiotic alternatives. Field and experimental evidence in the late 20th and early 21st centuries settled the debate. Petroleum with a composition like that of the major oilfields on Earth (which is markedly distinct from abiotic hydrocarbon accumulations) now stands as a biosignature under the proposed definition.

2.5.4. Table of Categorisation

The three case studies considered above show how the new definition might be used in practice and how it might help avoid hasty conclusions regarding the presence of life. Examples of neither biosignature nor potential biosignature (structures on ALH84001), potential biosignature (3.7Gyr stromatolites), and biosignature (petroleum) have been discussed. The following table sorts a wider selection of terrestrial and extraterrestrial biosignature claims throughout recent history into these three categorisations. A brief explanation for the categorisation is provided, though the decision over biosignature status is only the opinion of the authors and should alternatively, in practice, be arrived at collectively by the scientific community.

Phenomenon	Biosignature	Potential Biosignature	Neither	Justification
Combined features of meteorite ALH84001 (McKay et al., 1996)			✓	Some possible abiotic causes were known at the time, but could not be ruled out (e.g., due to technological limitations).
3.7 Gyr Stromatolites (Nutman et al., 2016, 2019, 2021; Allwood et al., 2018)		✓		Life has been proposed as a convincing explanation for these candidate stromatolites (Nutman et al., 2016). However, the debate is lively with possible abiotic explanations suggested (Allwood et al., 2018), alongside counter arguments (Nutman et al., 2019, 2021).
Petroleum (major accumulations)	✓			A biotic explanation exists, and any abiotic explanations have been well considered and ruled out.
"Sinton bands" indicating life on Mars (Sinton, 1959)			✓	The abiotic cause of telluric water vapor (Rea et al., 1965) was reasonably within reach at the time, yet was not ruled out.
Viking Labelled Release experiment		✓?	✓?	Abiotic explanations have been suggested (see, for example, Navarro-González et al., 2010), but there is still a lack of consensus; no account is completely satisfying as yet. Such grey-area cases are to be expected.

Oumuamua as an alien artefact			✓	A (somewhat outlandish) biotic explanation was proposed (Loeb, 2018; Bialy and Loeb, 2018) to explain Oumuamua's unusual properties. However, an abiotic explanation has been developed (Bergner & Seligman, 2023) and the scientific consensus sides with the abiotic explanation.
3.8 Gyr Carbon Isotopes		✓		A biotic explanation is available, but debate is ongoing, and the picture is unclear (see e.g., Lepland et al., 2005)
Venusian Phosphine (Greaves et al., 2021)			✓	Although phosphine above a certain abundance on Venus might well be considered a potential biosignature, its presence in the cloud decks of Venus has been disputed (Villanueva, 2021).

Table 2.1: Selection of biosignature claims and their suggested categorisation as either a 1) biosignature, 2) potential biosignature, or 3) neither, under the proposed definition of biosignature.

2.6. Broader Implications of the New Definition

The new definition requires the explicit exploration and ruling out of non-biological explanations for the term *biosignature* to be used. The examples in Table 2.1 of phenomena that are *not* biosignatures show that non-biological explanations often arise from differing areas of science. For example, the fossil-like structures on ALH84004 were elucidated by research into the conditions on an early Mars and the effects of fluids on rocks; open questions regarding Oumuamua's non-biological origin concern the outgassing of exotic volatiles and comparisons with the trajectories of unusual comets like 2P/Encke; and, in the case of Sinton bands, the prevailing explanation came from atmospheric science and physics.

What these varied explanations show is that an unconceived alternative explanation may be hiding in any domain of science. If we are truly to rule out the known or accessible abiotic causes of any observation, a multidisciplinary effort is required. Possible explanations need to be pulled from broad fields and explored collaboratively. Such an approach will more effectively uncover new abiotic processes, leading

to a more complete exhaustion of the causal processes at play in astrobiology. More generally, if the epistemic gap between a biosignature and life is to be narrowed, a multidisciplinary approach to abiotic explanations will be essential. The narrowing of the gap between known abiotic mimics and possible abiotic mimics will lead to the narrowing of the gap between a biosignature and an actual detection of life.

Hence, the new definition reaffirms the widely held view that, for astrobiology to flourish, collaboration between its composite disciplines is essential. This is a sentiment reflected in other scientific disciplines, for example, the eventual acceptance of Wegner's continental drift theory only resulted from the collaboration of paleontologists, geologists, paleobotanists, and others (Vickers, 2022, Chapter Five) and the collaboration of biologists, genealogists, engineers, and computer scientists bringing about the success of the Human Genome Project (Hood & Rowen, 2013). The new definition of biosignature, and the arguments made in defence of it in this chapter, endorse the benefit of multidisciplinary research in astrobiology.

2.7. Conclusions

This chapter has presented and discussed the various definitions of the term *biosignature* extant in the literature. Definitions that require all possible (known and unknown) abiotic causes for the observation to be ruled out are impracticable, particularly given the uncertainty over candidate hypothesis captured by the problem of unconceived alternatives. Definitions that simply require life to be *a* possible explanation are too weak to justify the term *biosignature*. I have considered a pluralistic account of the term but argued that astrobiologists working on the search for life in different contexts constitute a single community in which a multiplicity of definitions for the same term is bound to cause unnecessary confusion. I, therefore, claim that a single definition of biosignature is needed. I repeat here the definition I have proposed and defended:

A <u>biosignature</u> is any phenomenon for which biological processes are a known possible explanation and whose potential abiotic causes have been reasonably explored and ruled out.

This definition acknowledges the problem of unconceived alternatives: the need to rule out all *unknown* abiotic causes before identifying something as a biosignature is removed and the proposed definition is sympathetic to the call for a pragmatic definition pursued by Schwieterman et al. (2018) that strikes the balance between exclusivity and usability. The proposed definition has been applied to historical cases to

illustrate how the definition would work in practice.

A new definition for potential biosignature has also been proposed that complements the definition of biosignature. Again, this definition recognises the epistemic limitations imposed by the problem of unconceived alternatives and encourages the active search for abiotic mimics. Potential biosignatures are hence defined as so:

A potential biosignature is any phenomenon for which biological processes are a known possible explanation but whose potential abiotic causes have not yet been reasonably explored and ruled out.

Finally, I have acknowledged that discovering a biosignature is not the same as discovering life itself; biosignatures are limited, observable proxies for life and are inherently uncertain. Definitions of biosignature that overlook this gap (e.g., NASA Astrobiology Roadmap, 2003) are impracticable. For any observation, this gap should narrow as a greater understanding of abiotic processes builds and these abiotic alternative explanations can be explored and ruled out. The proposed new definition encourages such progress by requiring that, for something to be a biosignature, potential abiotic causes must have been reasonably explored and ruled out.

Chapter Three

A Defence of the Proposed Definition of Biosignature: Epistemic vs. Ontic Definitions

Since the publication of "The Call for a New Definition of Biosignature" (Gillen et al., 2023), criticisms of the paper's proposals have been made. Primarily, opposition to the epistemic nature of the Gillen et al. definition, in preference for an ontic definition of biosignature, exists (e.g., Cowie, 2023b). The motivations for abandoning the epistemic nature of the definition might initially appear intuitive. However, the costs of this outweigh any benefits that come from appealing to a more typical ontic definition. Consequently, this chapter will focus on a defence of the Gillen et al. definition, in light of recent literature that looks to undermine it.

3.1. Introduction

In a paper published in *Mind* and titled: "New Work on Biosignatures" (2023b), Cowie provides a critical survey of the oft-cited definitions of biosignature circulating the astrobiology literature. His goals are to 1) partition the definitions into two camps: subjective definitions and objective definitions, and 2) argue in favour of an objective definition of biosignature that appeals to the metaphysical posits of types and tokens.

By arguing against the very principle of the subjective definitions of biosignature, Cowie (2023b) stands to undermine the Gillen et al. (2023) definition. As a reminder of the proposed definition, Gillen et al. (2023) define a biosignature as: *any phenomenon for which biological processes are a known possible explanation and whose potential abiotic causes have been reasonably explored and ruled out* (Gillen et al., 2023, p.1228). The subjective nature of this definition comes directly from the word "known" and somewhat indirectly from the phrase

"reasonably explored and ruled out". What this subjectivity means for the definition is that the truth status of a particular biosignature claim is dependent on an individual's (or group's) knowledge and belief set. Under this definition of biosignature, an individual needs to explicitly *know* that the presence of biology can be a potential explanation, and they must have the knowledge that visible alternative, abiotic explanations are unviable.

In contrast, what Cowie (2023b) calls *objective* definitions are those whose correct application holds either false or true, independent of an individual's knowledge and beliefs. An oft-cited example of an objective biosignature definition comes from Des Marais et al. (2003): "A biosignature is an object, substance, and/or pattern whose origin specifically requires a biological agent" (p.234). With this definition, a biosignature is just something that has originated from a biological agent, whether or not an individual knows this or has even perceived it.

The distinction between subjective and objective definitions is a useful one. However, there is an analogous pairing that may be even more illuminating in this case. This is the distinction between epistemic and ontic definitions. Broadly construed, subjective definitions are epistemic ones, and objective definitions are ontic. The arguments raised by Cowie can instead be couched in terms of epistemic vs ontic definitions of biosignature. This chapter adopts this language as it allows a more pointed discussion of the advantages and disadvantages of the Gillen et al. definition while still directly addressing the concerns raised by Cowie (2023b).

The structure of the chapter then is as follows: §3.2 discusses the prevalence and intuitive appeal of ontic definitions. These definitions usually take the default position, especially within science, and hence the reasons for this will be discussed here. §3.3 then introduces the idea of epistemic (pertaining to knowledge) and doxastic (pertaining to belief) definitions. Following this, §3.4 provides two scenarios under which epistemic/doxastic definitions should be preferred. §3.5 makes the case for biosignatures falling under the second of the scenarios discussed. The benefits of adopting an epistemic/doxastic definition of biosignature are first presented and then defended here. §3.6 raises and responds to two potentially undesirable consequences of an epistemic definition of biosignature. A conclusion is drawn in §3.7 that the term *biosignature* should be reserved for the epistemic state of having gone as far as we can with the available evidence. Thus, an epistemic definition, as proposed in Gillen et al. (2023), is more defensible than Cowie's ontic one (Cowie, 2023b).

3.2. The Prevalence and Intuitive Appeal of Ontic Definitions

It is true that the majority of definitions in science, and more generally across other disciplines, are ontic. For example, Britannica's definition of a star: "Any massive self-luminous celestial body of gas that shines by radiation derived from its internal energy sources" (Aller et al., 2024). A star, by this definition, is just the ontological content of the definition. There is no requirement that the scientific community know that the criteria for a star has been satisfied in order for something to be a star.

Another example of an ontic definition is that of life. Although the actual convergence on a singular definition of life remains evasive, the popular Joyce definition states that: "Life is a self-sustaining chemical system capable of undergoing Darwinian evolution" (1994). Again, this represents an ontic definition whereby something is either alive or not just if it meets the conditions of this definition and regardless of whether it is known to be alive.

As a principle, ontic definitions remove, or at least ignore, any agent from the definition. Such definitions, therefore, will genuinely refer to things even in a world with no humans to observe them. A star is star even if a child believes it to be a firefly; an imaginary friend is not alive even if a child believes that it is. Ontic definitions have intuitive appeal, particularly within science, because they assume an underlying allegiance to an objective world. Such definitions imply that the world is something to be investigated and discovered. We can be correct or wrong in our knowledge about the world, but our knowledge does not actively change the way the world is.

It is an unproblematic, and even welcomed, aspect of science that we can say statements like the following: we used to think that the sun orbited the Earth, but we were wrong; it never did. Science assumes a separation between the way things are and our knowledge of those things. In a sympathetic view, our knowledge and beliefs mediate between us and the real world. In a more pessimistic view, they can obscure us and lead us astray. It seems to make sense, then, that many definitions in science are ontic; science aims at objective truth, and hence, knowledge and belief should not be baked into most scientific definitions.

3.3. Epistemic and Doxastic Definitions

In contrast with the ontic definitions discussed in the prior section, epistemic definitions are definitions with a knowledge state embedded into them. Doxastic definitions are similar, but instead of a knowledge state, they have a belief state embedded into them. The distinction between knowledge and belief has long represented a hotly fought-after grey area — e.g., with the proponents of knowledge as justified true

belief including Ayer (1956, pp. 9–11, 254–259) contrasting with Sartwell's knowledge as merely true belief (Sartwell, 1991). A position on this distinction is not needed for the following discussion and so will be wholeheartedly avoided. From here on out, epistemic and doxastic definitions will be grouped together. The important thing is that, although epistemic and doxastic definitions may differ, they are the same in the relevant way when compared to ontic definitions.

An interesting example of a doxastic definition is that of legal guilt. Cornell Law School defines it as "the finding by a judge or jury that the defendant has committed the crime" (Legal Information Institute, n.d.). Elsewhere, legal guilt is defined along the lines of a judge or jury having decided *beyond reasonable doubt* that a defendant has committed a crime (Crown Office & Procurator Fiscal Service, 2024). What these definitions of legal guilt rest on is that a doxastic state is needed in order for legal guilt to be ascribed. Someone is legally guilty if and only if a legal court believes them (in the relevant way) to be guilty. It is true that defendants once found guilty have since been exonerated, but this does not mean they were never legally guilty to begin with. The case of legal guilt is not synonymous with the aforementioned example of the Earth's position in the solar system.

Significantly, legal guilt is distinguished from factual guilt in both the UK and US judicial systems – the latter being defined as a defendant having actually committed a crime (Chamberlin Law Firm, 2018). This distinction manifests no clearer than with the peculiar Alford plea, which is a legal plea in the US judicial system in all but three states. In 2009, the Minnesota House of Representatives defined the Alford pleas as: "a form of a guilty plea in which the defendant asserts innocence but acknowledges on the record that the prosecutor could present enough evidence to prove guilt" (Minnesota House of Representatives, 2009). In this way, the defendant is claiming that they are factually innocent, yet they make a legal guilty plea on account of believing that any reasonable jury would find them guilty on the available evidence.

Double jeopardy laws represent a further quirk in the epistemic nature of legal guilt. Double jeopardy comes from Roman law *non bis in idem* — not twice against the same (Buckland, 1963) — and protects against an individual being tried more than once for the same crime. Most countries uphold some form of double jeopardy, though some have more conditions than others. Before 2005 and the introduction of sections 75-83 of the Criminal Justice Act 2003 (Ashworth & Player, 2003), England and Wales prohibited the re-prosecution of an individual, even in cases where new overwhelming evidence had been discovered. However, since 2005, such new evidence (e.g., the discovery of DNA) is sufficient to override double jeopardy.

The existence of double jeopardy laws shows that, pre-2005, an individual who has previously been found innocent in a court of law, could later confess to the same crime and even have DNA evidence against them, but they are still considered legally innocent. This is the case simply because a court of law once

believed it to be the case. This epistemic definition of legal guilt results in these peculiarities, and yet the judge-jury designation of legal guilt has been around since at least the Normans (Stephens, 1896). Why, then, has this epistemic definition remained so useful over the course of a millennium?

The success of *legal guilt* as an epistemic definition simply comes from it being highly pragmatic. The term is useful, and when it comes to the longevity of words, this is key. The rival term of factual guilt is actually not particularly useful when it comes to the legal justice system, as any determination of factual guilt is going to be obscured by uncertainties. It is rare to be able to definitively ascertain someone's factual guilt, and hence the benchmark of legal guilt is used instead. Legal guilt represents an epistemic position that best approximates the factual guilt status of a defendant, given the available evidence. Without the epistemic definition of legal guilt, there would be no ability to convict or acquit defendants, and hence the entire concept of a legal system would be upended; the reasonable doubt baked into legal guilt captures the lack of clarity with which the legal lens attempts to view the facts.

3.4. The Conditions under which Epistemic/Doxastic Definitions are Most Appropriate

The definition of legal guilt exemplifies the utility of a doxastic definition. But can we draw some general guidelines for when epistemic/doxastic definitions are more suitable than ontic ones? I propose that there are two scenarios whereby epistemic/doxastic definitions should be employed over ontic ones. The first of these is when the term in question does not refer to an objective truth but rather a subjective one. The second scenario is when we have limited information about the objective truth, and so, rather than having a definition that refers to the real state of the world, we settle for a definition that captures our best understanding of it.

An example of this first scenario is one popular conception of beauty. Hume writes of beauty: "Beauty is no quality in things themselves: It exists merely in the mind which contemplates them; and each mind perceives a different beauty" (Hume, 1757, p.136). This captures a general sentiment that beauty is subjective; something is beautiful to someone just if they believe it to be beautiful. In cases such as Hume's conception of beauty — where there is no objective, agent-independent truth — striving for an ontic definition is bound to fail. An ontic definition would simply miss a crucial part of what the common-sense meaning of "beauty" captures: that is that people find the thing beautiful. An epistemic/doxastic definition of beauty is hence appropriate here on account of the central role that individuals' beliefs play. Other such definitions might arguably include *faith, hope,* and *despair*.

Let us turn now to the second scenario in which epistemic/doxastic definitions are more appropriate than ontic ones. This concerns cases where there may well be an objective state of affairs, but our access to it is significantly limited. Moreover, with such cases, there is a pragmatic need for an epistemic or doxastic definition. As an example, the definition of legal guilt meets these criteria well. The ability of a jury to access factual guilt is largely compromised. They are instead working with a limited data set. Their determination of legal guilt is their best approximation to factual guilt, given the available evidence. Furthermore, there is a real need for this epistemic definition. The legal system needs a word to represent our best approximation, and legal guilt captures this epistemic state.

Other examples of this scenario in which epistemic or doxastic definitions should be employed can be found in highly uncertain areas of science. If it is the case that, at present, no more progress can be made on the objective truth of a proposition, but there is utility in naming and communicating the epistemic status, then epistemic/doxastic definitions are appropriate. Some definitions of disease stand as interesting examples. Kukla (2022), for example, defends a pragmatic account of disease whereby the determination of some diseases is context-dependent. More specifically, one component of Kukla's conception of disease is that something should be classified as a disease if there is pragmatic utility in doing so. In such a way, an individual's changing beliefs about the utility of a diagnosis can, alone, change a determination of 'no disease' to 'disease'.

Another pertinent example of an epistemic definition can be found within palaeontology. The term 'fossil' is reasonably well understood to refer to the preserved remains of a once-living form, often in rock. Such a term is best understood as ontic. However, the term 'dubiofossil' has been used and defended (e.g., McMahon et al., 2021; Buick, 1990; Hofmann, 1972), and this is strictly epistemic. Dubiofosils are things which resemble fossils, but the evidence is dubious; we are unsure if the object in question is a fossil or a pseudofossil (that is, something which looks like a fossil but is, in fact, not one), and hence it is a dubiofossil. It is not the case that dubiofossils exist objectively in the world, waiting to be discovered, but rather dubiofossils come into existence on account of the uncertain epistemic state of researchers.

Moreover, something considered a dubiofossil at one time may not be considered a dubiofossil at a later time if new information arises to show that it is, in fact, just a pseudofossil. This does not mean that we were wrong to consider it a dubiofossil in the first place; the term captured the epistemic state of researchers at the time, even though the state changed at a later time. This feature is shared with the epistemic definition of biosignature, whereby the term is used to represent the community's evaluation of the available evidence *at the time*. The epistemic nature of the proposed definition of biosignature, although not as common as ontic definitions, is still in good company with other astrobiology terms such as dubiofossil.

At this point, we can return to the issue that prompted this chapter. The definition of biosignature proposed by Gillen et al. (2023) fits very well into the criteria for the second scenario outlined. The uncertainties of biosignature research and the pragmatic need for a term that communicates the epistemic state in this case provide strong support for the Gillen et al. definition over an objective one. The remainder of this chapter now turns to furthering this argument.

3.5. Why a Doxastic Definition of Biosignature?

There exists a real gap for a term that does the work that the Gillen et al. definition of biosignature does; there is a need for an epistemic definition of biosignature. Ontic definitions of biosignature have been proposed, e.g. Des Marais et al. (2003): "A biosignature is an object, substance, and/or pattern whose origin specifically requires a biological agent" (p.234). This definition is objective and, again, is intuitive as it is in alignment with many objective, ontic definitions in science. However, on account of the high uncertainty surrounding biosignature research, it is not particularly useful.

By taking the Des Marais et al. definition, biosignatures are real-world things that exist whether or not we have detected them – as is the case with stars. However, we can be reasonably certain when we have detected a star, and hence, we can feel justified when we are using the term *star* to refer to some particular spectroscopy data. The same cannot be said for the search for signs of life. Biosignature research is inherently uncertain. The cautionary tales discussed in the previous chapter of this thesis help drive the revisionist history of biosignature claims home. Ultimately, we are working with incomplete information when evaluating whether a signal is a sign of life, and this compromises our ability to make definitive statements. This scenario is captured well by Stanford's problem of unconceived alternatives (Stanford, 2006a; Vickers et al., 2023), and a deeper discussion of this can be found in Chapter Two of this thesis.

There is, therefore, an epistemic distance between our knowledge of the cause of a signal and the actual cause of a signal, and we might anticipate that this distance is even larger than that between legal guilt and factual guilt. To be able to determine definitively that an extraterrestrial signal has a biotic origin, we would need a quantity and quality of data beyond what is often available. It is a seldom occurrence to be able to check these claims in situ (e.g., with sources outside our solar system or at the outer edges), and even those that can be tested in situ (e.g., with some objects within our solar system) face significant uncertainties in the interpretation of the results or concerning possible contamination. The interpretation of the Labeled Release experiment – carried out on the surface of Mars in 1976 – is still debated today and stands as a paradigmatic example of the uncertainties surrounding data garnered from some

astrobiology missions (see Navarro-González et al., 2006; Bianciardi et al., 2012; as well as Chapter Two of this thesis).

Consequently, there are few scenarios where the use of an ontic definition of biosignature is applied correctly. Exceptions, of course, are the uninteresting cases of known products of life on Earth, for example: roads, contrails, and city lights. These biosignatures are biosignatures in a somewhat reductive sense; we are not communicating anything new by branding these as biosignatures. Conversely, astrobiology strives to push the boundaries of what is known, and consequently, many claims will be uncertain. If an ontic definition of biosignature were chosen over an epistemic one, the result would be that, at least for many centuries, it could not be used in the cases in which astrobiology is most interested (e.g. for cases beyond Earth).

Moreover, the ontic definition of biosignature arguably occupies a space that is already filled. The phrases s*ign of life* or *consequence of life* effectively have the same meaning as the Des Marias et al. ontic definition of biosignature. We might ask, therefore, why preserve the term *biosignature* for something with which there is already a phrase? Instead, we should save the term *biosignature* for a gap in our lexicon that has a pragmatic need to be filled.

What the epistemic definition of biosignature offers is a statement of an epistemic position. It sits at the end of a journey of investigation into the possible causes of the signal in question and represents the point where we say: *at present, we can go no further with the available evidence, but our best guess is that this signal is of biotic origin.* As such, a biosignature is not necessarily a sign of life, but it is certainly the next best thing, given the available information. The communication of the culmination of research into the origin of a substance or signal is of great utility to the scientific community and the public. The term *biosignature* should hence be preserved for this purpose, and only the epistemic definition achieves this.

3.6. Potentially Undesirable Consequences of An Epistemic Definition of Biosignature

With a definition of biosignature tied to the scientific community's knowledge and belief set, it is impossible for something to be a biosignature until after it is detected. For example, say there was simple life living on Mars, only it lives deep below the surface, and so we have never detected any sign of it. An ontic definition of biosignature would assert that a plethora of biosignatures exists under the surface of Mars (and likely on the surface), waiting to be detected. In contrast, an epistemic definition of biosignature would only render these phenomena biosignatures when they are detected and evaluated. This might initially seem counterintuitive. But this is only on account of how prevalent ontic definitions are compared

to epistemic ones. There is no practical reason why we should view these signs of Martian life as biosignatures before they are observed. The intuition is only on account of a linguistic norm, but familiarity should not trump utility.

There is a way of conceiving of these yet-to-be-discovered signs of life as *potential* biosignatures. This is to say, they could become biosignatures, were they to be discovered and suitably evaluated. However, an epistemic definition of *potential biosignature* has been proposed by Gillen et al. (2023) as well as in §2.4.1 of this thesis. Here, a potential biosignature is defined as: "any phenomenon for which biological processes are a known possible explanation but whose potential abiotic causes have not yet been reasonably explored and ruled out" (Gillen et al., 2023, p.5; this thesis, §2.4.1). Therefore, at the risk of providing multiple definitions for one term, the Gillen et al. definition of potential biosignature is endorsed here on account of its practicality; by definition, potential biosignatures that just haven't been found yet constitute a set of size zero. There are certainly phenomena, waiting to be discovered, that could one day meet the criteria for biosignature or potential biosignature. However, it would be wrong to call these things biosignatures or potential biosignatures until after they have been discovered and suitably evaluated.

A second, but related, potentially undesirable consequence of the epistemic definition of biosignature is how its status can change without any physical change in the phenomenon. What this means is that, at one time, a detection may genuinely meet the criteria of having a known possible biological origin and its potential abiotic alternative explanations reasonably explored and ruled out, only for it to lose its biosignature status in light of a previously unknown abiotic mimic. In this case, the discovery of a non-biological cause renders the phenomenon in question no longer a biosignature. But crucially, it does not invalidate its biosignature status prior to the new discovery. In such a way, the discovery of new evidence is not retroactive; we would be correct in saying this was a biosignature, but due to new information, it can no longer be called a biosignature. This contrasts with an ontic definition of biosignature, which would claim that we were simply wrong in designating the observation as a biosignature. We thought it was a biosignature, but it never actually was.

Cowie (2023b) picks up on both of these quirks of the epistemic definition of biosignature. To recap, these are: 1) the status of a biosignature depends on whether we know about it, and 2) the status of a biosignature can change, and this change is not retroactive. Cowie (2023b) gives an example to elucidate his argument against the intuitiveness of the epistemic definition: he asks the reader to imagine that planet F has a chemical disequilibrium that is, in actuality, caused by life; however, we do not have the knowledge that this combination of gases is uniquely caused by life. Now, Cowie argues, if we were to look at this planet, the objective account of biosignature would dictate that we are looking at a biosignature, though we do not know it. In contrast, adopting an epistemic definition of biosignature would mean that we are

not looking at a biosignature at all — satisfying the epistemic definition of biosignature requires that observers make an evidence-based conclusion that the only remaining, known, and viable explanation for the phenomenon in question is life. Accidentally observing a consequence of life, but being unaware of this, is therefore insufficient for the epistemic definition of biosignature.

Cowie (2023b) highlights how, on the ontic (though he prefers *objective*) definition, we are able to *discover* biosignatures. They are real things out there, independent of our knowledge of them. Contrastingly, epistemic (or subjective) biosignatures can be created: "So we haven't — or haven't merely — discovered that it is a biosignature. Rather, we have made it into or created it to be a biosignature. This seems wrong. We did not create or cause a biosignature to come into being. It was extraterrestrial life that did *that*! Rather, we discovered that something that was always a biosignature is a biosignature" (Cowie, 2023b, p.17).

Cowie's attack on epistemic definitions of biosignature being able to *create* their subject matter does initially jar with intuition. However, his portrayal of epistemic biosignatures is slightly misguided. Cowie's concern makes the assumption that epistemic definitions actively bring new entities into existence — scientists noticing the chemical disequilibrium on planet F creates the biosignature that is the chemical disequilibrium. However, what is really occurring is that a term, biosignature, is being applied that captures our best knowledge state at a given time. In the same way that a court verdict does not create a crime but rather classifies an event under a legal framework, a biosignature designation classifies a phenomenon on the basis of what is epistemically justifiable at the time.

Notwithstanding this defence against Cowie's critique over creating a biosignature, Cowie's (2023b) arguments do capture an intuition about preferring ontic over epistemic definitions. However, again, this intuition comes only from the high prevalence of ontic definitions in science and elsewhere. The methodological reason for this, as discussed in §3.2, is that science aims at describing an objective world. However, this does not mean there is no room for subjective definitions. In the case of Cowie's planet F, we may well know that a certain chemical disequilibrium is often caused by life. But we have no way of confirming whether *this* chemical disequilibrium is caused by life or by an unknown abiotic mimic. Hence, there is little pragmatic use for an ontic definition here, whereas a strict and well-defined epistemic one (such as the Gillen et al. definition) would be useful.

Ultimately, we want our words to be useful. And hence the charge against an epistemic definition of biosignature of being unusual or even counter-intuitive is not enough to force an unhelpful and unusable ontic definition of biosignature.

3.7. Conclusions

This chapter has defended the Gillen et al. (2023) definition of biosignature against a recent rebuttal in the literature (Cowie, 2023b). This rebuttal challenges the epistemic nature of the definition and argues that an ontic definition should be alternatively adopted. Ontic and epistemic definitions have been defined: ontic definitions refer to phenomena independent of our knowledge or beliefs about them; epistemic/doxastic definitions require a knowledge or belief state in order to be made true.

The prevalence of ontic definitions in science and elsewhere has been highlighted and broadly justified. Science strives to describe a mind-independent world (or, at least, a perspective-independent one). It, therefore, makes sense for the majority of definitions to fall into the ontic camp. However, there is a requirement for some definitions to be epistemic/doxastic. This chapter has proposed two scenarios under which epistemic/doxastic definitions are more appropriate than ontic ones. These scenarios are: 1) there is no objective fact of the matter (e.g. what falls under Hume's definition of beauty) or, 2) there is a high degree of uncertainty when attempting to access the objective fact of the matter, so describing our epistemic stance is the best we can do.

The definition of legal guilt is one such definition that satisfies the second of these scenarios. Legal guilt is an approximation of factual guilt and represents the point where we say: we have gone as far as we can with the available evidence, and this is our epistemic position. There is a real need for a word that describes this belief state. Biosignatures are analogous to legal guilt. With most biosignature research, the actual origin of a phenomenon under question is clouded by uncertainty and blindness to unforeseen abiotic explanations. And yet, the community needs a word to differentiate dubious observations from those that have been heavily scrutinised and have garnered a consensus that they are likely biotic in origin, as per the IPCC framework defended in §2.4.1. of this thesis. The epistemic Gillen et al. (2023) definition provides this word.

Finally, rebuttals to an epistemic definition of biosignature have been raised and addressed. The first of these concerns the requirement that a phenomenon be *observed* in order for it to be a biosignature. Cowie (2023b) refers to this as *creating* a biosignature. I counter this by arguing that we are *categorising*, not creating. The second rebuttal concerns how the discovery of new information can cause a phenomenon that was a biosignature one moment to not be the next. And this demotion to non-biosignature is not retroactive. Broadly speaking, both of these rebuttals make an appeal to intuition and can ultimately be summarised as follows: epistemic definitions are rare and so appear strange.

The response to both of these rebuttals is to admit the unusual nature of an epistemic definition. However, the pragmatic need for a definition of biosignature that captures our knowledge status

outweighs the argument that ontic definitions are more common. A phrase already exists for something that is definitively a consequence of life, but there is a gap and a dire need for what the Gillen et al. definition offers. An epistemic definition of biosignature may be slightly unorthodox, but if we want the words we use to be useful, we need to be open to definitions that encompass our epistemic limitations.

Thus far, Chapters Two and Three have proposed and defended a new definition of biosignature that is strong enough to be useful, whilst responding to the uncertainty over the list of hypotheses that the problem of unconceived alternatives captures. The remainder of this thesis will now focus on how this type of uncertainty, alongside uncertainty over probabilities and payoffs, affects the use of mathematical and computational models in astrobiology. Chapters Four and Five will presently make the case for funding high-uncertainty, high-payoff research in astrobiology (under the framework of maximising expected utility), before evaluating whether SETI falls within this category. Following this, Chapters Six and Seven will consider the value of disagreement in astrobiology. Chapter Six will argue that disagreement over prior probabilities may be at the root of much dispute in astrobiology, before Chapter Seven rounds off the thesis with an analysis of computational community modelling of diversity in astrobiology.

Chapter Four

Risk Vs Uncertainty in Astrobiology: More Than Semantics and Not Pedantic

Risk and uncertainty in decision making and financial markets are clearly defined. These terms are also increasingly being used in astrobiology in the evaluation of projects in funding calls. In recent years, NASA has specifically encouraged the call for higher-risk projects with their Research Opportunities in Space and Earth Sciences (ROSES). This has resulted in the design of projects considered high-risk/high-payoff, for example, the Enceladus Orbilander mission.

However, there has been a mix-up between the use of the words "risk" and "uncertainty". This chapter will discuss this and the significant implications that it could have on the funding of projects with a known low probability of success (risk) when what is desired is an unknown probability of success (uncertainty). The overarching aim is to argue that the misuse of risk and uncertainty is more than a mere semantic issue. The underlying theory of why high-uncertainty/high-payoff projects should be funded will also be explored.

4.1. Introduction

In terms of its wealth of exciting and fruitful research proposals, astrobiology is one of the most fertile fields in science. Astrobiology is a fast-developing field rich in front-page-worthy phenomena waiting to be discovered. And there is no shortage of well-designed project proposals with the potential to make huge breakthroughs. The bottleneck to this torrent of discovery is funding. Selection criteria are, therefore, carefully designed to optimise the output of a field whose potential far exceeds its financial and technological limitations. Typical selection criteria stipulated by funding bodies include the degree of impact that the project could have and the cost of the project (in terms of finances, personnel, equipment,

time, etc). However, recently, there has been interest in a surprising criterion; there has been a call for specifically funding *high-risk/high-payoff* projects.

The National Academies of Sciences, Engineering, and Medicine's (NASEM) 2017 review explicitly makes this call: "Recommendation: NASA needs to investigate appropriate mechanisms to ensure that high-risk/high-payoff fundamental research and advanced technology-development activities receive appropriate consideration during the review process." (NASEM, 2017, p. 31). And, in response to an FAQ asking how NASA has responded to emerging topics and shifting programmatic priorities, a NASA webpage notes that "The Astrobiology Program also partnered with NSF and KnowInnovation to use an alternative solicitation and review mechanism to encourage high risk, high reward proposals" (NASA astrobiology, n.d.).

Moreover, in *An Astrobiology Strategy for the Search for Life in the Universe* (2019) – a detailed strategy that outlines central questions in astrobiology and highlights promising research areas – the growing interest in high-risk/high-payoff projects is noted: "One high-risk/high-payoff area for which philanthropic and, increasingly, international investments have entirely supported the search for life is the search for technosignatures, or the signature of technologically advanced life." (NASEM, 2019, p.6).

NASA's interest in funding high-risk/high-payoff projects is also seen in the 2017 Research Opportunities in Space and Earth Sciences (ROSES) funding competition. Submissions were assessed, in part, on their level of risk, with the flagging of those considered high-risk/high-impact. The risk associated with a proposal is defined by the ROSES funding competition as "to what extent would this proposal test novel or significant hypotheses, for which there is scant precedent or preliminary data or could run counter to the existing scientific consensus?" (NASEM, 2017) — this definition will be returned to and unpicked in §4.4 of this chapter.

Of those flagged as high-risk/high-impact, 35% were selected for funding, compared to an overall selection rate of 24% (Vickers, 2020). This is coarse data with both high-risk and high-impact being clumped together and, without finer data on how risk individually impacted the selection, definitive conclusions cannot be drawn about risk. For example, the disproportionate funding of high-risk/high-impact projects may reasonably be attributed to the high-impact alone. However, at the very least, high-risk does not seem to be detrimental to the selection of a project, as intuition might suggest.

In a recently published paper, we suggest that high-risk research projects in astrobiology are disproportionately high-impact (Jeancolas et al., 2024a). By carrying out interviews with the authors of the most cited astrobiology papers of the past 20 years, we argue that "the majority of the selected

breakthrough results derive from endeavours considered medium- or high-risk, risk is significantly correlated with impact, and most of the discussed projects adopt exploratory approaches" (Jeancolas et al., 2024a, p.1). These findings might suggest that the funding of high-risk research in astrobiology is proving beneficial. However, although not challenging the substance of Jeancolas et al. (2024a), I make the case in this chapter for clearly dividing high-risk projects from high-uncertainty ones. As such, I show that the findings of Jeancolas et al. (2024a) are consistent with the call for high-uncertainty projects in astrobiology, rather than high-risk ones.

Since the call for more high-risk/high-payoff research in astrobiology, swarms of exciting and unorthodox projects have been funded. The recent work from the SETI Institute is one clear example, as is the planned Enceladus Orbilander, which will sample the water plumes of this moon of Saturn before landing to analyse materials for evidence of life (MacKenzie et al., 2020). It seems that the direction of astrobiology has been nudged by the recent interest in high-risk projects. But caution is due over what we are asking for here. Just what do these calls for high-risk really mean? Is it risk we should be focussing on or something else? And what is the theoretical backdrop for why it might be rational to fund unorthodox projects?

This chapter, therefore, argues that there has been a mix-up between the words uncertainty and risk in some funding calls in astrobiology. Although this may initially seem like an innocuous substitution of one word for another very similar one, the implications for optimising the output of astrobiology are significant. It is not rational to fund high-risk projects, though it may be rational to fund high-uncertainty ones. The chapter takes the following form: §4.2 discusses the distinction between risk and uncertainty. First, this is done in decision theory generally and then in financial markets and medicine. §4.3 then delves into the distinction between risk and uncertainty in projects within astrobiology by considering real-world examples of each. §4.4 contains the key argument of the chapter in making the case for funding high-uncertainty projects in astrobiology and explicitly *not* funding high-risk ones. A theoretical explanation of this is also provided. §4.5 acknowledges a potential area of exploitation in calling for high-uncertainty projects over high-risk ones and proposes a way to mitigate this. Finally, §4.6 concludes the arguments made in the chapter.

4.2. The Distinction Between Risk and Uncertainty

Risk and uncertainty are clearly defined in decision theory and game theory. However, the distinction is not restricted to the theoretical arena. Pinning down risk and uncertainty is essential also in practical fields

including behavioural economics, financial markets more generally, and even medicine. It is not an exaggeration to say that human life and the economy have hinged on this distinction. What, then, is this crucial difference between risk and uncertainty?

4.2.1. Risk and Uncertainty in Rational Decision Making

A game of Russian roulette, where it is known how many bullets are in the revolver, offers a paradigmatic example of decision making under risk. Specifically, the risk concerns the probabilities. Let us say, it is known that there is one bullet in the gun. Imagine now a situation where person X is considering taking a spin on Russian roulette. They are able to make an informed decision on the probability of getting the bullet – one in six – as the probabilities are *known*. This would be a decision under *risk*. Consider now the case where the number of bullets in the revolver is *unknown*. Person X is simply unable to weigh up the risks and would hence be making a decision under *uncertainty*. In summary: when the probabilities of an outcome are known, the decision is made under risk. When the probabilities are unknown, the decision is made under uncertainty.

Orthodox rational choice theory hinges on this distinction largely due to the central role of expected utility theory. Formalised in the mid-20th century by the fathers of game theory, John von Neumann and Oskar Morgenstern (1953), expected utility theory offers a normative framework for optimising the outcome of any particular choice under risk. Expected utility theory emerges from the rational choice preference axioms of continuity and transitivity, alongside the axiom of independence of irrelevant alternatives. Resultantly, rationality is ensured (von Neumann & Morgenstern, 1953). The theory prescribes the following: when confronted with a decision under risk, a rational agent should opt for the outcome which maximises their expected utility. This is to say that, given a pair of lotteries $M = [A_1, A_2 \ldots . A_n]$ and $L = [B_1, B_2 \ldots . B_n]$ with corresponding probabilities $p_{Li=}[0,1]$ and $p_{Mi=}[0,1]$, a rational agent should select the lottery that gives the largest expected utility, given by:

$$E(U) = \sum_{i=1}^{n} A_i p_i$$

Where A_i represent the outcomes within the lottery and p_i represent the probabilities corresponding to the lottery outcomes.

Let us consider an example application of expected utility theory. Consider the choice between the two following lotteries.

Lottery 1:

0.9 probability of winning £0

0.1 probability of winning £1000

Lottery 2:

1.0 probability of winning £100

An individual considering which of these lotteries to take should employ expected utility theory to optimise their outcome. In this case, the expected utility of lottery 1 would be:

$$E(U)_1 = \sum_{i=1}^{n} A_i p_i = £0 \times 0.9 + £1000 \times 0.1 = £100$$

And the expected utility of lottery 2 would be:

$$E(U)_1 = \sum_{i=1}^{n} A_i p_i = £100 \times 1.0 = £100$$

To make a rational choice, our individual should choose the option with the highest expected utility. In this case, lottery 1 and lottery 2 have the same expected utility – £100 – and hence they should be indifferent between the two choices.

As a side, it is well documented that people do not actually act in compliance with expected utility theory and instead apply a weighting function to probabilities which over-weigh low probabilities and under-weigh high ones (Kahneman & Tversky, 1979). People are also risk-averse when it comes to gains and risk-loving when it comes to losses (Kahneman, 2011). As such, when it comes to gaining money, people want the certainty. But if they have to lose, people a prepared to take the risk of a gamble.

The modelling of this descriptively accurate account of decision making is a fundamental feature of Daniel Kahneman and Amos Tversky's Nobel prize winning prospect theory (1979). When individuals are asked which of the above lotteries they would prefer, they are usually not indifferent (as expected utility would prescribe) and would rather take the sure bet of winning £100 (Kahneman & Tversky, 1979, p.266). This behaviour is explained nicely by Kahneman and Tversky's risk aversion to gains.

Nonetheless, expected utility theory is a *normative* account that can be utilised when making decisions under risk, i.e., when the probabilities of the possible outcomes are known.

In contrast to the above example of decision making under risk, consider the choice between the following lotteries:

Lottery 1:

Unknown probability of winning £0

Unknown probability of winning £1000

Lottery 2:

Unknown probability of winning £100

This would be an example of decision making under uncertainty and, due to the lack of known probabilities, expected utility theory can do nothing to advise. Indeed, there are few, if any, prescriptive accounts for decision making under uncertainty, with the exception of first attempting to narrow down the uncertainty! There may also be in-between cases such as decision making under partial uncertainty. Such cases might include betting on horse racing. The bookie provides odds for each horse, and these odds can be used to inform the probability of each horse winning. However, the odds have arisen from informed speculation and hence are only estimates. The bookie might, in actuality, ascribe a range of probabilities (say, between 10% and 20%) for which he is reasonably confident represents the probability of a particular horse winning the race. Horse betting thus provides an example of decision making under partial uncertainty.

Consequently, the implications of risk and uncertainty are not confined to theoretical lotteries. Distinguishing between decision making in which the probabilities for each potential outcome are known and in which the probabilities are unknown has great importance to real world scenarios.

4.2.2. Applications of Risk and Uncertainty in Financial Markets

Investors are very concerned with risk. Consider a company where, in a particular year, stocks may be expected to increase in value by 5%. This figure is often arrived at by taking historical data – the company may have lost 10% one year but gained 20% the next, such that the average growth is 5%. Additional

features, like information about global market trends, may also help inform the decision on expected interest. What this 5% growth means is that the expected utility of a £100 investment would be £105.

There may indeed be some uncertainty about this risk assignment (as with horse racing odds). But the attempts by investors to pin down these probabilities show a preference for risk over uncertainty. Of course, it is not always the case that the expected utility materialises. An investor may indeed lose the whole £100. This does not mean that the risk assignment was false and that we were really making a decision under uncertainty. Unlikely outcomes happen all the time. But, if the investor were to invest in a variety of companies, their *average* earnings should comply with the average projected interest rate. This is analogous to flipping a coin; you would not expect the coin to always land end on, between heads and tails, but after 100 flips, the average outcome becomes clear.

Whereas risk makes up the backbone of financial markets, uncertainty is somewhat shunned. To invest in an uncertain investment means that the probability of each outcome is unknown. It is a truly blind investment, and there is no data to suggest what your investment will tend to over time. Uncertainty in financial markets is indicative of more research needing to be done to change that uncertainty into risk, and much research focuses on how best to mitigate this uncertainty (e.g., Rigotti & Shannon, 2005).

4.2.3. Applications of Risk and Uncertainty in Medicine

Risk and uncertainty are central concepts in the decision making that underpins medical diagnoses and prescribed courses of action. For instance, a doctor may appeal to data on past sufferers of heart disease to inform her diagnosis of a current patient. Her patient is a smoker, has high blood pressure, and is overweight: three key risk factors for heart disease. Hence the doctor increases her estimated probability that her patient falls under this diagnosis. When considering which course of treatment to prescribe, the doctor discusses the side effects with the patient. Statin, a cholesterol-lowering drug, lists a range of common, uncommon, and rare side effects. Each of these will have resulted from medical trials whereby, for example, myopathy was reported in only 0.1% of participants and is therefore classified as rare. In summary: statistical risks honed over time are central to diagnoses and prescriptions in medicine.

Policy choice, too, is often informed by medical risk and an interesting example of this is found in the COVID-19 pandemic. In spring 2021, the UK vaccination programme was in full swing, with vaccines (including the Oxford/AstraZeneca) being administered to all willing adults (Office for Budget Responsibility, 2021). However, in May 2021, the Joint Committee on Vaccination and Immunisation published a press release advising people under 40 not to take the Oxford/AstraZeneca vaccine and to instead take an alternative COVID-19 vaccine. The reasoning for this came from the very small number

of concurrent thrombosis (blood clots) and thrombocytopenia (low platelet count) cases reported following the first dose of the Oxford/AstraZeneca vaccine. However, this was not new information. The risks for blood clotting and low platelet count were known prior to the policy change. So, what spurred this reversal?

The change was in the *comparative* risk. The risks for blood clotting and low platelet count remained incredibly low – by the end of April 2021 the incident rate of thromboembolic events after a first or unknown number of doses was at 10.5 cases per million doses (Public Health England, 2021). A risk calculation was made in early spring, when COVID-19 cases were incredibly high, that the risk of the vaccine was lower than the risk of getting the virus and being harmed by it. However, come May, COVID-19 cases had fallen, and so, although the risk from the vaccine and the risk from having COVID-19 had remained the same, the risk from being an unvaccinated young person was lower than the risk of taking the vaccine because you were simply less likely to get the virus in the first place. Hence, the relative risks shifted, and it was no longer considered individually beneficial for someone under 40 to take the Oxford/AstraZeneca vaccine.

It goes without saying that a proper grasp of risk in medicine is crucial for good public health and policy. It was because the researchers had the data to inform the risk assessment that they were able to advise. It is true that the risks could not be wholly constrained, and policymakers and health professionals had to wrangle with some degree of uncertainty. But, if instead, there was total uncertainty about the risks of the COVID-19 virus and vaccine, it would not have been possible to respond to the changing situation in an informed, rational way.

4.3. Risk and Uncertainty in Astrobiology

The above examples elucidate the role of risk and uncertainty in decision theory, financial markets, and medicine. This section will now turn to their presence in astrobiology. Analogous to how risk is defined in decision theory, high-risk, high-payoff projects in astrobiology should be defined as projects for which there is a known, low probability of the desired outcome actualising, but the payoff of the desired outcome is high. For example, sending a rover to specifically look for extant life on the moon would be considered high-risk on account of what is already known about the moon. The lack of atmosphere, alongside the moon's distinct lack of elements considered essential for life (e.g. carbon and nitrogen) (Wordsworth, 2016; Westall et al., 2013) means there is good reason to believe the probability of life on the moon is very small. In contrast, a project would be considered high-uncertainty, high-payoff if the

probability of success is unknown. Such a project might include directing the James Webb Space Telescope at a red supergiant right at the end of its life and hoping to catch it as it goes supernova. Watching the forging of all the heavier elements would be of high value, but we might struggle to predict when, exactly, the star will die.

4.3.1. The Enceladus Orbilander

In practice, projects in astrobiology usually comprise a suite of subprojects with varying degrees of risk and/or uncertainty in probabilities. Funders must, therefore, consider the entire set of subprojects when determining the overall promise of the project. The Enceladus Orbilander mission, planned for launch in 2038, is a great example of such a project comprised of subprojects with varying degrees of risk and uncertainty. In brief, the Orbilander is designed to have an orbital phase around Enceladus and then subsequently land on this moon of Saturn. It will orbit Enceladus for a year and a half, sampling the water plumes as it does so. Following this, it will spend two years on the surface of the moon, studying Enceladus's surface materials for evidence of life (MacKenzie et al., 2020).

The Enceladus Orbilander's primary goal is to search for life in the subsurface ocean of this exciting moon of Saturn. Enceladus is now widely considered habitable; the Cassini mission identified Enceladus as an active moon with regular plumes and a liquid subsurface ocean, all of which suggest Enceladus's habitability (McKay et al., 2014). Furthermore, a recent analysis of the Cassini data suggests that phosphates (considered essential for life) might be relatively abundant in Enceladus' oceans (Postberg et al., 2023). As such, the excitement about the proposed Orbilander mission is palpable.

This primary goal in searching for life might be considered highly uncertain. As mentioned, there is good reason to think Enceladus is *habitable,* but this is far from saying it is *inhabited.* The fundamental probability of abiogenesis is just too poorly defined to comment on the probability of Enceladus actually harbouring life. Chapter five of this thesis provides a more in-depth discussion of the difficulty in defining a probability for abiogenesis.

Despite the uncertainty associated with searching for life, the Orbilander mission includes several sub-projects with better-defined probabilities of success. One of these includes a microscope and nanopore sequencer. The authors of a paper discussing the balance of return and resources state, "We also considered the likelihood of success with the high-risk, high-reward microscope and nanopore sequencer. In addition to potentially providing unambiguous evidence of life, both systems could also reveal specific aspects of the nature of detected lifeforms (high reward). However, the likelihood of a positive result from these systems was considered lower than for the rest of the life detection payload (high risk)"

(MacKenzie et al., 2021, p.5). Due to the known technological limitations of the equipment, this particular part of the overall project has been identified as high-risk, that is, a known low probability of success.

Finally, an aspect of the Enceladus Orbilander mission that might be considered low risk would be simply characterising the composition of the moon's plumes and subsurface ocean. As there would be value in the detection of any particular composition, the only way this goal could fail would be if there were a technical fault. So long as the instruments run correctly, this goal has a high probability of succeeding. The mixed suite of high-uncertainty, high-risk, and low-risk projects within the Enceladus Orbilander presents a great example of the differences between risk and uncertainty in astrobiology.

4.3.2. The Detection of the First Exoplanet Orbiting a Main Sequence Star

An example of a high-risk project that actually paid off was the detection of the first exoplanet orbiting a main sequence star by Nobel laureates Mayor and Queloz in 1995. The planet, 51 Pegasi b, is a gas giant that falls under the description of a 'hot Jupiter' and orbits so close to its host star that a year on 51 Pegasi b lasts four Earth days. The existence of exoplanets was known prior to Mayor and Queloz's discovery; in 1992, Wolszczan and Frail detected two rocky planets in orbit around a pulsar in the constellation Virgo (Wolszczan & Frail, 1992). But Mayor and Queloz's discovery was the first to bring hope for other potentially habitable worlds orbiting sun-like stars.

The search for 51 Pegasi b was a high-risk pursuit on account of where the planet was detected. The mode of detection used the radial velocity method which looked for a wobble in the star caused by the gravitational tugging of the planet as it orbited with a small radius. Such a method is only effective at detecting large planets at close distances to their star (for the gravitational force to be strong enough to noticeably wobble the star).

However, the feasibility of such large, close-orbiting planets was highly disputed at the time. The existing theories of planet formation were based on our solar system, where the gas giants occupy the outer solar system. Mayor and Queloz's detection method would only work if they found a large, close-orbiting planet. But the existence of such a planet was considered unlikely.

The community thought they had reason to think that the probability of detecting a planet so close to its host star was very small. And so, the exploration would have been considered high-risk at the time. Nonetheless, the project paid off. So, does this case study mean we should expect high-risk projects to pay off?

It is true that, *sometimes*, high-risk projects will pay off, in spite of their justifiably low probability of success. Eventually, a coin might land heads 100 times in a row; someone will win the lottery; and lightning occasionally strikes twice. However, in the case of Mayor and Queloz's detection of 51 Pegasi b, they did not spot the exoplanet in spite of a known low probability of doing so. They spotted it when, in actuality, there was quite a high probability of them doing so. This is because their ascribed low probability of success came from a now false theory of planet formation. Consequently, the detection of 51 Pegasi b resulted in a reworking of planetary formation theory that has large planets forming far from their star and migrating in over millions of years (Lin et al., 1996).

The detection of 51 Pegasi b offers an interesting example of how human fallibility can play into the determination of risk itself. The existing incorrect theories about planet formation rendered the search high-risk. In reality, although their incidence is rare at ~ 0.9 – 1.2% (e.g., Mayor et al., 2011; Wright et al., 2012), they are still vast in number and due to their ease of detection, they make up ~20% of all detected exoplanets to date (Wright et al., 2012).

Nonetheless, scientists make use of the data they have at the time to assess a risk value, and so what was genuinely considered high-risk before the detection of 51 Pegasi b would now be considered low-risk. Indeed, if we are correct in our assignments of risk values, we should, by definition, expect high-risk projects to fail most of the time.

4.3.3. Table Categorising High-Risk and High-Uncertainty Projects in Astrobiology

Table 4.1 gives an overview of a selection of projects considered either high-risk, high-uncertainty, or somewhere in between, alongside the outcome of each project. This in between classification refers to projects in which the probabilities of success are reasonably uncertain, though it is estimated that they are low.

It is worth considering the publishing bias when looking at the outcome of past projects. Null results are less likely to get published, so there is a bias in the literature towards those projects that succeeded (Kepes et al., 2014). Moreover, Table 4.1 is a mere selection of the incredibly rich set of missions in astrobiology, starting from Lederberg initiating a research programme within NASA to explore the distribution of life in the cosmos (Lederberg, 1960). It is, therefore, not possible to draw firm conclusions from the examples in the table. Nonetheless, the trend shown in the table of high-risk projects usually failing, and high-uncertainty projects sometimes failing and sometimes not, is what would be expected, given our definitions of high-risk and high-uncertainty.

Project	High-Risk? High-Uncertainty? Or Somewhere in Between?	Success?
Detection of the first exoplanet orbiting a sun-like star (Mayor & Queloz, 1995).	High-risk. Such large and close-orbiting planets were believed to be very unlikely at the time. However, this reasoning was later found to be false.	Yes – the first exoplanet orbiting a sun-like star was detected.
Testing samples from the Apollo 11 mission for signs of life on the moon.	High-risk. It was highly suspected that the moon was uninhabitable. Testing the samples, however, was relatively low cost.	No – no signs of life found.
Testing samples from the Apollo mission for water on the moon.	High-risk. There was already a consensus at the time that the moon likely did not contain water.	No – no signs of water were found at that time.
JWST detection of the oldest found supermassive black hole (e.g. Kocevski et al., 2023; Witze, 2023).	High-Uncertainty. Before this detection, the prevalence of relatively low-mass, old supermassive black holes was not confirmed. Hence, it was difficult to put a probability on the likelihood of detecting one. Now, they are thought to be abundant.	Yes – the existence of these black holes is now confirmed, and valuable data are being collected.
JWST analysis of exoplanet VHS 1256b, a super Jupiter, 40 light years away (Miles et al., 2023).	High-uncertainty. Previous observations by the Vista Telescope in Chile showed the planet to appear red, thus *hinting* that there *might* be dust in its atmosphere. However, the extent and nature of this were unknown.	Yes. Sandstorms have been detected on this exoplanet alongside water, methane, carbon monoxide, and carbon dioxide — the largest number of molecules found on an exoplanet (Miles et al., 2023).
1976 Viking Labeled Release experiment searching for life on Mars.	Somewhere in between. The previous habitability of Mars was hotly debated (Averner & MacElroy, 1976; Ponnamperuma & Klein, 1970), and there existed concerns over the sensitivity of the	Contentious. The interpretation of the LR experiment is still not unanimous (e.g., Bianciardi

	Viking equipment to detect signs of life (Navarro-González, 2006).	et al., 2012; Navarro-Gonzalez et al., 2006).

Table 4.1. selection of notable discoveries in astrobiology and their probability risk or uncertainty classification.

4.4. Prioritising High-Uncertainty Over High-Risk Projects

High-risk projects, by definition, fail more often. In contrast, high-uncertainty projects are a mixed bag. The thrust of the remainder of this chapter is to argue that funding bodies should be calling for high-uncertainty projects and certainly not calling for high-risk ones. Moreover, the distinction between these categories should be made explicit to prevent high-risk projects from sneaking in under the label of high-uncertainty. The theoretical foundation for funding high-uncertainty projects will also be outlined.

As a reminder of the ROSES definition of high-risk projects, funding selectors are encouraged to ask: "to what extent would this proposal test novel or significant hypotheses, for which there is *scant precedent or preliminary data* or could *run counter to the existing scientific consensus*?" (NASEM, 2017, emphasis added). With the distinctions between risk and uncertainty presented so far in this chapter, an ambiguity within this definition of high-risk/high-payoff becomes apparent. The ROSES definition of risk here is a combination of both high-risk *and* high-uncertainty. The requirement for "scant precedent or preliminary data" indicates a project residing within an unexplored epistemic space with little data to help constrain the probability of success — a high-uncertainty project. Conversely, the requirement for a project which runs "counter to the existing scientific consensus" indicates a high-risk project where there are indeed data to provide probabilities of success, and success is expected to be unlikely.

The overly broad nature of the ROSES funding competition definition of risk has also been discussed in Vickers (2020). Here, Vickers points out that many projects which fall under the category of scant evidence or preliminary data should not be considered to have a known low probability of producing high-impact results — the examples of the WFIRST microlensing survey and the Rosetta mission to land a module on a comet are given (Vickers, 2020, p.489). Resultingly, Vickers (2020) recommends that future funding calls comparable to the ROSES call should define risk as "when the probability of a significant scientific return is significantly below average" (Vickers, 2020, p.489).

Although I agree wholeheartedly with Vickers' (2020) definition of risk and suggestion that such a definition should be precisely employed when referring to risk, I aim to make the case in the remainder

of this chapter that funding bodies should *not* be calling for more high-risk research but rather for more high-uncertainty research.

4.4.1. The Low-Risk, High-Payoff Jackpot

Resources are tight in astrobiology, so in order to maximise the output of the field, the projects with the highest expected utility should be funded over others. In a white paper in support of the *Planetary Science Decadal Survey 2013-2022,* Schingler et al. (2009) present a mission evaluation metric to guide and optimise mission selection. They propose the following mission evaluation metric, *V*:

$$V = S \times P_S / \$$$

where *S* is the potential scientific value of the mission, P_S is the probability of success for the mission, and *$* is the mission cost. This metric takes into account cost and so would prioritise a low-cost mission over a high-cost one with identical risk and value inputs. This is fully consistent with expected utility, only in Schingler et al.'s metric, they take the value to be inversely proportional to mission cost, whereas in expected utility, it would make more sense to subtract the cost from the expected utility. This distinction is minor, and either the mission evaluation metric (2009) or simply employing expected utility would be highly effective for comparing competing missions. For completeness, the formula for employing expected utility for one outcome of interest, accounting for mission costs, is:

$$E(U) = (A \times P) - \$$$

where *E(U)* is the expected utility, *A* is the payoff of the outcome of interest, *P* is the probability of the outcome of interest materialising, and *$* is the mission cost.

An argument can certainly be made that we should, therefore, solely fund low-risk, high-payoff projects. Indeed, there are so many of these excellent candidate projects out there that the whole budget could be spent on them, and a significant scientific output would be reliably garnered. This is discussed in Vicker's *Expecting the Unexpected in the Search for Extraterrestrial Life* (2020). Here, Vickers clearly states the disanalogy with financial markets where high-payoff is usually tied to high-risk. Examples of these prolific low-risk, high-payoff projects in astrobiology include the LUVOIR telescope concept, which could reliably image exoplanets. Another such example is tasking JWST with analysing a range of exciting objects like sets of planets known as hot Jupiters and mini-Neptunes, the TRAPPIST-1 system, and K2-22b – a rocky planet that is currently disintegrating around its host star (Vickers, 2020, pp.17-18).

All that has been discussed thus far would agree with this. The most rational decision seems to be to exclusively fund these low-risk, high-payoff projects. So why pay any attention to high-risk or high-uncertainty projects? The only reason why the funding of these unlikely or unpredictable projects might be rational is if the payoffs far exceed those of the low-risk, high-payoff projects; that is to say, a project is high-risk or high-uncertainty but has an *astronomically* high payoff. The question thus arises: is there reason to believe this would be the case?

4.4.2. High-Uncertainty, Astronomical Payoff

There may be theoretical reasons to believe that, lurking within the most uncertain projects, lie truly revolutionary outcomes. The field of astrobiology is relatively new; unexplored physical and theoretical spaces remain rich. We could indeed restrict ourselves to our small area of known phenomena and continue to mine this for exciting breakthroughs that gradually expand the boundaries of what we know. We could opt to ignore the mysterious, poorly understood, and so poorly quantified area outside what our theories can provide probabilities for. But, by avoiding this uncertain area, we may be missing discoveries that upend our current framework. Low-risk, high-payoff projects may be highly beneficial in bolstering the existing framework, but high-uncertainty projects may rewrite the framework itself.

Surely, the highest payoff is that which tells us something completely new. Projects with well-defined, high probabilities of success (low risk) are likely to sit within a theoretical space that is already well-explored. This does not mean they cannot yield huge payoffs – analysing the disintegrating planet K2-22b would undoubtedly produce valuable data. However, projects whose probabilities are difficult to determine are likely so because they sit in a theoretical space yet to be explored. It is these high-uncertainty projects that can shake the operating framework and introduce new fruitful lines of research. As such, there is a theoretical motivation for funding high-uncertainty projects alongside low-risk, high-payoff projects.

Broad support for the link between uncertainty and major scientific breakthroughs can be found in the 'science of science' literature. These findings suggest that exploratory research which ventures into unexplored areas tends to produce disproportionately high-impact discoveries (Fortunato et al., 2018). This link is further supported in the paper "Bias against novelty in science: A cautionary tale for the users of bibliometric indicator" (Wang et al., 2017). The opening sentence of the abstract states that: "Research which explored unchartered waters has a high potential for major impact but also carries a higher uncertainty of having impact" (Wang et al., 2017, p.1416). The paper goes on to present data regarding the relationship between the impact of published papers and what is called *novelty*. Novelty is defined via

an analysis of the bibliography of each research paper. Papers that contain an unusual combination of references would be regarded as highly novel. For example, a paper whose references span from behavioural economics to marine biology to medieval history would score very highly on the proposed novelty index. The rationale is that unusual fusions of previous research are representative of a less explored research area.

I am inclined to say that Wang et al.'s novelty index tracks with uncertainty, as opposed to risk. This is because a project that spans previously ill-connected disciplines is feasibly an unusual project whereby its probability of success is difficult to pin down. Different disciplines are being fused in unique ways, and hence the project sits in a somewhat unknown space. This contrasts with high-risk projects, whereby we feel confident in assigning a low probability of that project succeeding.

What Wang et al. (2017) find is that papers defined as highly novel are more likely, in the long run, to be in the top 1% of highly cited papers. They also find that the variance of the most novel research is higher than for research that is low in the novelty index. However, these findings are stated as so: "We find novel papers to have a larger variance in their citation distribution and be more likely to populate both the tail of high impact and the tail of low impact, reflecting their 'high risk' profile" (p.1417) and "While novel research faces a higher level of risk, we also expect novel research to have a higher probability of making a significant contribution to research" (p.1420) and, finally, "novel papers, in particular highly novel papers, exhibit citation patterns consistent with the 'high risk/high gain' profile associated with breakthrough research… they have a significantly higher chance of being top 1% highly cited" (p.1424).

With the definitions of risk and uncertainty used in this chapter, these conclusions are a little confusing. If a paper is deemed high-risk, it is, by definition, unlikely to yield its potential impact. We should expect the highest-risk papers to, on average, fail. However, the comments about high variance and unpredictability about impact are exactly what would be expected for high-*uncertainty* papers. This distinction is an important one, as Wang et al. go on to call for less risk-averse behaviour on the part of funding agencies. The research and theory, though, suggest that high-risk projects should, by and large, not be funded, though perhaps high-uncertainty projects should. It is important to distinguish these so as to prevent projects with a known low probability of success from being smuggled in alongside high-uncertainty ones.

This approach of funding high-*uncertainty* projects is consistent with the findings of "Breakthrough results in astrobiology: is 'high risk' research needed?" (Jeancolas et al., 2024a). By interviewing the authors of the most impactful papers in astrobiology of the past 20 years, we find that "most of the discussed projects adopt exploratory approaches" (ibid, p.1). We take this exploratory approach, whereby the

potential payoffs and likely probabilities of success are quite unknown and categorise it as a facet of risk. As such, a central conclusion of the paper is that "high-risk research seems to be a key lever to generate breakthrough results" (ibid, p.17). I do not wish to push back on the substance of this conclusion; we should expect highly exploratory (uncertain) projects to generate high-impact results in astrobiology for the reasons argued in this chapter. However, I would stress the utility of separating our language when dealing with risk and uncertainty. As such, if the reader is inclined to make this distinction in vocabulary, they would join me in interpreting the results of Jeancolas et al. (2024a) as showing how high-*uncertainty* projects (rather than high-risk projects) are a key lever to generating breakthrough results in astrobiology.

Finally, a call for funding more high-*uncertainty* projects is also consistent with the ongoing call for astrobiology to 'expect the unexpected' (NASEM, 2018, p.10) insofar as we should expect there to be huge dark areas in our knowledge, and the stumbling across unforeseen phenomena should be anticipated. This should be distinguished from expecting the probably-not-there. If we really do have good reason to believe that a project will likely fail, then other projects should be preferred.

It is true that sometimes our modes of assessing probabilities are false as our theories are still being developed (as was the case with Mayor and Queloz's exoplanet detection). But a case could be made that these examples have been falsely characterised as high-risk when we should have been more uncertain about the probabilities. It is only with hindsight that we can see cases were falsely characterised as high-risk due to the employment of an incorrect theory. So, it is difficult to say that all high-risk projects are genuinely high-risk. However, it is expected that projects with huge, revolutionary outcomes track better with high-uncertainty than high-risk on account of the target area being already better explored in the latter compared to the former.

4.5. Avoiding Exploitation of the Call for High-Uncertainty Projects

Embracing high uncertainty is generally not something encouraged in other fields. Normative decision theory, for example, stipulates that when faced with high uncertainty, an agent should first try to gain more information such that they can narrow the uncertainty into known risks before acting; there exist few models that deal with rational decision making under high uncertainty (Kochenderfer, 2015). However, in astrobiology, sometimes it is simply not possible to constrain our uncertainties into risks on account of how vast and relatively unexplored the field is. Sometimes, it is hard to even conceive of what the possible outcomes of a project could be, let alone ascribe tight probabilities to these outcomes. And this is just a fact of many projects in astrobiology at present.

Having said this, for maximising the output of the field, it is important that projects which should be classified as high-risk are not falsely presented as high-uncertainty so as to gain funding from organisations that encourage the expecting of the unexpected. Funding calls for high-uncertainty projects should guard against projects where the probabilities have been intentionally not pinned down. It is a well-known feature of research funding structures that scientists actively brand their projects in alignment with the specifications of the funding call. This is, of course, inevitable and by no means itself a negative thing.

However, given a call for high-uncertainty projects, it is conceivable that researchers may tailor their applications by simply leaving out known low probabilities and instead presenting the likelihood of outcomes as uncertain. They may hold off on doing preliminary research into, for example, the likelihood of a particular instrument working under certain weather conditions. Probabilities that can be constrained should be. It is the genuinely uncertain projects that might represent a poorly explored research area that are to be endorsed under the reasoning in this chapter. Resultingly, calls for high-uncertainty research should include a requirement that the uncertainty be justified as coming as a direct result of the field being highly exploratory and new. Any uncertainties that could be converted into known risks should be addressed.

4.6. Conclusions

The language surrounding the recent calls for more high-risk/high-payoff research is ambiguous and misleading when high-risk and high-uncertainty are used interchangeably. This is more than a pedantic observation; the confusion can lead to a non-optimal output for the field of astrobiology. There exist reasons to believe there is genuine merit in funding high-uncertainty projects, but funding high-risk projects should be strictly detrimental over time.

By adopting the framework of expected utility, this chapter has argued for the rationality of sometimes funding high-uncertainty projects in astrobiology. This is an approach that would not be considered normative in other fields, such as finance and medicine; it is the highly unexplored nature of astrobiology that uniquely justifies this approach. In contrast, the irrationality of funding genuinely high-risk projects has been laid out. In a field rich in low-risk/high-payoff and high-uncertainty/high-payoff projects, the projects with a known low probability of success should be bottom rung.

Finally, some of the practical implications of encouraging high-uncertainty projects have been discussed. Calls for high-uncertainty projects could mitigate efforts to masquerade high-risk projects as high-uncertainty by requiring applications to highlight how attempts have been made to define probabilities wherever possible. Applications should justify the uncertainty in their project as originating from a field that is largely unexplored, and reviewers should keep a keen eye out for projects with genuinely undefinable probabilities, as opposed to optionally undefined probabilities. The following chapter will now take this endorsement for the funding of high-uncertainty, high-payoff projects in astrobiology and apply it to a contentious area of research. Namely, the question of whether SETI falls within this high-uncertainty, high-payoff camp will now be evaluated in light of the US Congress's historical decline to offer public funding to SETI (Garber, 1999).

Chapter Five

Is SETI High-Uncertainty, High-Payoff Research?

The previous chapter made the case for funding high-uncertainty projects, which are defined as projects with poorly defined probabilities of success or unknown payoffs. This chapter takes the case study of the highly speculative search for extraterrestrial intelligence (SETI) and asks whether it falls within the camp of high-uncertainty projects. This is a notable example as SETI is currently not funded by NASA and is rather funded by private benefactors. If indeed an argument can be made that SETI is highly uncertain exploratory research, do the arguments of the previous chapter mean it should receive public funds?

This chapter first makes the case for SETI being high-uncertainty research. This is done by utilising the famous Drake equation as a framework for identifying the level of uncertainty associated with the detection of extraterrestrial intelligence. On our path to ascertaining the prevalence of extraterrestrial intelligence for which we can communicate, we hit up against questions like how likely is it for life to emerge on a particular planet? And, given this abiogenesis, how likely is it for life to become intelligent, or even technologically advanced? In this chapter, I move through each of the seven parameters of the Drake equation and review literature on the pinning down of these parameters, finding that several suffer from the N = 1 problem. I conclude that, indeed, SETI research is high-uncertainty research.

So, should NASA fund SETI? Having established SETI as high-uncertainty research, I turn to the potential payoffs associated with the search for extraterrestrial intelligence. A project should not be funded solely because it is high-uncertainty. It should also be high-payoff (or at least not have a negative payoff). I will discuss the state of public opinion on the matter of SETI's potential payoffs and will conclude that this matter is not settled. SETI must compete with other high-uncertainty projects, many of which will be high-payoff. And as such, there are innumerable research projects that should be ahead in the funding line. The chapter concludes that SETI research does not yet fall into the camp of high-uncertainty, high-payoff research that has been defended in Chapter Four. NASA should prioritise the funding of alternative research whereby there is more consensus over positive payoffs.

5.1. Introduction

Since the US Congress halted funding the search for extraterrestrial intelligence (SETI) in 1993 (Garber, 1999), SETI has been privately funded. Despite being cut off from NASA's deep pockets, SETI's finances are healthy; it is kept buoyant by a wealth of enthusiastic and generous philanthropists. And with an annual budget that fluctuates around the $20-30 million dollar mark (SETI Institute, n.d.), teamed with a recent $200 million philanthropic gift (Feehly, 2023), the future finances of SETI continue to look secure.

The reasons for NASA's severance from SETI are discussed in detail in Garber (1999). Some of the contributing factors cited in Garber (1999) as 1) a deficit in the federal budget and SETI being an easy target, 2) unfounded associations with unscientific, sci-fi notions, and 3) a lack of support from other scientists and politicians – e.g., Senator William Proxmire who was sceptical about NASA funding generally; in response to a discussion on space colonies, he stated that "it's the best argument yet for chopping NASA's funding to the bone… I say not a penny for this nutty fantasy" (Lovell, 1977).

Chapter Four of this thesis has made the case for funding some high-uncertainty projects in astrobiology. And yet, the highly speculative search for extraterrestrial intelligence has been banished from the public purse. This chapter takes the prescriptions of Chapter Four and applies them to the case of SETI. I evaluate whether an argument in favour of funding high-uncertainty, high-payoff research compels us to favour the public funding of SETI. §5.2 of this chapter makes the case that SETI represents an excellent example of high-uncertainty research, and this uncertainty truly comes from the research area being highly unexplored and novel. This is done by evaluating the uncertainty present in almost every parameter of the famous Drake equation.

On the way to this finding, §5.2 will also provide a literature review of scientific opinion on several fundamental probabilities in astrobiology. These probabilities concern the likelihood of detecting extraterrestrial intelligence, and I assess to what extent it is possible to pin these down. I take the famous Drake equation as a framework for what variables might be relevant for ascertaining such a probability. The most poorly constrained final four variables receive special focus. These are: the fraction of habitable planets that develop life (abiogenesis), the fraction of those that develop intelligent life, the fraction of those that then develop comprehensible communication, and finally, the average lifetime of such civilisations. It is found that our sample size of one (the N = 1 problem) is largely responsible for our inability to pin down these variables. Hence, SETI research is an honest example of high-uncertainty research.

However, the jury is out as to whether the successful detection of extraterrestrial intelligence would be high-payoff. §5.3 evaluates whether this is the case and hence whether SETI should fall under the high-uncertainty, high-payoff projects I have defended the funding of in Chapter Four of this thesis. In the face of concerns over the potentially negative cultural impact of detecting extraterrestrial intelligence and the prevalence of innumerable high-uncertainty, high-payoff project proposals in astrobiology, SETI should not be top of the list. §5.4 concludes that, although an excellent example of high-uncertainty research, SETI cannot yet be considered sufficiently high-payoff to compete with other candidate astrobiology projects.

5.2. The Drake Equation

When evaluating whether SETI research should be considered high-uncertainty, the Drake equation offers a fantastic place to start. Proposed in 1961 by Frank Drake, the Drake equation provides a framework to discuss how many intelligent, communicating civilisations might exist within our galaxy. The equation was not intended to advance any particular number of such civilisations, but rather to highlight what relevant parameters scientists should consider when coming to their own estimates about detecting life (Drake, 1961). The equation is as follows:

$$N = R_* \cdot f_p \cdot n_e \cdot f_1 \cdot f_i \cdot f_c \cdot L$$

Where: *N* is the number of civilisations in the Milky Way with which we could communicate; R_* is the average rate of star formation in our galaxy; f_p is the fraction of such stars that host planets; n_e is the average number of habitable planets per star hosting planets; f_1 is the fraction of those planets where life actually emerged; f_i is the fraction of *those* planets where life developed intelligence; f_c is the fraction of *those* planets that developed comprehensible communication methods; and finally, *L* is the average length of time that such civilisations are communicating their existence (Drake, 1961).

With a framework provided for estimating the number of civilisations with which we could communicate, it remains for individuals to substitute in their own estimates for each parameter. The issue, though, is that estimates range dramatically. Drake and his colleagues gave their best guess as between 20 and 5×10^7 with the parameter assignments: $R_* = 1\ \mathrm{yr}^{-1}$, $f_p = 0.2$ to 0.5, $n_e = 1$ to 5, $f_1 = 1$, $f_i = 1$, $f_c = 0.1$ to 0.2, $L = 1000$ to 100,000,000 years (Drake, 1961). And this vast disagreement is within one group alone! Other estimates increase the variance of detectable civilisations even more (as discussed in Wilson, 2001; Dick, 2020, pp.81-99). For example, some of the most conservative estimates for these parameters include the rare Earth hypothesis value of $f_p \cdot n_e \cdot f_1 = 10^{-12}$ (Shapley, 1959), Ernest Mayr's pessimistic

value of $f_i = 10^{-9}$ (Mayr, 2010), Drake's lower bound of $f_c = 0.1$ (Drake, 1961), and Shermer's $L = 300$ (Shermer, 2002). Inputting these into the Drake equation gives a value of $N = 3 \times 10^{-20}$. Such a value indicates that we are very likely alone in our galaxy and that even our own existence is somewhat miraculous. The disagreement then in the range of this pessimistic estimate and Drake's upper bound estimate is huge: of the order 10^{27}**!**

It is apparent, then, that there is huge disagreement over this fundamental figure in the search for extraterrestrial intelligence, that is, the number of civilisations with which we can currently communicate. But where might this uncertainty in the value of *N* come into the equation? Is there just one particular parameter causing all the uncertainty, or is each and every part of the equation poorly constrained?

To better understand this uncertainty and justify that we are dealing with genuine uncertainty due to a lack of data rather than risk masquerading as uncertainty, let us walk through each parameter to evaluate what the current literature has to say on how well-defined it is. We will start off with the first three less problematic parameters, before moving on to the more slippery final four parameters.

5.2.1. The First Three, Less Problematic Parameters

At first glance of the Drake equation, its ability to provide an objective value for *N* initially looks promising. This is because the first three parameters are relatively easy to constrain. Empirical data exist on each of these to minimise speculation. Take the first parameter: R_*, the mean rate of star formation. This is something that we can observe and track over time using ground and space telescopes. Data collected from NASA and the European Space Agency put the rate of stellar formation at 0.68 to 1.45 solar masses per year in our galaxy (Robitaille & Whitney, 2010). More recent analyses put this slightly higher at 1.65 (Licquia & Newman, 2015) and 2.0 ± 0.7 solar masses per year (Elia et al., 2022).

Just a note about these estimates: these values are technically for the *solar mass* formation per year, not the number of stars formed per year. To convert from the former to the latter, we must divide by the ratio of the average star in our galaxy's mass to the sun's mass. This has been suggested to be ~ 0.5 (Kennicutt & Evans, 2012). Hence, we effectively double the above estimates of the formation of solar masses per year to get the stellar formation rate.

The discrepancy in the solar mass formation (and hence stellar formation) might seem significant. However, in the context of the Drake equation, the range in values is small. There may be some slight disagreement over the exact value, but all mainstream suggestions are within the same order of magnitude,

that is, 1 to 10 stars per year. Hence, this first parameter of the Drake equation is reasonably well-defined with empirical data. It is thus not significantly contributing to the uncertainty in the overall equation.

The second parameter is f_p – the fraction of stars that host planets. This, too, is similarly well-constrained. A 2012 study using gravitational microlensing puts this at about 1: "Stars are orbited by planets as a rule, rather than the exception" (Cassan et al., 2012, p.2). It has been highlighted that we should expect ~100% of *sun-like* stars to host planets (Lineweaver & Grether, 2003), but beyond this caveat, estimates of f_p are again of the same order of magnitude. This second parameter is, therefore, not significantly contributing to the uncertainty in *N*.

Finally, n_e represents the number of habitable planets per star with planets. This value provides the most uncertainty of the first three. However, it is being categorised with the other two as the uncertainty associated with this parameter pales in comparison to the uncertainty of the remaining parameters. Having said this, the uncertainty with this parameter comes in with the question of what *habitable* means. Continued research into extremophiles gradually pushes the edges of what is considered habitable with ever more extreme temperatures, pH, radiation, pressure, and salinity (e.g., Heuer et al., 2020; Clarke et al., 2013; Marion et al., 2003) being considered consistent with life on Earth.

Moreover, if we are just looking for planets that can host life similar to that on Earth, we may be missing myriad exotic civilisations. Life may have evolved to be unrecognisable on other planets. With different evolutionary pressures, it is difficult to conceive of the creative way in which life may have adapted and thrived. A pH that is too low for any lifeform on Earth may be optimal for a theorised organism that evolved in the acidic rain of Venus. As such, we should expect the number of habitable planets to be larger than the number of planets that might support life comparable to that on Earth.

Resultingly, we reasonably have a lower bound with Earth-like planets in the habitable zone of sun-like stars, and so even just considering these alone as the habitable planets gives us a minimum of ~ 22% (Petigura et al., 2013). Assuming such a lower bound again ensures a range of values for n_e that are within the same order of magnitude (i.e., we might expect n_e to be somewhere between 0.22 and 1.00). To summarise then, although the first three parameters of the Drake equation are the focus of much research and lively debate, they do not seem to account for the astronomical disagreement (of the order 10^{27}) in the output of the Drake equation. It is then in the final four parameters that we might find the root of the uncertainty.

5.2.2. f_l, The Probability of Abiogenesis

Nestled within the middle of the Drake equation is a parameter highly elusive. f_l, the fraction of habitable planets where life emerges, has historically evaded pinning down. The root of the issue is the N = 1 problem: we have a sample size of one; life emerged here on Earth once, so we cannot infer probabilities about abiogenesis in our galaxy more generally. We cannot know that our sample size of one is representative in any way. It is indeed possible that all habitable planets eventually harbour life ($f_l = 1$), or it is possible that we were an absolute exception ($f_l \approx 0$).

Attempts to quantify f_l, therefore, do not rely on statistical inference from empirical observation in the same way that the first three parameters do. England (2013), for example, suggests that life is actually thermodynamically preferred; molecules have a tendency to self-organise so as to better regulate heat exchanges between themselves and their environment. This theory implies that f_l is quite high – perhaps close to unity. In contrast to this is the idea that abiogenesis occurs when nucleotides combine completely randomly in just the right way to form RNA. Totani (2020) calculates how big the universe would need to be for the minimum number of 20 nucleotides to have randomly combined at some point in the history of the universe. It is consequently found that the probability of abiogenesis randomly occurring in the observable universe is negligible, $<<1$.

The principle of mediocrity is cited as a principled reason for believing that f_l is close to unity (or at the very least, $f_l \approx 1$ on Earth-like planets). The principle states that, if an object is drawn at random from a class of objects, it is more likely to have characteristics representative of the class than not, i.e., it is likely mediocre in its class, not exceptional. This has led some (e.g., Sagan, 1994) to argue that the same logic applies to life in the universe. We should employ the principle of mediocrity to conclude that, if life emerged here on Earth, it probably emerged elsewhere: "We have not been given the lead in the cosmic drama" (Sagan, 1994, p.28).

The problem, though, with this reasoning is that it might not be valid to say that our datum (the existence of life on Earth — us) has been randomly drawn. The principle of mediocrity stands in opposition to the anthropic principle. The anthropic principle, developed by Carter (1974) but proposed earlier by Dicke (1957), asserts that there exists a selection bias when making conclusions about life in the cosmos. When considering the Earth's position in the Goldilocks zone of a friendly star, the age of the universe and even all the way down to the fundamental constants of nature, it seems as if the universe and our place within it have been fine-tuned for our existence. Stephen Hawking stressed this: "The remarkable fact is that the values of these numbers seem to have been very finely adjusted to make possible the development

of life" (Hawking, 1988, p.125). If, for example, the strong nuclear force were 2% stronger, stable stellar fusion might be impossible (Davies, 1993), and thus life could never have begun.

The anthropic principle applied to the frequency of life in the universe may be summarised as so: our existence tells us almost nothing about the prevalence of life elsewhere. It is a necessary condition that the universe and our place in it appear fine-tuned to our existence, as otherwise, we would not be here to ask the question. Our existence does provide a very weak lower constraint: we know that the probability of abiogenesis must not be zero. But beyond this, we cannot infer any more detail. If there was quite literally one species in the entire universe asking the question of abiogenesis likelihood, we would necessarily be it. Similarly, the universe could be teeming with life. Our existence alone is wholly uninformative regarding which of the two scenarios we could be in.

The anthropic principle highlights a common mixing up of conditional probabilities. We may ask two questions: 1) how likely is it that life emerged, given that the universe is seemingly fine-tuned for life? Or 2) how likely is it for the universe to be so seemingly fine-tuned for life, given that life emerged (i.e., we are here asking the question)? The latter question is better constrained. Indeed, the probability of the universe being fine-tuned to life, given that life exists, is unity. In contrast, the first question is very poorly constrained due to our singular data point. The anthropic principle stresses that our existence only provides a definitive answer to the second question, not the first one.

Attempts at pinning down f_l using Bayesian analyses have had mixed receptions. For example, Carter (1974, 1983), Spiegel and Turner (2012), and Bostrom (2002) have analysed the use of Bayesian statistics to constrain the probability of abiogenesis. A central problem though of employing Bayes' theorem to update our existing prior based on the knowledge that life exists on Earth is that we quickly run into the notorious "problem of old evidence" in Bayesian statistics. To summarise the argument: a Bayesian may wish to update her prior probability that abiogenesis is common $P(A_c)$, given the knowledge that life emerged here on Earth, L_E. She will then do the following:

$$P(A_c|L_E) = \frac{P(L_E|A_c)P(A_c)}{P(L_E)}$$

Using the law of total probability, we can rewrite $P(L_E) = P(L_E|A_c)P(A_c) + P(L_E|not\ A_c)P(not\ A_c)$. This then gives:

$$P(A_c|L_E) = \frac{P(L_E|A_c)P(A_c)}{P(L_E|A_c)P(A_c) + P(L_E|not\ A_c)P(not\ A_c)}$$

Because we are using the "old evidence", that is, the existence of life on Earth, L_E, we will see that this update is wholly uninformative. This is to say, $P(A_C|L_E) = P(A_C)$. This is because the probability that life emerged on Earth, given that abiogenesis is common, $P(L_E|A_c)$, equals the probability that life emerged on Earth, given that abiogenesis is not common, $P(L_E|not\ A_c)$, equals 1. We know that life did indeed emerge on Earth, regardless of how unlikely abiogenesis was! That is: $P(L_E|A_c), = P(L_E|not\ A_c) = 1$. Equating these conditional probabilities to 1 gives:

$$P(A_c|L_E) = \frac{P(A_c)}{P(A_c) + P(not\ A_c)}$$

And since $P(A_c) + P(not\ A_c) = 1$, the equation collapses to $P(A_c|L_E) = P(A_c)$ and we see that the update has been completely uninformative. Whitmire (2022) similarly presents the problem of old evidence applied here to the probability of abiogenesis being common.

The proposed solution to the problem of old evidence in Bayesian statistics is to abandon the omniscient view we have when faced with old evidence and instead present the conditional probabilities as hypothetical. What this would mean for the above application is that, instead of asking what the probability is that life emerged on earth, given abiogenesis is common or not common (which would, with our omniscient view, be 1 as this is old evidence — we know life did emerge), we should ask hypothetically, what would be the probability if we did not know that life did in fact emerge.

The effect of this shift in perspective is analogous to the shift in perspective resulting from the anthropic principle. We cannot use our existence in any way to make probabilistic conclusions about the likelihood of abiogenesis more generally, with the very minor exception of installing a lower bound of not zero. By using these hypothetical conditional probabilities (as opposed to the uninformative "old evidence"), we could theoretically get an update on our prior probability of abiogenesis.

However, we are now right up against the problem of the priors in Bayesian statistics. We find that our posterior probability for abiogenesis is largely determined by our priors, including our prior probability for abiogenesis. Balbi and Lingam (2023) offer a detailed analysis of this and conclude that "the single datum that life has appeared at least once on Earth merely sets weak constraints on the minimal probability of abiogenesis. In fact, the *a priori* probability assigned to this event (viz., optimistic, pessimistic, or agnostic prior) exerts the strongest influence on the final result" (Balbi and Lingam, 2023, p.3117).

This result, that we cannot turn to Bayes to reveal new information about the probability of abiogenesis, should not be surprising. Bayes' theorem offers a mechanism to assimilate data into your existing data

set. It cannot generate new information. Our outputs cannot contain less uncertainty than our inputs: statistical trickery cannot provide an alternative to empirical data.

An alternative argument used to motivate a high value for f_l has some intuitive appeal. This is the argument from early emergence. Life emerged on Earth just about as soon as it feasibly could. Although still in its fiery Hadean Eon, around 4.0 billion years ago, the Earth had cooled just enough to be considered habitable. Then just a couple hundred million years later (a mere blip in the history of the Earth), life was resident on Earth. This is seen in the earliest discovered fossils dating back around 3.7 billion years (Nutman et al., 2016). Although the validity of these particular stromatolites has sparked debate (e.g., Allwood et al., 2019), less contentious stromatolites date back 3.5 billion years (Furnes et al., 2004). It appears, therefore, that life emerged on Earth just about as early as it could – intuitively seeming as if abiogenesis were common.

One could imagine a scenario where life did not emerge nearly as soon as it could on Earth, and instead, the Earth remained sterile for a couple of billion years. This might give the impression that abiogenesis is not so common – it took a couple of billion years to eventually emerge on a friendly planet. Such an argument is proposed as a way of wiggling out of the N = 1 problem we face when trying to extrapolate our existence to the probability of abiogenesis more generally. Lineweaver and Davis (2002), for example, take the rapid emergence of life on Earth and use this to provide a lower bound for abiogenesis elsewhere. Referring to terrestrial planets, they conclude that "We find that on such planets, older than approximately 1 Gyr, the probability of biogenesis is > 13% at the 95% confidence level. This quantifies an important term in the Drake Equation" (p.293).

Indeed, this reasoning might be taken to provide weak evidence in favour of a non-negligible value for f_l, however, any quantification of this is overstating the available data. It is true that life emerged early on Earth, but we cannot assume this was typical. Moreover, some have argued for a selection effect that favoured the earlier emergence of life on Earth. This is on account of intelligent life necessarily needing to have emerged before the increasing solar intensity results in an end to Earth's habitable state (Totani, 2020). As a result of the early emergence of life on Earth not necessarily being representative of life on other planets and potentially being subject to a selection effect, the early emergence of life on Earth is not especially informative regarding the probability of abiogenesis.

Of final consideration before concluding this survey of recent literature on the probability of abiogenesis is an acknowledgement of the rather fringe theory of panspermia. This proposes that life started elsewhere in the universe and has been distributed throughout space via meteoroids and space dust. In the case of Earth, life was then seeded as some of this material fell to Earth. Although the theory has some proponents (for example, Rampelotto, 2010; Wickramasinghe, 2011; Siraj & Loeb, 2020), it remains far

from the mainstream. Nonetheless, it is worth noting that, if panspermia were correct and life on Earth is just one seed of many in the universe, then f_l would need to be modified to include the probability of life being seeded on a planet. I will not develop this further here, though, and for convenience, I will proceed with the assumption that panspermia is false.

To conclude then this section on f_l: our sample size of one for the emergence of life is hugely pernicious to our abilities to constrain f_l. Attempts to extrapolate from our single datum are problematic; the principle of mediocrity is undermined by a selection effect highlighted by the anthropic principle. Moreover, attempts at updating f_l with the information that life emerged here on Earth using a Bayesian framework are unhelpful. Either we face the problem of old evidence, or the problem of the priors when attempting to utilise Bayes' theorem in this way. Resultingly, the update is uninformative in either case. Attempts to appeal to the rapid emergence of life on Earth to indicate a high value for f_l might also be undermined by a selection effect problem (Totani, 2020)

Our sample size of one for the emergence of life, therefore, leaves us unable to make unspeculative conclusions about the probability of abiogenesis elsewhere in the universe. Bayesian statistics cannot assist in this as a statistical alternative to genuine empirical data, and we simply do not have this data. Hence, the probability of abiogenesis, f_l remains highly unconstrained and is a contributing parameter to the overall uncertainty of *N;* this lends to the position that SETI research is high-uncertainty research.

5.2.3. f_i, The Probability of Intelligence Developing

Moving on now to the fifth parameter in the Drake equation and things do not get more straightforward. f_i, the probability of life becoming intelligent, smuggles in yet more uncertainty into the equation. Having said this, attempts have been made in the literature to constrain this. This section will address some of these and offer rebuttals such that we are left with another wildly unconstrained parameter.

To address the question of how likely it is for life in the universe to become intelligent, we may have to first take on the inevitable job of defining intelligence. The question of what counts as animal intelligence is at the core of rich and impassioned research. One popular measure is that of tool use, which has been well documented in several species — e.g., from chimpanzees (Goodall, 1964), to ants (Maák et al. 2017), to New Caledonian crows (McGrew, 2013). Other measures of animal intelligence include: mirror self-recognition, as documented in fish (Kohda et al., 2019); simple mathematical skills, as seen in bees (Howard et al., 2019); and the ability to feel empathy, as observed in rats (Bartal et al., 2011).

However, debate surrounds which list of measures best captures animal intelligence, and at the heart of these debates are considerations of value-ladenness. It can be difficult for humans to separate out our perspective when designing experiments to test animal intelligence. One particular example is the mirror self-recognition test (initially designed by Gordon Gallup, 1970). Animals who, upon looking in a mirror, notice a mark on their forehead, are considered to pass the mirror self-recognition test. Many animals are said to fail this test, but really it is the test that fails. For example, the failure of gorillas has been attributed to the fact that gorillas avoid eye contact with members of their own species, as doing so is considered threatening (Shillito et al., 1999). Similarly, the failure of dogs to pass the test has been linked to their primary dependence on smell as opposed to sight as a means of interacting with the world (Horowitz, 2017). Interestingly, dogs pass an analogous test that focuses on smell self-recognition, as opposed to sight (Horowitz, 2017).

Searching for a universal definition of intelligence, then, appears problematic. However, there may be a way to wriggle out of needing to land on such a definition. What we really need from this parameter is not necessarily the probability of life becoming *intelligent* (and here we substitute in whichever problematic and subjective definition we may hold), but rather, we need the probability of sufficient complexity associated with organisms that might then have the capability of communicating with species beyond their own planet. Whether or not that complexity is in such a direction that the organism has the *inclination* (as well as baseline capability) to communicate is a question for the sixth parameter.

This interpretation of f_i is sufficient for the Drake equation to be valid. So, this will be the interpretation I will refer back to throughout this section. To reiterate then: f_i, is the fraction of planets harbouring life where the life developed sufficient complexity such that it has the *capability* of communicating beyond its host planet. This parameter does not also necessitate that the species in question has human-like intelligence, i.e., that the *type* of intelligence lends itself to human-like pursuits. Meeting the requirements of f_i, is contingent on the species having sufficient mental complexity; it is not concerned with what type of tasks the species is mentally equipped for.

To help ground this definition of intelligence, we would consider that a goldfish would not have sufficient complexity to communicate, whereas humans and conceivably chimpanzees, dolphins, and octopuses might. Indeed, chimpanzees, dolphins, and octopuses do not attempt to communicate with extraterrestrials; the specifics of their historical evolutionary pressures have not led to prioritising this type of intelligence (again, this will be addressed with the sixth parameter). However, we might reasonably regard such animals to have evolved the sufficient complexity of mind associated with extraterrestrial communication, even if their intelligence evolved in a different direction than that which would facilitate extraterrestrial communication. Given this, it seems as if life has become sufficiently complex on Earth

at least once and likely more than once. How, then, might we determine a figure for the overall likelihood of such complexity?

First though, it is quite tempting to adopt the human-centric and fallacious view of evolution, with humans occupying the top spot of a pyramid of evolution. In this view, humans stand atop all other animals, with other mammals usually one rung below, then fish and birds, followed by insects, and finally, at the base of the pyramid, are the plants. This anthropocentric view of evolution panders to a false sentiment that humans are the end goal of evolution and that all the other species on earth are mere stepping stones (Miller, 2012). Such a sentiment has been termed *ladder thinking* and has been found to be a prevalent misconception amongst US university students (Kummer et al., 2016). This ladder view of evolution has been attributed as far back as Aristotle's *The History of Animals* (1991). See Archibald (2014, Chapter One) and Mayr (1982) for discussions of this.

The issue with this, however, is that it completely contradicts the workings of evolution. As all life on Earth shares a universal common ancestor, all existing life has been evolving for the same amount of time: we are all just as evolved as each other. Thus, asking the question of who is the "most evolved" is vacuous; the idea that humans sit atop an evolutionary pyramid is false and harmful (Werth, 2022). A relevant question to ask could be which species is best suited to their current environment. And the answer to this will quite literally change with the wind, as evolutionary fitness is tied to and relative to what the environment is doing at that time in history. As such, we must avoid such pyramid ideologies that imply that evolution marches towards human intelligence over time.

So, with the view established that humans are not the end point of evolution, and rather *all* life on Earth has been evolving for the same length of time, we are now in the position to assess what fraction of species develop sufficient complexity. Life may not tend to become more human-like, but does it tend to become more complex such that, given time, the level of complexity we require is an inevitable event?

Frank Drake himself was very optimistic about this parameter; his and his colleagues' initial estimate at N put f_i at 1 (Drake, 1961). Lineweaver (2007) cites an on-flight conversation with Drake about his optimistic opinion on the convergence of intelligence. Lineweaver cites:

> "The Earth's fossil record is quite clear in showing that the complexity of the central nervous system — particularly the capabilities of the brain — has steadily increased in the course of evolution… creatures with superior brains are better able to save themselves from the sudden change in environment. Thus smarter creatures are selected, and the growth of intelligence accelerates… This picture suggests strongly that, given enough time, a biota can evolve not just one intelligent species, but many" (Drake, 2006, cited in Lineweaver, 2007, p.357).

This reasoning has intuitive appeal. It is one that Sagan concurs with: "Other things being equal, it is better to be smart than to be stupid, and the overall trend toward intelligence can be perceived in the

fossil record" (Sagan, 1995). The evolutionary benefit of intelligence appears reasonable, but it is a soft argument premised on intuition. So, what is the hard evidence for this increasing complexity in the fossil record?

Harry Jerison (1955, 1973) has documented the increase in what he coins the Encephalization Quotient (E.Q.), defined as the ratio of brain weight to some power of body weight. From this, Jerison argues that E.Q. tracks with complexity, controlled for organism size (Lineweaver, 2007). And the data are reasonably convincing. Relative brain size (and thus reasonably and roughly complexity) does seem to have increased over the evolutionary history of life on Earth. This conclusion is similarly drawn elsewhere (e.g., see Russell, 1983; Rospars, 2013). This might suggest that the guiding hand of evolution selects for ever-increasing complexity, and hence, our required level of complexity might be an evolutionary trend.

If such a conclusion is true, that intelligence is a convergent feature of evolution, then f_i might be set to approximately 1. Moreover, the quality of the evidence for this assertion appears to be of a different kind to the highly speculative evidence for f_l, namely because it *seems* as if we do not have a sample size of one here. It looks as if complexity evolved independently on Earth numerous times, and hence we can confidently infer from our vast pool of data that complexity is inevitable. However, Lineweaver (2007) poses an argument that just might undermine this conclusion and bring us right back to the N = 1 problem.

This critique from Lineweaver undermines the independence of the numerous instances of evolved complexity. Lineweaver (2007) points to the shared evolutionary history that all species on Earth have (to varying extents) and argues that this limits the amount of independence that can be attributed to the evolution of complexity. He concludes: "When considering convergence, a basic principle is often ignored: the extent of convergence cannot be larger than the extent of divergence from the common ancestor" (p.361).

To explicate this, Lineweaver considers the seemingly convergent feature that is the eye. From looking at the multitude of animals from different evolutionary branches that have eyes, you might conclude that eyes have independently evolved numerous times, and hence, there exists some evolutionary guiding pressure for eyes analogous to the argument for complexity/intelligence. However, to assess whether a feature is truly convergent, we must take the species that possess eyes and compare their independent evolutionary histories with their shared evolutionary histories. In the case of eyes, Lineweaver points out that eye-possessing organisms have been evolving independently for about 500 million years, whereas their shared evolutionary history spans about 3500 million years (Lineweaver, 2007, p.361). This means that their independent histories make up only about 15% of their total evolutionary history.

Lineweaver highlights how the evolution of features is a gradual, stepwise process and, in the case of the eye, "the common ancestor of the independent eyes had, during ~3500 million years, already evolved the complex biochemical pathways for photoreception" (2007, p.361). The extent to which the eye existed 500 million years ago before the various species diverged need not be particularly significant. All that matters is that a groove had been made (or "toggle switch" as Lineweaver calls it) that then set the subsequent species on a path. The very first bases of photoreception made the continued development of eyes an easier pathway for evolutionary prowess. To conclude Lineweaver's argument on shared evolutionary history, he summarises as so:

> "What Drake, Sagan and Conway-Morris have done is interpret correlated parallel moves in evolution as if they were unconstrained by shared evolution but highly constrained by a universal selection pressure towards intelligence that could be extrapolated to extraterrestrials. I am arguing just the opposite – that the apparently independent evolution toward higher E.Q. is largely constrained by shared evolution with no evidence for some universal selection pressure towards intelligence." (2007, p.362)

Lineweaver, then, challenges the notion that there is an evolutionary drive towards higher and higher complexity. He argues that the *apparent* convergence on complexity on Earth is really largely the result of an early shared ancestor evolving the beginnings of intelligence and then evolution continued to carve into this groove as one possible evolutionary pathway. Moreover, intelligence may have just been one of many possible evolutionarily beneficial pathways. The happenstance that a far-distant ancestor developed some rudimentary intelligence might best be described as a fluke which evolution latched on to as intelligence can provide an edge in survival.

But intelligence is certainly not necessary for survival, and numerous organisms here on Earth thrive without it. Bacteria, for example, would not be considered sufficiently complex, yet are approximately 5×10^{30} in number (Lehman, 2017). Bacteria are thriving and are very proficient at passing on their genetic material. It is difficult to conceive of alternative ways that evolution might have run, but the contingency of evolution has been noted widely (e.g. Gould, 1990). If, to use Lineweaver's language, the toggle for intelligence had never been switched, perhaps life on Earth would have similarly thrived as simple organisms.

Lineweaver's arguments have merit, especially when considering concrete features like the eye. The phrase *the extent of convergence cannot be larger than the extent of divergence from the common ancestor* (Lineweaver, 2007, p.361) truly captures an intuitive logic. However, there is space to push back on his claims when it comes to the more nuanced idea of intelligence. It is relatively straightforward to look back in the ancestral trees of eye-possessing organisms to point out that their shared common ancestor had rudimentary photoreception. But how might this work for intelligence-possessing organisms? What would be the beginnings of intelligence when the concept seems so nuanced and multifaceted?

Both sight and intelligence can come in degrees, but this appears more so with intelligence. Both humans and octopuses possess significant intelligence. However, the common ancestor of vertebrates and cephalopods existed approximately 750 million years ago and resembled something of a simple flatworm (Godfrey-Smith, 2013). It is difficult to say that an intelligence toggle had been switched (in a binary way) within this flatworm in the same way that some very rudimentary photoreception might have popped up via mutation. Intelligence strikes as even more gradual and cumulative than eyesight. Hence, it appears arbitrary to point to any one common ancestor and claim that the first toggle was switched on here. Perhaps it is more intuitive to say that intelligence goes all the way down.

If we are to completely buy Lineweaver's argument that intelligence is not a convergent feature of evolution because intelligent species share a somewhat intelligent ancestor, we would need to argue that this 750-million-year-old flatworm possessed a degree of intelligence that set it apart from the ancestors of animals we would not deem intelligent. This is a tall order on account of the blurred lines of intelligence.

Hence, Lineweaver's argument is mixed in its persuasiveness. He rightly points out that careful analysis of common ancestors must play a central role in any determination of convergent features of evolution. Any degree of convergence must be prefaced on the degree of divergence from a common ancestor. The eye is an excellent example of this. However, the vagueness of intelligence makes it difficult to point to any individual ancestor as having the "intelligence toggle" switched on. The shared ancestor of humans and octopuses strikes as an odd place to do this.

Nonetheless, Lineweaver's arguments are enough to throw into question the assertion that intelligence is a convergent feature of evolution. If the reader is convinced of Lineweaver's argument, we again hit an N = 1 problem; the initial groove for intelligence may have been carved once, and subsequent species have since leaned increasingly into that. If so, we cannot use the multitude of intelligent species on Earth as independent lines of evidence for the likelihood of intelligence emerging generally. f_i then evades pinning down and more uncertainty is brought into the Drake equation.

5.2.4. f_c, The Probability of Intelligent Life Developing Comprehensible Communication Techniques

Of all the parameters in the Drake equation, it is perhaps the most difficult to ascertain the meaning of the sixth, let alone determine a value for it. f_c is often defined as the proportion of intelligent civilisations

that go on to develop communication methods that we could detect (Drake, 1961). These communication methods are expected to centre around radio telescopes and, on account of humans being the only species on Earth that have built such technology, f_c is sometimes alternatively defined as the proportion of intelligent civilisations that go on to evolve into humanoids, that is: they develop human-like intelligence (Lingam and Loeb, 2019). The work that this parameter really needs to do is to take the set of intelligent civilisations in the galaxy and pick out the ones that are actively communicating in a way that we will understand.

On Earth, humans are the only species that even come close to meeting this criterion. It is true that countless other species communicate. But because of how the Drake equation is formulated, we need the parameter to isolate the species which we could detect even from across the galaxy. It might be argued that we could detect signs of life — biosignatures — from organisms on different planets that are not necessarily sending out radio waves and thus not actively communicating their existence. Indeed, this pursuit is one of the differential factors that separates biosignature research from SETI. The problem with this, though, is that the Drake equation concerns the number of detectable civilisations in the entire galaxy, and so the vast majority of these would be far too far away for us to detect without the use of radio waves — the distances are just too vast.

Even more troubling is the suggestion that even we humans have not reached this level of human-like intelligence! We have only recently become detectable within our own solar system, and even this is caveated. We are not quite the all-singing and dancing smog show we might think we are. With resolutions of 1km, few signs of life can be detected on Earth, even on cloud-free days (Kilston et al., 1966; Sagan et al., 1993). It is at this resolution that agricultural fields and roads become detectable. For a reliable detection of intelligent life, it has been proposed that an entire planet would need to be imaged at 2km resolution (Kilston et al., 1966). It is not surprising, then, that humanity may be invisible to civilisations beyond our solar system. The radiation we emit would be too weak to be detected by our current SETI equipment at distances beyond 100 light-years (Shostak, 2015). And although these distances sound vast, they are not by galactic scales: these are the distances of the relatively nearby stars in our galaxy. As such, even if there are many civilisations advanced enough to develop comprehensible communication techniques, they may not meet the detectability threshold for us to notice them.

So, what does the current literature say about the proportion of intelligent civilisations that develop human-like intelligence (or develop communication technologies that we can detect and comprehend)? In short, the literature is mixed. Lineweaver (2007) again gives a persuasive argument from evolution for a low value of f_c. He notes that the octopus and dolphin are of similar E.Q. to humans and yet have not had the inclination to develop radio telescopes and advertise their existence to the cosmos. He concludes

that "this strongly suggests that high E.Q. may be a necessary, but not a sufficient condition for the construction of radio telescopes… if you live underwater and have no hands, no matter how high your E.Q., you may not be able to build, or be interested in building, a radio telescope" (Lineweaver, 2007, p.363).

There are many ways for a species to be intelligent, but the inclination to build radio telescopes and search for life beyond Earth is a rather specific form of intelligence. Even we humans lacked this inclination until after experiencing a scientific revolution. This argument is made even more powerful by the fact that humans, octopuses, and dolphins all share (at least some) evolutionary history. If there were to be any planet where another example of human-like intelligence was to be found, you might expect it to be on the planet harbouring all the species with a shared evolutionary history with humans! And yet this is not the case. A relatively low value for f_c might, therefore, be justified.

Having said this, proposals have been made to argue for a more optimistic value for f_c. Lingam and Loeb (2019), for example, adopt a value of 0.1. Interestingly, this is for the combined parameters of f_i and f_c. That is, Lingham and Loeb (2019) take the product of the two, and this is done "for the sake of brevity" (p.30). They justify this (and indeed even justify approximating $f_i \sim 1$) by appealing to Darwin (1874) and Marino (2015) and the continuity between human and nonhuman minds. However, this argument has a fatal flaw. It is true that the barrier between humans and nonhumans is one of degree, not of kind, but that does not lead to the conclusion that non-human species tend to human species, given time. There is a sense of humans being the end point of evolution that is doing the work in closing the gap between continuity in our case and a general law of evolution. There does not exist convincing evidence that all evolutionary roads lead to humans, and thus, a high value of f_c has not been motivated.

The implication of these discussions for determining f_c is that there does appear to be some rather weak evidence for taking f_c to be small. We are the only example of life on Earth which developed sufficient visibility to other intelligent life forms in the galaxy — and even our status of possessing this human-like intelligence is under question (Shostak, 2015; Kilston et al., 1966; Sagan et al., 1993). Moreover, the lack of other species on Earth possessing this kind of intelligence is despite our shared evolutionary history.

However, the discussion remains highly speculative. Other features might come into play, such as Planet of the Apes hypotheses (Lineweaver, 2007), which imply that there is always a human-like intelligence niche waiting to be filled. It is possible that, were humans not on Earth, another species would step up to fill the niche and subsequently evolve similarly to humans. However, the evidence for this hypothesis is lacking.

There may exist reason to believe that f_c is not as poorly constrained as f_l and f_i. There have been countless opportunities for life on Earth to develop the type of intelligence that would lend to building radio telescopes, yet it happened just once. However, weakly indicative is as far as the evidence will take us. Hence, the parameter f_c remains largely unconstrained.

5.2.5. *L,* The Average Length of Time that Intelligent Civilisations Communicate Their Existence

The final parameter, *L*, represents the average length of time that an intelligent civilisation communicates its existence. This parameter is needed because, as observers, we require that an intelligent civilisation's existence overlaps with our own (or, technically, the information about their existence reaches us while we are still listening). The presence of this parameter corrects for possible communicating civilisations long in the past or in the future that we could not detect. Hence, the product of the Drake equation, *N,* is the number of communicating civilisations whose presence we can *currently* detect.

The value of this is just simply unknown. In this case, we do not even have an N = 1 problem, but rather an N = 0 problem as we have yet to go extinct. Theoretical speculation, of course, exists. Great filter theories with the filter still in front of us have been proposed to overcome the Fermi paradox (e.g., as discussed in Schulze-Makuch & Bains, 2017). These include notions about intelligent life necessarily self-destructing, either by internal wars or by destroying their host planet. These hypotheses do not attempt to put a number on *L* but suggest that it is relatively small.

Conversely, Lingam and Loeb (2019) have argued for a relatively large value for *L,* putting it in the order of 10^4 years. Attempts have been made elsewhere to pin *L* down. Von Hoerner (1961, p.1840) similarly estimates this in the order of 10^4; Gott (1993, p.317) estimates it at 5.1×10^3 to 7.8×10^6 years. However, these are just estimations from a wholly incomplete (and indeed even empty) data set. In short, we simply do not know *L,* and any attempts to pin it down are limited in empirical guidance and are rather the product of speculation.

5.3. High-Uncertainty, High-Payoff?

Having given a literature review of the current understanding of the seven parameters of the Drake equation, it is apparent that uncertainty is baked into almost every parameter, especially the final four. Such is this uncertainty that the range of possible values for *N,* the number of communicating civilisations

that we can currently detect in our galaxy, is unsurprisingly vast. It is easy to see now how the aforementioned range in the order of 10^{27} comes about.

Moreover, this uncertainty has been shown to come from a genuine lack of data, rather than ignorance of available evidence. It is not the case that we can carry out simple experiments to better constrain, for example, the proportion of habitable planets that go on to develop life. Hence, this uncertainty irises from genuine unknowns and cannot be readily reduced to quantifiable risks. As such, SETI research fits well with the type of high-uncertainty projects defended in Chapter Four of this thesis. The subject matter of SETI research sits within an obscured and relatively unknown part of astrobiology, and the potential findings could be significant.

So, does this mean SETI research is high-uncertainty, high-payoff? If it can be presented as such, the findings of this thesis thus far could be used in favour of NASA re-funding SETI. Certainly, if extraterrestrial intelligence were to be detected, the impact would be highly significant. But whether the impact would be incredibly *high* or incredibly *low* is unclear. It is difficult to predict the fallout from such a detection, and much has been written on how best to manage the public communication of such: most notably the International Academy of Astronautics post-detection and reply protocols (Tennen and Forgan, 2018) and the Rio Scale, with its later amendments (Almár, 2001; Almár & Tarter, 2011; Forgan et al., 2019). The scientific interest might be huge, but the social implications are hard to predict.

To elaborate on this point, the Rio Scale was designed to handle the communication of potential extraterrestrial intelligence (ETI) detections (Almár, 2001; Almár & Tarter, 2011; Forgan et al., 2019). The two-dimensional scale quantifies the significance of a claimed discovery of ETI and is given by the product of the consequence of detection, multiplied by the credibility of the detection. The motivation for this scale came from the Torino Scale, which quantifies the significance of an asteroid impact claim. Similarly, the Torino Scale is defined as the product of the potential damage from the asteroid impact, multiplied by the probability of its collision with the Earth (Binzel, 1997).

Notably, though, the language surrounding the potential damage of an asteroid impact is exclusively and understandably negative. However, the language surrounding the "consequences" of an ETI detection is ambiguous; the authors do not refer to these consequences as either positive or negative. Instead, they group the consequences into three categories: minor, moderate, and substantial (Almár & Tarter, 2011) and these are subjectively determined by considering the class of phenomena, type of discovery, and distance to us. The authors recognise the subjectivity and suggest that the determination of which category an ETI claim should fall into should be informed by future research by social scientists (Almár & Tarter, 2011, p.3). The notable silence in commenting on whether a successful ETI detection would

be a positive event, or a negative event supports the claim that, although SETI research could be highly significant, it is not clear whether this would be significantly good or significantly bad.

Yet more divisiveness comes with whether we are simply listening for signals or advertising our existence. At present, SETI's actions are restricted to only listening for extraterrestrial signals, and with this, the potentially negative payoffs may be less than if SETI were attempting to communicate. It is true that the receiving of a signal may cause social unrest. However, the risks of this should be less than the risks associated with sending a signal back. Indeed, some might believe that advertising our existence to extraterrestrial intelligence would be a giant leap towards humanity's extinction. In this way, the "payoff" would be immeasurably negative. Stephen Hawking's infamous quote comes to mind: "If aliens visit us, the outcome would be much as when Columbus landed in America, which didn't turn out well for the Native Americans" (Jha, 2010).

On the other hand, some might view the potential payoffs of communicating with extraterrestrial life as the best thing that could happen to humanity due to, for example, being taught millennia worth of technological advancements, e.g., as concluded in the "Billingham Report" (Billingham, 1999). Hence, some may believe that SETI is simply the most potentially high-payoff research out there, and the scope of SETI should be increased to include advertising our existence. Finally, there are those somewhat in the middle who express cautious optimism about the psychological, sociological, and political implications of successfully detecting ETI (Harrison, 1997).

An interesting cross-cultural survey reported the breadth of potential reactions to the detection of ETI. Vakoch and Lee (2000) asked American and Chinese undergraduate students to quantify their opinions regarding a set of questions falling under six categories: (1) that extraterrestrial life *exists*, (2) that ETI would be *benevolent* and that we *should respond* to a message, (3) that ETI would be *malevolent*, (4) that communication would be *unsettling*, (5) that communication would be *religiously significant* and (6) that *experts* should be in charge of any reply (ibid, p.737).

A range of opinion was reported, and trends could be linked to the respondents' levels of alienation, optimism, anthropocentrism, and religiosity. For example, less religious American students were more likely to believe that ETI would be benevolent and that we should reply. The same sentiment was held by Chinese students who scored low on the anthropocentrism scale — that is, students less likely to think that humans are at the centre of the universe (Vakoch & Lee, 2000, p.740). Additionally, both American and Chinese students who felt more alienated were more likely to think ETI would be hostile (ibid, p.741), and those who were less anthropocentric were more likely to want to rely on experts if humanity were to reply to ETI communications (ibid, p.741). What this study shows is that there is a range of views regarding whether ETI would be a positive thing or a negative thing, and these opinions are tied to our

individual beliefs and values. An attempt, therefore, to argue that SETI research is unanimously high-payoff would be a contentious one.

A discussion of what lessons can be learnt from history to help us anticipate the societal consequences of the receipt of an ETI signal is given in "Consequences of Success in SETI: Lessons from the History of Science" in *Space, Time, and Aliens* (Dick, 2020, pp. 129-142). In this, Dick draws parallels to the Copernican and Darwinian revolutions to suggest that, although turmoil should be expected soon after a successful ETI detection, the social-scientific landscape will eventually settle (Dick, 2020, pp.129-142). See Dick (2018, especially Chapter 9), Baum et al. (2011) and Tarter (1992) for further speculation on the potential positive and negative payoffs of successfully detecting extraterrestrial life.

As things stand, there is no consensus on the positive or negative impact of communicating with extraterrestrial intelligence (e.g., Vakoch & Lee, 2000). The story is also unclear with merely listening for extraterrestrial signals; some may view this as harmless, while others may be concerned about the cultural consequences of a detection. There is reason to fund high-uncertainty, high-impact research. But not if the impact is negative. Some individuals consider it beneficial to pursue the goals of SETI (Billingham, 1999), while others consider it potentially harmful (Jha, 2010). The potential negative payoffs would not be directed solely at those in favour of SETI; they would be experienced by everyone. The divisiveness, therefore, of the potential payoffs of SETI invalidates it as high-uncertainty, high-payoff research, even as things stand with SETI only listening.

This state of affairs is even more pronounced by the numerous high-uncertainty, high-payoff projects within astrobiology that exist. And these projects are the ones SETI must compete with for funding. A notable class of these projects are in-situ projects involved with the search for simple life, or projects tangentially associated with this. An example of such a project is the recently launched Jupiter Icy Moons Explorer (JUICE). JUICE has the primary aim of characterising three of Jupiter's moons: Ganymede, Callisto, and Europa, with the view of determining their possible habitability (Grasset et al., 2013; Plaut et al., 2014). To this end, JUICE boasts a suite of instruments: spectrometers to analyse the composition of ices and minerals on the moon surfaces, a magnetometer to analyse the interaction between Jupiter's magnetic field and that of Ganymede, and an interferometer to characterise the gravitational fields of Jupiter and its moons, to name but three (Plaut et al., 2014).

In the case of JUICE, there is an overarching goal of determining whether these icy moons of Jupiter could be habitable. But within this goal, numerous sub-goals have value regardless of their implications for habitability. For example, the utility in determining the chemical composition of the moons' surfaces extends into fields beyond astrobiology — findings can be applied directly to Earth. As such, JUICE offers a somewhat high-uncertainty project as the payoffs are unknown; we do not know exactly what

the instruments will find. However, whatever payoffs will result, we would expect these to be positive — even determining explicitly that these three moons of Jupiter are not habitable would be of huge scientific value.

JUICE therefore represents an exploratory mission whereby the hoped-for payoffs are unproblematically positive. Other such examples of multi-faceted high-uncertainty, high-payoff astrobiology projects include JUNO (which studies the atmospheric composition and gravitational fields of Jupiter (Bolton & Juno Science Team, 2010)), and the future missions of Dragonfly (sampling materials from multiple sites on the Saturn moon of Titan (Lorenz et al, 2018) and the Nancy Grace Roman Space Telescope (targeting fundamental questions about dark energy and exoplanets (Mosby et al., 2020)).

These projects stand in contrast to the single-goal-oriented SETI projects which are similarly high-uncertainty, but not similarly high-payoff. It is difficult, therefore for SETI projects to compete for funding with these compound, high-uncertainty, high-payoff projects. Such projects continually churn out high-impact results, whereas SETI projects are more binary in their outputs. The search for extraterrestrial intelligence is almost exclusively carried out on Earth via radio telescopes. These telescopes are directed at tiny sections of the sky, many light years away. They watch and wait. The aim is to stumble across a signal: an electromagnetic wave within the frequency range of the telescopes. But to date, no detection of extraterrestrial origin has been made with any real confidence. The "Wow! Signal" has been acclaimed as the strongest candidate for extraterrestrial intelligence (Ehman, 2019), but, in part due to its one-off detection, it remains a yet-to-be-explained curiosity.

A counter argument might be raised that, although SETI projects are largely single-goal-orientated, they do result in serendipitous, unforeseen discoveries. Such unexpected discoveries are exactly what is hoped to be mopped up by the advocation of funding high-uncertainty projects. Jocelyn Bell's discovery of pulsars in 1968 is a notable example of such a serendipitous discovery. Whilst looking at data from the Mullard Radio Astronomy Observatory, Bell and her advisor, Anthony Hewish, noticed a series of highly regular sharp pulses coming from the same patch of sky. They initially jokingly called the signal "little green men" and thought they might have detected extraterrestrial intelligence (Penny, 2013). However, later research determined that these pulses were the result of a yet-unknown star, and thus pulsars were discovered (Wade, 1975).

Although pulsars were detected 16 years before the establishment of SETI, their unanticipated discovery does stand as an unforeseen benefit from scanning the skies for radio waves. But it also stands as the exception that proves the rule. The wealth of unanticipated discoveries made via in-situ astrobiology missions dwarfs the unanticipated discoveries from SETI. Some of the most significant of these discoveries include the discovery of water ice on Mars by the Mars Phoenix lander (Hecht et al., 2009);

the contentious detection of metabolism on Mars in 1976 (Klein, 1978); and the range of discoveries from the Apollo missions, including the historical lunar magma ocean (Wood et al., 1970). Consequently, SETI projects struggle to compete with in-situ high-uncertainty, high-payoff projects in astrobiology.

To summarise, for SETI to be characterised as high-uncertainty, high-payoff, more discussion is needed over the nature of the payoffs if SETI were to successfully detect an extraterrestrial intelligence signal. Only if a consensus is formed that such an outcome would be beneficial could the current goals of SETI be considered high-payoff. Even more discussion would be needed if SETI were to expand into communicating our existence, and it is difficult to see how a consensus on the positive or negative impact of this could be achieved. Less problematic projects exist within astrobiology, in which there is more consensus over the payoffs. These are the projects that SETI must compete with if it wishes to receive public NASA funds and, at present, SETI's contentious payoffs leave it uncompetitive.

5.4. Conclusions

This chapter has evaluated whether the contentious research of SETI might fall under the category of high-uncertainty, high-payoff research that has been defended in Chapter Four of this thesis. Such a question is significant as SETI does not presently receive public funds from NASA and is rather funded privately. If I were to find that SETI does fall under this desirable high-uncertainty, high-risk category, I may be obliged to call for the public funding of the search for extraterrestrial intelligence. Such a call is unpopular with many, with some notable examples including Senator Willam Proxmire who initially pushed for its defunding in 1981 (Garber, 1999) and George Basalla, Emeritus Professor of History at the University of Delaware who wrote that "extraterrestrials discussed by scientists are as imaginary as the spirits and gods of religion or myth" (Basalla, 2006, p.14).

Indeed, this chapter has found that SETI research is certainly highly uncertain, and specifically, it is uncertain in the necessary way. This is because the uncertainty is due to SETI being situated in a highly unexplored and relatively new field. The cause of this uncertainty has been broken down by the use of the Drake equation. Via a literature review of each of the seven parameters of the Drake equation, it has been shown that many orders of uncertainty are built into at least four of the seven parameters. Hence, the range in the number of intelligent civilisations with which we could communicate might be of the order 10^{27}.

However, I have argued that SETI research is not unproblematically high-payoff, and it is this that invalidates it from being classified as high-uncertainty, high-payoff. The significance of successfully detecting an extraterrestrial signal would likely be vast, but to be classified as high-payoff, a consensus

would need to be met over whether this significance is good or bad. It is not clear that such a consensus has been met at present. As such, more communication over the potential payoffs of SETI is required to ascertain whether a consensus can be met on the matter. This is especially the case when SETI must compete with other projects in astrobiology whose goals are less problematically positive. The JUICE mission has been described as an excellent example of this, as are the JUNO and Dragonfly missions, alongside the Nancy Grace Roman Space Telescope.

With all these things considered, SETI is unable to compete for funding with the wealth of high-uncertainty, high-payoff projects in astrobiology. Its contentious payoffs invalidate it from entering the ranks of high-uncertainty, high-payoff projects defended in Chapter Four of this thesis. As such, the arguments of this thesis do not call for any special treatment regarding the funding of SETI research.

Chapter Six

Revisiting the Oumuamua Debate: The 'Problem' of the Priors and its Value for Astrobiology

The previous chapter of this thesis has made the case that there is high uncertainty surrounding the probability of detecting extraterrestrial intelligence. This chapter will delve into how this uncertainty impacts the inferences made by individual scientists working within the field of astrobiology. At the very least, we would expect uncertainty over such a fundamental probability as the probability of detecting extraterrestrial life to lead to disagreement over the nature of extraterrestrial objects.

This chapter takes the case study of a mysterious object, Oumuamua, and analyses the fallout within parts of the astrobiology community that ensued after its detection. I consider three accounts of why there was disagreement within the astrobiology community over the origin of Oumuamua and I argue that the cause of the disagreement lies with differing prior probabilities for detecting extraterrestrial intelligence. The high uncertainty associated with the detection of extraterrestrial intelligence led to scientists arriving at different conclusions regarding the origin of Oumuamua.

This disagreement, however, need not be a bad thing for the astrobiology community. I conclude the chapter with an argument in favour of disagreement amongst scientists working within an uncertain field. This argument is made in the present chapter but will be developed and assessed quantitatively in Chapter Seven.

6.1. Introduction

The origin of 'Oumuamua' (meaning *first distant messenger)* remains a hotly debated topic. The orthodox explanation, that Oumuamua is an unusual comet, prevails with most scientists. But the now infamous

"LH" — Avi Loeb's hypothesis that Oumuamua is a light sail — has sparked a flurry of papers attempting to establish and settle the disagreement. Yet neither side has managed to convince the other, and there is a distinct sense of the two sides talking past each other.

This chapter attempts to diagnose what is at the heart of the divergent conclusions drawn about Oumuamua's origin and discusses the value of this divergence to the field of astrobiology. As such, §6.2 outlines the key details of the detection of Oumuamua and presents two contending hypotheses about its origin: Loeb's extraterrestrial hypothesis (LH) (Loeb, 2021; Loeb, 2018, p.2; Bialy & Loeb, 2018, p.4) and the orthodox comet hypothesis (Bergner & Seligman, 2023; Bannister et al., 2019). §6.3 then addresses three recent analyses of the Oumuamua disagreement: Cowie's (2023a) undefined priors, Matarese's (2022) meta-empirical arguments, and Lineweaver's (2022b) differing priors. Only Lineweaver's account is found to be a convincing explanation of how both sides ended up with their conflicting conclusions. §6.4 then expands on Lineweaver's identification of differing prior probabilities and concludes that LH is rationally arrived at if one's prior probability is sufficiently non-negligible. Having evaluated the crux of the debate, §6.5 then argues for the value of differing priors for the progress of astrobiology by referencing community modelling (in particular, Weisberg & Muldoon, 2009; Avin, 2019). The conclusions of the chapter are given in §6.6.

6.2. The Curious Case of Oumuamua

In 2017, the University of Hawaii's Pan-STARRS1 telescope detected an interstellar object entering our solar system; it looped around the sun with a perihelion of 0.25AU before exiting back into interstellar space. Beyond its exotic origin, Oumuamua presented three curious properties. Firstly, the object's trajectory was shown to divert from that which would be expected if it were moving solely under the gravitational forces of the solar system — it was detected to have some additional acceleration that followed an inverse square law as it approached the sun. Consequently, propositions that Oumuamua was an asteroid were undermined. Secondly, Oumuamua's shape was highly irregular. The object was too small to have its shape directly imaged, but light curves have restricted Oumuamua's shape to either a "cigar" like shape with an extreme major-minor axis ratio or an oblate spheroid ("pancake" like) (Belton et al., 2018). Both such shapes would be considered highly unusual for a non-artificial object. Finally, the very detection of Oumuamua is a striking peculiarity. Given the short time that Pan-STARRS1 had been operational, and the estimated number density of interstellar objects (guided by planet formation theory), the actual detection of such an object would be considered incredibly unlikely.

6.2.1. The Comet Hypothesis

The prevailing and conventional explanation for Oumuamua classifies it as a comet with unusual outgassing (Bergner & Seligman, 2023; Bannister et al., 2019). Such an explanation can account for the excess acceleration observed as backwards outgassing exhibited by comets provides forward propulsion. The magnitude of this excess acceleration would also increase similarly to that observed by Oumuamua as the comet approaches a star, on account of an increase in radiation resulting in more cometary material being evaporated. However, no such outgassing was observed (Loeb, 2018, p. 2; Trilling et al., 2018, p.2).

It has been suggested that an exotic combination of volatiles might be able to save the comet hypothesis as the required forward propulsion could result from outgassed particles unable to be detected by our telescopes (The 'Oumuamua ISSI Team, 2019). Specifically, if the ejected particles were exclusively large (being in the range of 0.1-1mm), then they would have been beyond the detectable limit of our telescopes, although the mechanism for such large particle outgassing has yet to be understood. More recent research consistent with the comet hypothesis has proposed that Oumuamua's acceleration is caused by the outgassing of molecular hydrogen formed via Oumuamua's H_2O-rich icy body (Bergner & Seligman, 2023).

Yet the comet hypothesis struggles to account for two other unusual features of Oumuamua: its unusual shape and the unexpected detection of such an object in the first place, given the predicted number density of objects in interstellar space. Regarding the first of these, most comets have far less extreme major-to-minor axis ratios. Certainly, comets are not regular — it is not unusual for comets to have one axis twice the length of the other — but Oumuamua displayed an extreme aspect ratio of between 6:1 and 8:1 (Mashchenko, 2019). Moreover, the description of Oumuamua as a comet fails to provide a convincing reason for us having detected it in the first place — either it was luck, or our current models for predicting the number density of objects in interstellar space need major revisions.

Notwithstanding these hurdles, the prevailing explanation among mainstream science is that Oumuamua is a comet with unusual outgassing — the details of which might reveal themselves with further research (and indeed may already have been, e.g., Bergner & Seligman, 2023). Many maintain this position despite the serious problems raised above and with due acknowledgement of the compound low probability of all these features manifesting at the same time.

6.2.2. Loeb's Extraterrestrial Technology Hypothesis

Although the classification of Oumuamua as a highly unusually shaped comet with never-seen-before outgassing is adopted by most scientists involved in the Oumuamua debate, dissenters exist. Prominent Harvard astrophysicist Avi Loeb advocates for a more exotic explanation. He proposes that Oumuamua is an extraterrestrial artifact: specifically, a light sail (Loeb, 2021; Loeb, 2018, p.2; Bialy & Loeb, 2018, p.4). The light sail hypothesis proposes that Oumuamua harnesses solar pressure to propel itself forward. Indeed, this pressure would also result in an excess acceleration in line with the inverse square excess acceleration observed (and would evidently not necessitate any outgassing). Moreover, the required shape of a light sail would have a high surface area to volume ratio, such as an oblate spheroid — just the kind of flat, pancake shape that is consistent with data from Oumuamua. Finally, the number density issue is alleviated, at least in its present form, as, by denying that Oumuamua is an interstellar comet, our statistical models do not need revising.

The nature of this proposed light sail is speculated by LH advocates; It could be space debris from the transportation of cargo between stars by alien civilisations (Lingam & Loeb, 2017). One might ask how prevalent such debris is in the universe and how likely it is to stumble into our solar system. Alternatively, Oumuamua could be a targeted and active probe intentionally sent to our solar system (Loeb & Bialy, 2018). Such an explanation would completely bypass the number density issue as the presence of targeted probes, with their characteristically *non-random* trajectories, cannot be extrapolated to comment on their prevalence generally.

6.2.3. The Tension

These radically different conclusions drawn about the nature of Oumuamua have sparked tension on each side of the debate. The following extracts help to capture the variance in views:

"Assertions that 'Oumuamua may be artificial are not justified when the wide body of current knowledge about solar system minor bodies and planetary formation is considered" (The Oumuamua ISSI Team, 2019, p.600).

"There exists no plausible reason why a technological civilization would build and launch 'Oumuamua type probes of the sort described by Bialy and Loeb (2018)" (Zuckerman, 2022, p.1417).

"Using very conservative probabilities, based on its shape, rotation and luminosity alone, a cometary 'Oumuamua would be a one-in-a-million naturally occurring object. Attempt to

explain its composition so that we can explain its deviation beyond solar gravity by outgassing that was invisible to our instruments and you still have an object that is as rare as one in thousands. But that's not all… 'Oumuamua's spin rate didn't change… Maybe just one in every thousand comets keeps a steady spin… we're now talking about a one-in-a-billion object. Then there's its lack of jerks. If there was naturally occurring outgassing… that's another one-in-a-thousand coincidence, 'Oumuamua is now one in a trillion." (Loeb, 2021, p.86)

"'Oumuamua's anomalies suggest that it might have been a thin craft—with a large area per unit mass—pushed by the reflection of sunlight" (Loeb, 2022, p.1392).

The crux of this discrepancy cannot be attributed to a lack of attention given to Oumuamua's origin. A wave of papers followed Oumuamua's detection in 2017, each proposing a new analysis and interpretation of Oumuamua's unusual features (e.g., Bergner & Seligman, 2023; Loeb, 2022; The Oumuamua Team, 2019; Bannister et al., 2019). Specialist groups were gathered and funded to settle the debate, for example, the ISSI Oumuamua Team. And six years since Pan-STARRS1 first detected the object, the debate continues; the December 2022 issue of *Astrobiology* contained articles titled: "On the Possibility of an Artificial Origin for 'Oumuamua" (Loeb, 2022), alongside "'Oumuamua Is Not a Probe Sent to Our Solar System by an Alien Civilization" (Zuckerman, 2022).

Both sides of the debate have access to the same information regarding Oumuamua. Details of Oumuamua's unusual features are universally agreed and both sides are aware of the other's arguments. Yet no resolution between the sides has been made and there is a real sense of opponents talking past each other. The arguments of one side simply do not land with the other side. Loeb's assertion that a likelihood of around 1 in a trillion for Oumuamua being of non-intelligent origin is left largely unchallenged. There are attempts to show that a non-intelligent explanation is *possible* (The Oumuamua Team, 2019), but this would not assuage Loeb, who concedes that such an explanation is possible — it is just not the more probable explanation.

Disagreement is to be expected, given the uncertainty associated with the data gathered from Pan-STARRS1, alongside uncertainty in the probability of extraterrestrial life existing more generally. But if either side of the debate is to have any hope of changing their opponents' minds, they need to know what it is, exactly, they are disagreeing over. Are false premises at the heart of the divergence in views? Might it be the case that both sides agree on the evidence and background information, but someone is updating their beliefs irrationally? Or perhaps something else, and altogether less problematic, is happening?

6.3. Existing Resolutions to the Debate

Proposed diagnoses of why such a broad disagreement prevails have been published. I will discuss and build upon three recent ones: Cowie (2021, 2023a), Matarese (2022), and Lineweaver (2022b).

6.3.1. Cowie's Undefined Priors

Turning first to Cowie (2023a). Cowie presents a dichotomy between the philosophical style of argument that advocates of Loeb's extraterrestrial hypothesis (LH) can appeal to. These are either 1) a contrastive argument (otherwise stated as an inference to the best explanation), or 2) an eliminativist argument. Cowie argues that Loeb is not justified in making either type of argument in favour of LH, and hence, the misunderstanding within the Oumuamua debate centres on a flaw on Loeb's and his proponents' part.

To focus first on the second style of argument, Cowie argues that employing an eliminativist argument is problematic on account of the problem of unconceived alternative explanations. LH cannot be convincingly argued for by claiming that all non-intelligent explanations have been ruled out, as the history of astrobiology (and science more generally) shows an embarrassment of such premature claims made, only for an unforeseen explanation to be later discovered and upend the previous claim (Dick & Strick, 2005).

Such reasoning is sound; LH cannot be supported via an eliminativist argument. However, it should be noted that the problem of unconceived alternatives undermines both sides of the debate — an eliminativist argument cannot be employed to support LH or the comet hypothesis as unconceived alternatives may stand in opposition to both hypotheses.

Turning now to the first of Cowie's proposed argument styles, the contrastive argument. The threshold of evidence for a contrastive argument is lower than that of an eliminativist argument, and so it initially appears more fruitful for advocates of LH. A contrastive argument only requires that one explanation be *more likely* than the opposing explanation. As such, a contrastive argument cannot be employed to argue that any explanation is definitely correct but rather that it is the most likely correct explanation on the table. Therefore, to conclude in favour of Oumuamua's intelligent origin, Loeb sympathisers need only show that their explanation is more likely than the orthodox one.

However, Cowie (2023a) argues that Loeb is not justified in employing a contrastive argument as he is unable to define his prior probability of detecting extraterrestrial life. Cowie summarises: "We cannot sensibly estimate the prospects for encountering extraterrestrial artefacts in the solar system. So, we lack justification for thinking the prospects are fair or good. Note that the claim being made here is very weak.

It is *not* that the prospects are poor. It is that they are unknown. Yet this very weak claim is still enough to show that one ought not to accept LH given our current information." (2023, p.73). In short, if Loeb cannot fix his prior probability of detecting extraterrestrial life, he cannot arrive at a posterior probability of Oumuamua being of intelligent origin in order to compare to the orthodox explanation.

In this way, uncertainty over the prior probability of detecting extraterrestrial life leaves Loeb unable to employ a contrastive argument in favour of LH. However, such an argument only works when dealing with *total* uncertainty and when all actors agree that this uncertainty is total. Cowie's reasoning may be robust in precluding his own use of a contrastive argument, but it need not preclude Loeb's use of it.

Cowie's pessimistic view of defining the prior probability of detecting extraterrestrial intelligence need not extend to Loeb's reasoning. For Loeb to be justified in employing a valid contrastive argument in favour of LH, only *he* needs to be able to specify his prior probabilities. Loeb is quick to write of how he would give a reasonable value for this: "It is very presumptuous for us to assume that we are the only intelligence in the vast cosmos… [I]t is most likely that we will encounter relics of extraterrestrial technologies before establishing contact with any living civilisation. This must be kept in mind as we contemplate explanations for the mysterious properties of… 'Oumuamua." (Loeb, 2021, p.115). It is apparent, then, that although Loeb does not go out on a limb and specify a particular prior, he values this reasonably high.

Indeed, it is true that even Loeb would associate some uncertainty with his prior probability of detecting extraterrestrial life. However, the above quote indicates that he would not ascribe total uncertainty. He may instead estimate a range (perhaps following a distribution skewed to the higher probabilities). The key point is that, so long as Loeb's priors are not *completely* uninformative, he will be justified in carrying out a contrastive argument. To give an example of this, let us suppose that Loeb believes the probability of detecting extraterrestrial life is somewhere between 0.5 and 0.8. He could then use contrastive arguments with his lower and upper bounds to determine whether LH wins out in either case.

Such subjectivity should not be viewed as problematic. The group of Bayesians called subjective Bayesians endorse just this. They employ Bayes theorem with subjective priors, and hence regard the framework as a means by which individuals can rationally update their priors — whether these priors are well-founded is beyond the remit of the theory, though the priors should still satisfy some minimal requirements such as coherence (De Finetti, 2017; Savage, 1972). Loeb, therefore, is completely warranted in carrying out a contrastive argument to determine which explanation is more likely, according to his own priors.

Moreover, there is a second undesirable consequence of Cowie's wholly undefined priors. If we were to embrace fully undefined priors about the probability of detecting extraterrestrial life, and hence we

accepted that we could not carry out contrastive arguments, we find ourselves unable to make seemingly unproblematic statements. Specifically, we will be left unable to conclude in favour of detecting extraterrestrial life, *regardless* of the evidence presented. Even if we stumbled across the body of an alien organism with entirely non-terrestrial biochemistry, Cowie's argument from scepticism leaves us unable to say anything about our posterior probability that we have detected alien life.

A final flaw in Cowie's argument that undefined priors invalidate Loeb's use of a contrastive argument is that the same logic can be used to invalidate the use of a contrastive argument in favour of the comet hypothesis. By flipping the burden of proof to the proponents of the orthodox explanation, for them to carry out a contrastive argument against LH, they need to explicitly show that their posterior probability is higher than that of LH. However, due to uncertainty in the latter, they would not be justified in making such an assumption. The uncertainty works both ways; indeed, if such complete uncertainty existed, Loeb would be unjustified in asserting that his posterior is higher than the orthodox one. But equally proponents of the orthodox explanation could not claim that their posterior is higher than LH.

Hence, to summarise this review of Cowie's analysis, the problem of unconceived alternatives is, as Cowie points out, detrimental to any argument by elimination. However, an attempt to undermine a contrastive argument by claiming that Loeb's prior probability of detecting extraterrestrial life is undefined falters. For Loeb to employ a valid contrastive argument, he need not utilise any universally agreed upon prior probabilities of detecting extraterrestrial life; he need only employ his own priors. There is, therefore, nothing inherently wrong about Loeb's argument structure. It just might be the case that his opponents do not find his premises convincing; they may disagree with his optimistic prior probability of detecting extraterrestrial life.

Moreover, Cowie's defence of embracing a wholly undefined prior results in the unintuitive outcome of being unable to ever conclude in favour of detecting extraterrestrial life. Finally, even if both sides of the debate were to agree that the prior is completely undefined, and indeed, Loeb would be unable to make a contrastive argument in favour of LH, the proponents of the orthodox explanation would similarly be unable to make a contrastive argument *against* LH. Hence, appealing merely to uncertainty in the priors is not sufficient to undermine LH. Cowie's analysis of the Oumuamua controversy, therefore, appears to not fully capture the rationale of each side and the arguments they employ.

6.3.2. Matarese's Meta-Empirical Arguments

A second analysis of the Oumuamua debate comes from Matarese (2022). This different approach instead focuses on meta-empirical arguments to bolster the *viability* of LH. Matarese points out that additional

evidence about the nature of Oumuamua will likely not surface (unless Project Lyra were to actualise (Hein et al., 2019)). Thus, in this lack of further empirical evidence, a meta-empirical approach may be helpful.

Before diving into Matarese's description of meta-empirical arguments and their application to the Oumuamua debate, it is important to note that Matarese does not identify disagreement over meta-empirical arguments as the crux of the disagreement over Oumuamua's origin. Rather, Matarese argues that meta-empirical arguments can be used to support LH (indeed, both sides may agree with this statement). In this subsection, I will present Matarese's case before appealing to quotes from Loeb to evaluate to what extent meta-empirical arguments would have affected Loeb's position. I will conclude that they would likely have had little impact, as Loeb does not merely think his position is *viable* but rather thinks it is *likely*. Disagreement over the interpretation of meta-empirical arguments will be shown not to be at the heart of disagreement over Oumuamua's origin.

Meta-empirical arguments have been especially championed by Dawid (2006, 2007, 2009, 2013, 2015, 2016, 2018, 2019, 2022). He defines them as "lines of reasoning that aim to generate a significant degree of trust in a theory's viability in the absence of empirical confirmation" (Dawid, 2022, p.1). Dawid identifies three key meta-empirical arguments employed by scientists in the absence of empirical data (Dawid, 2013, 2022). The first of these is the No Alternatives Argument (NAA): scientists are inclined to trust a theory more if no alternative theory is able to account for the observations, despite considerable effort being made to find such an alternative theory. The second meta-empirical argument is coined as the unexpected explanation argument (UEA). This captures the phenomenon whereby scientists tend to trust a theory more if the theory is capable of explaining more phenomena than it was developed to explain. Finally, the meta-inductive argument (MIA) strays somewhat into empirical evidence. It is described as follows: scientists tend to have increased trust in theories which satisfy the NAA and the UEA, if previous theories within the same field, which satisfied the NAA and the UEA, were empirically successful.

As a slight aside, these meta-empirical arguments can find their basis in the vast literature on theoretical virtues. Kuhn's seminal paper "Objectivity, Value Judgment, and Theory Choice" (Kuhn, 1977) outlines five criteria by which candidate scientific theories should be evaluated. Often, empirical data is limited when the promise of a fledgling theory is being assessed, and so theoretic virtues are relevant. Kuhn (1977) identifies the following five criteria: accuracy, consistency, scope, simplicity, and fruitfulness. Some criteria are more empirical than others, e.g., accuracy requires more empirical input than simplicity or (internal) consistency. Additional theoretic virtues have been suggested; for example, in his *Conjectures and refutations: the growth of scientific knowledge,* Popper defends the theoretical virtue of falsifiability and non-ad

hocness (Popper, 2014, Chapter One). See also Schindler's *Theoretical Virtues in Science* for a deep discussion of these virtues and their relation to concrete examples of historical scientific theories (Schindler, 2018, especially Chapter One). These theoretical virtues help form the foundation of the non-empirical arguments of Dawid (2013, 2022) and Matarese (2022). As an example, the unexpected explanation argument (UEA) captures the virtues of scope (Kuhn, 1977) and non-ad hocness (Popper, 2014, Chapter One).

Turning back now to Matarese's application of Dawid's meta-empirical arguments to the case of Oumuamua. Matarese (2022) argues that the first two meta-empirical arguments (the NAA and the UEA) apply to LH and hence increase the viability of the hypothesis. Briefly, The NAA applied to LH states the following: despite a long and careful search for alternatives to LH, none have been found that are more empirically adequate than LH, and this is confirmatory to LH. Concerning the UEA, Mataese argues that: LH is able to account for additional observations that LH was not initially designed to explain, and this is confirmatory of LH. To expand a little on this second non-empirical argument, Matarese argues that LH was initially proposed to account for the excess acceleration of Oumuamua, only for later reflections to show that LH would also explain Oumuamua's shape, colour and high reflectivity (Matarese, 2022, p.16).

Matarese convincingly applies the NAA and UEA to LH. It is apparent that Loeb *could* employ such meta-inductive reasoning to bolster the viability of his position. However, the question arises of whether Loeb's reasoning rests on these meta-inductive arguments in a way that his opponents have ignored. Could it be disagreement over meta-indictive evidence, as opposed to empirical evidence, that is at the root of the disagreement over Oumuamua's origin?

To answer this question, it is necessary to dissect what it is that meta-empirical arguments lend to theories; in what way do they bolster them? Both Matarese and Dawid are especially careful and precise in their language. Mata-empirical arguments are said to increase the trust in or viability of a scientific theory (Dawid, 2022; Matarese, 2022). Other phrases adopted by Matarese are restricted to: LH gaining "epistemic significance" (Matarese, 2022, p.18) in light of meta-empirical arguments; meta-empirical arguments "could be naturally used to support LH" (Matarese, 2022, p.17); and meta-empirical arguments can "check the epistemic status of LH" (Matarese, 2022, p.17).

Matarese is distinctly careful not to mix up 'viable' with 'probable', and this is significant. The above language highlights the limitations of meta-inductive arguments. Meta-inductive arguments, being detached from prior probabilities, are only able to show how capable an explanation is, not how likely it is. By not grounding your explanation with probability assignments for your premises, you can end up with a very tidy and complete explanation, but one that is very unlikely to be true.

An extreme example of this would be a God hypothesis. An omnipotent being could be employed to explain the most complex of phenomena neatly. Meta-empirical arguments could be employed to bolster the viability of such an explanation. A God hypothesis might be capable of explaining things that physics has yet to provide an answer to. Moreover, the God hypothesis is flexible enough to explain ever-new phenomena. However, viable explanations have little external meaning unless they are grounded by reasonable probabilities — in this case, the probability of God existing in the first place.

So, meta-empirical arguments (especially the NAA and the UEA) strive to do something different than empirical arguments. They do not attempt to comment on the likelihood of a theory being correct; rather, they comment on its viability. This is akin to how much explanatory power the theory has, or whether rival theories exist. These intuitively *might* lend to an increase in the probability that the theory is correct, but there is no certainty in this. Hence, the adoption of language like *viable* as opposed to *likely* is important. It is this key distinction that leaves disagreement over meta-empirical arguments as an unlikely cause of the disagreement over Oumuamua's origin.

Loeb recognises that LH is a good explanation – it explains and is consistent with the numerous unusual properties of Oumuamua. If he wished to adopt Dawid and Matarese's terminology, he would likely have no qualms in describing LH as 'viable' and rendering it meta-empirically confirmed by theoretical virtues. However, Loeb goes further and makes a stronger claim. He believes that the empirical evidence is confirmatory to LH. LH is not only viable, but it is also sufficiently probable. Hence, Matarese's analysis is illuminating in addressing a form of confirmation not often discussed in science — that of meta-empirical confirmation — but the suggestion that Loeb employs this to help his case understates Loeb's goal. Proving LH is viable is small fry in comparison to his real goal in proving that it is sufficiently probable.

6.3.3. Lineweaver's Differing Priors

The third and final analysis of the crux of the Oumuamua debate that I will discuss comes from Lineweaver (2022b). Lineweaver frames LH and the orthodox explanation in terms of Bayesian updates and concludes that "the primary source of the 'Oumuamua controversy is disagreement about the prior odds" (2022b, p.1424). I believe this is correct. Lineweaver shows how sensitive posterior probabilities are to the prior probabilities in a Bayesian update. As such, it is because Loeb has a higher prior probability of detecting extraterrestrial life than his orthodox opponents that the two sides disagree in their posteriors.

Further support for this analysis of the Oumuamua debate can be interpreted in the findings of Vickers et al. (2025). Here, my co-authors and I report the results of four surveys of researchers in the astrobiology community. Three of these surveys asked respondents to position themselves on a Likert scale of how likely they think the existence of basic, complex, and intelligent life is. A fourth survey asked non-astrobiologist scientists the first of these questions. General agreement over the likely existence of basic life was found across the board. However, disagreement within the astrobiology community was evident regarding the likely existence of complex or intelligent life. The results of these surveys are given in Figure 6.1.

The data, therefore, illustrate that disagreements over prior probabilities for the existence of extraterrestrial intelligence are extant in the astrobiology community. We should expect, then, that different researchers will arrive at different conclusions when weighing up which hypothesis is more likely, one involving extraterrestrial intelligence and the other not. This supports Lineweaver's (2022b) argument that differing prior probabilities are at the heart of the disagreement over Oumuamua's origin.

I am inclined to agree with Lineweaver (2022b); it is these differing starting points that result in the differing endpoints. It is not that the opponents disagree on how confirmatory the evidence is to either position, but rather that the Bayesian updates being carried out start and thus end in different places. However, Lineweaver makes a further prescriptive claim on what value we should assign to our prior probability of other intelligent, technologically advanced civilisations existing elsewhere in the universe. By considering the evolution of the millions of lineages of life on Earth besides ours and the lack of any of these producing human-like technological intelligence, Lineweaver concludes that "the probability of technological alien civilisations is somewhere between zero and tiny" (Lineweaver, 2022b, p.1424). A deeper discussion of various attempts at pinning down the probability of civilisations becoming technologically advanced can be found in Chapter Five of this thesis — one central criticism of this conclusion is that it assumes Earth-like evolutionary constraints apply universally, which may not be justified given our single data point.

Consequently, Lineweaver (2022b) may be read as advocating for the orthodox explanation, as opposed to LH being the more justified one. It is this sense of objectivity that I would like to push back on. We should keep open the idea of the scientific community at large embracing differing priors as a means of ensuring that all possible explanations are explored whilst individuals can maintain individual rationality. However, this is a discussion I would like to return to in §6.5. At present, let us refocus on describing the root of the debate: the differing priors within a Bayesian update.

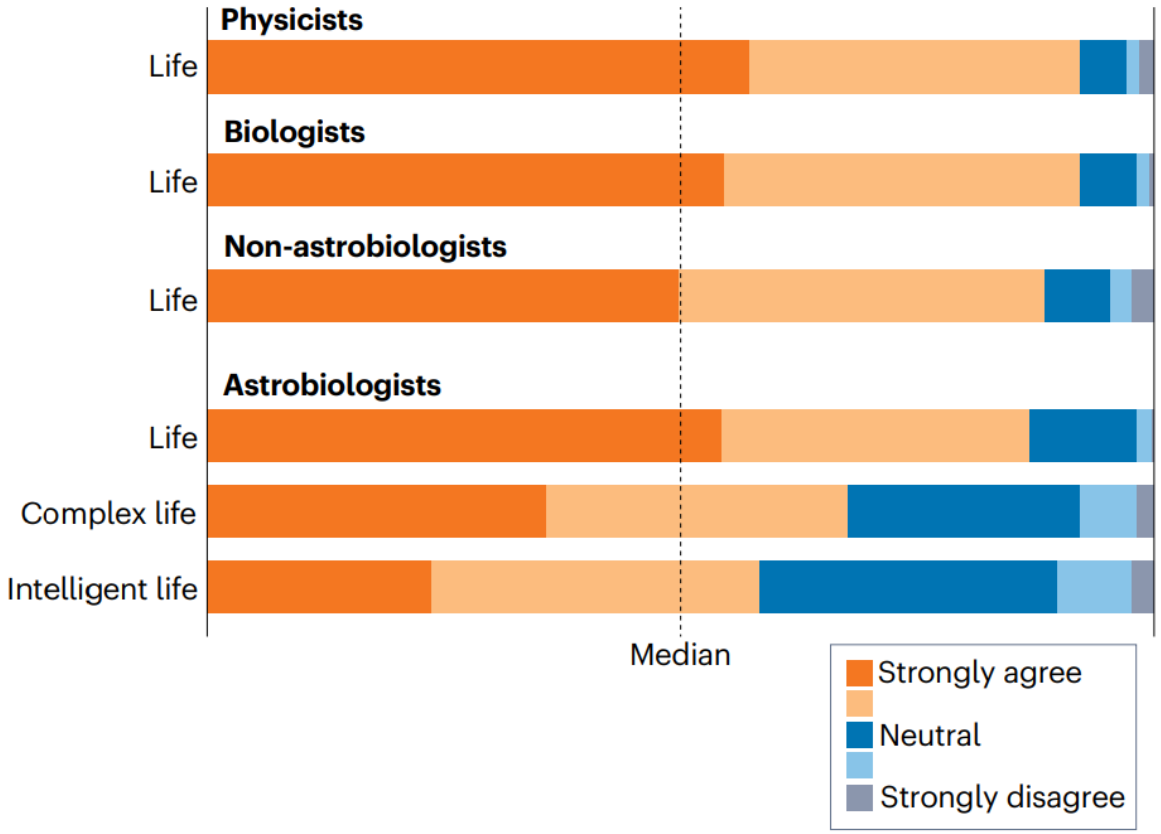


Figure 6.1. Results from four surveys into community opinion regarding the likely existence of extraterrestrial life (Vickers et al., 2025, p.17). The survey question corresponding to "Life" was: *It is likely that extraterrestrial life (of at least a basic kind) exists somewhere in the universe.* The survey question corresponding to "Complex life" was: *It is likely that extraterrestrial organisms significantly larger and more complex than bacteria exist somewhere in the universe.* The survey question corresponding to "Intelligent life" was: *It is likely that extraterrestrial organisms with advanced cognitive abilities comparable to or superior to those of humans exist somewhere in the universe.*

6.4. The Pivot of the Differing (Yet Defined) Priors

6.4.1. The Bayesian Framework

Lineweaver (2022b) provides an account of the Bayesian update involved when integrating evidence from Oumuamua. I will give my own account of this below using a slightly different formulation of Bayes' formula from Lineweaver's. First though, I will briefly motivate the application of the Bayesian framework to new evidence in astrobiology. Employing Bayesian statistics ensures that an individual updates their credences (upon learning of new information) rationally. This is to say that they will end up

with posterior probabilities that comply with the axioms of probability, which is a desideratum beyond dispute.

Bayes' theorem is employed in cases with probabilistic evidence. It provides a means of rationally updating, with an individual's prior beliefs providing an anchor. Bayes' formula is often given as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

where *A* and *B* are events. *P(A)* and *P(B)* are the probabilities of *A* and *B* occurring respectively and $P(A|B)$ is a conditional probability – the probability of *A given* that *B* has already occurred. Similarly, $P(A|B)$ is the probability of *B given* that *A* has already occurred. The probabilities on the right-hand side represent the priors and hence are probabilities assumed to be known. The conditional probability on the left-hand side is our posterior and is the probability to be determined.

Let us now apply Bayes' theorem to Oumuamua to evaluate the posterior probability that we have detected a light sail *P(L)*, given the evidence, *E,* observed from Oumuamua. This would be equal to the probability of the evidence, given that we have detected a light sail, $P(E|L)$, multiplied by the prior probability of detecting a light sail *P(L)*, all divided by the probability of the evidence, *P(E)*. Hence our posterior probability will be given by:

$$P(L|E) = \frac{P(E|L)P(L)}{P(E)} = \frac{P(E|L)}{P(E)}P(L).$$

Now, we are presented with a concise formula in which, to output powerful information about the probability of Oumuamua being a light sail, we need only substitute in our priors.

The posterior probability is shown to be sensitive to two key probabilities: the probability of the evidence, given the detection of a light sail, $P(E|L)$, divided by the probability of the evidence, *P(E)* and the prior probability of detecting a light sail in the first place, *P(L)*. The first of these is a measure of how confirmatory the evidence is to the hypothesis *L*. However, it is crucially grounded by the prior probability of detecting a light sail in the first place. The light sail hypothesis may be highly confirmed by the evidence (the quotient is high), yet if the prior probability of detecting a light sail is too low, the posterior will not sufficiently increase.

Alternatively, if the prior is non-negligible and the evidence is highly confirmatory, the posterior probability can be significant. So how is this prior probability determined? In practice, such a probability is only arrived at individually through conjecture and estimation. It is likely to vary greatly from one

individual to the next. It is disagreement over priors such as this that is largely responsible for the disagreement in Oumuamua's origin.

6.4.2. Divergent Priors Leading to Divergent Posteriors

It has so far been shown that the probability of any hypothesis regarding Oumuamua's origin is dependent on 1) how confirmatory the evidence is of the hypothesis and 2) the prior probability of the hypothesis. *Both* of these features are important, and I argue that both of these are actually being employed by the various players involved in the Oumuamua debate (whether knowingly or not). I do not think that either side of the debate is leaning on meta-empirical arguments in the absence of empirical arguments (Matarese, 2022). And I do not think it helpful, nor necessary, to assume total uncertainty over the prior probability of detecting extraterrestrial life (Cowie, 2023a).

Of the two features constraining the posterior probability, it is the prior probability of detecting a light sail that may present more disagreement. Indeed, how confirmatory the evidence is of a particular hypothesis is something that individuals could reasonably agree on. However, the prior probability of detecting a light sail is likely difficult to objectively pin down due to a lack of data. It is clear, though, that Loeb ascribes a reasonable value to this as the previously cited quotes show, e.g., "Given so many worlds – fifty billion in our own galaxy! – with similar life-friendly conditions, it's very likely that intelligent organisms have evolved elsewhere" (Loeb, 2021, p.50).

In contrast, the reasoning of advocates of the orthodox explanation implies a much smaller prior probability of detecting extraterrestrial life. The methodology of this side is to assume intelligence as the last resort explanation, indeed a recent report by researchers at NASA concluded that "extraterrestrial life itself must be the hypothesis of last resort" (Spergel et al., 2023, p.25). The extraterrestrial hypothesis can only be taken seriously when all other explanations have been shown to be implausible. This rationale is consistent with an implied vanishingly small prior probability for detecting extraterrestrial life. As such, both sides of the debate may carry out a Bayesian update with the same value for how confirmatory the evidence is, yet Loeb ends up with a posterior for LH that is larger than that for the orthodox explanation, whilst the reverse is true for his opponents.

Finally, it is noteworthy that Loeb does not need a significant or even likely posterior for LH. He only needs it to be larger than the posterior of the orthodox explanation; it need only have a comparative advantage. Moreover, if Loeb's goal is only for the community to take LH as a serious contender, he may only need his posterior to come close to the orthodox posterior. Notably, these posteriors can be very small.

6.5. The Problem of the Priors?

6.5.1. Subjective vs. Objective Bayesian Epistemology

'The Problem of the Priors' is an issue discussed widely within Bayesian epistemology (see, for example, Barrett, 2014; Lin, 2022; Pettigrew, 2016; Press, 2009). It highlights the dependence of any posterior probability on the inputted prior probabilities: it is all very well having a means for integrating new information in compliance with Kolmogorov's axioms of probability, but how does one determine the initial priors? The problem has initiated the carving of Bayesians into two camps: the subjective Bayesians and the objective Bayesians. Both employ Bayes' theorem, though they differ in what they claim the theorem can provide.

Subjective Bayesians only use the theorem to ensure that they update rationally. They relinquish the requirement for objective probabilities in preference for subjective ones. Prior beliefs need not reflect the *real-world* probabilities; rather, priors are determined by one's subjective beliefs. These subjective beliefs must still comply with the axioms of probability, but an individual's prior beliefs need not map onto the real world. Consequently, the corresponding posterior probabilities need not map onto the real world. Hence, subjective Bayesian updating does not strive to provide objective information about the world, but rather solely provides a means for an individual to rationally update on their existing subjective beliefs. In contrast, objective Bayesians strive to use well-motivated priors that track the real world. As a result, an objective Bayesian aims to have posteriors that accurately reflect reality.

Given the underdetermination of priors in astrobiology, the above discussion of the case of Oumuamua describes a subjective Bayesian story. It affords rationality to both sides of the debate whilst condoning differing, subjective priors. Both Loeb and his opponents may well be updating in a Bayesian manner while nonetheless arriving at different conclusions. They could both have updated in compliance with Kolmogorov's axioms of probability (by way of complying with Bayes' formula), it is just that their inputted prior probabilities for detecting extraterrestrial life differ. Hence, to summarise, the disagreement in the origin of Oumuamua is not due to individual irrationality or a misunderstanding of how confirmatory the evidence is to any hypothesis, but rather due to differing priors.

6.5.2. Collective Rationality

Thus far, it has been argued that both Loeb and his opponents may be updating rationally upon their (differing) prior beliefs. The argument that either side is being irrational is hence criticised. However, the question may arise of whether this divergence of priors leading to a divergence of posteriors poses a

problem for the rationality of the scientific community *as a whole*. The answer to this, I will argue, is that disagreement in the priors might actually result in a more optimal collective scientific structure. A community of disagreeing individuals may collectively be more rational (broadly construed) than a community of like-minded individuals.

The rationale for this is that the existence of agents with unorthodox prior beliefs helps to ensure that all avenues are honestly explored. Loeb's sufficiently high prior results in his genuine exploration of his hypothesis. Similarly, the advocates of the orthodox explanation are motivated to explore and develop the comet hypothesis, and all of this drives science forward. In a field where the prior probabilities are likely poorly defined, it is a good thing that there exists fervent disagreement, as it means a range of posterior hypotheses are investigated.

An interesting parallel to this overall benefit of individual disagreement can be seen in the partial migration of birds. Many bird species exhibit this partial migration, whereby some members of the species migrate to warmer climates, whereas some stay behind. This diversification reduces the risk of the entire species being wiped out due to an unpredictable environmental disaster, as well as allowing the species to exploit multiple food resources (Kaitala et al., 1993; Chapman et al., 2011).

Similar ideas to this diversification in the context of optimal scientific community structure have been discussed (see, for example, Kitcher, 1990; Zollman, 2007; Avin, 2019; Goldman & Shaked, 1991; Weisberg, & Muldoon, 2009; Muldoon, 2013). Many of these community models are agent-based and represent scientific knowledge as a landscape with the agents (the scientists) exploring and discovering the landscape (in particular, Weisberg & Muldoon, 2009; Avin, 2019). A common theme arises in these models that a diverse community regularly outperforms a homogenous one. The importance of 'mavericks' (Weisberg & Muldoon, 2009) in exploring unorthodox and unpopular ideas is defended; a community solely comprised of followers who share the same beliefs and methodologies is less likely to exhaust the epistemic landscape.

It is significant and interesting that a community of agents with different priors likely preserves *both* individual rationality *and* collective rationality. Other models of optimal community structure inevitably sacrifice the former for the latter, or at the very least, external influences must be brought in to make the options that are best for the community also the most attractive to the individual. Kitcher's (1990) and Strevens' (2003) reward system is a paradigmatic example of this. Here, funding incentives for exploring unchartered epistemic spaces are suggested to ensure fringe projects are given due credit — although the collective success of such reward systems has since been questioned (Heesen, 2019).

Alternatively, if the existence of the kind of varied prior probabilities seen in the Oumuamua debate were encouraged, instead of regarded as problematic, individual scientists would be internally motivated to research their own fringe project. They would be exhibiting individual rationality in their choice of research project. Moreover, this may lead to collective rationality as a more optimal scientific community emerges as a result.

These models support the claim that mavericks such as Loeb, with differing priors that result in the advocation and exploration of different hypotheses, are an important part of an optimal scientific community. Science benefits because of disagreement in priors. This result is expected to be even more supported when the epistemic landscape in question is the highly unexplored and vast field of astrobiology.

It is important for the community that even these rogue agents update their priors rationally to ensure that a hypothesis is abandoned if the evidence sufficiently disconfirms even the most generous prior probability assignment of it. In this way, it is expected that many rogue theories will fall by the wayside as disconfirming evidence continually reduces the posteriors held by these rogue agents.

To reiterate, all agents should ideally agree on how confirmatory a piece of evidence is to a particular hypothesis. So, given enough updates on a prior, all posteriors should tend to the same value. But *some* rogue theories just might continually get confirmed, and it is the mavericks that show early support of these theories that will seek their confirmatory evidence.

In summary, so long as there is reasonable agreement on the degree of confirmation for any piece of evidence, differing priors are not only not a sign of individual irrationality, but they may be at the heart of ensuring a diverse scientific community that is more successful at expanding collective knowledge.

6.6. Conclusions

This chapter has revisited the Oumuamua debate with a view to 1) identifying the crux of the disagreement about the origin of Oumuamua as being differing priors and 2) making a normative claim about the benefit of these differing priors. Regarding the first of these, previous diagnoses of the Oumuamua debate are somewhat lacking. Cowie's (2023a) argument of undefined priors invalidating a contrastive argument falls down in three keyways. Firstly, Loeb need only employ his own priors. Secondly, embracing undefined priors would leave us unable to conclude in favour of life, regardless of the evidence. Thirdly, Cowie's argument could be flipped and levelled against the orthodox position with just as much force.

A second analysis paper was then discussed, and this was Matarese's (2022) application of meta-empirical arguments to support LH. Although Matarese makes a strong case that meta-empirical arguments may be applicable in the case of LH, it is unlikely that disagreement over the application of meta-empirical arguments is at the heart of the disagreement over Oumuamua's origin. This is because Loeb does not believe LH is merely viable; he believes LH to be *likely*.

Finally, Lineweaver's account of the debate has been discussed and found to correctly highlight the crux of the disagreement to be differing prior probabilities. The claim that astrobiologists may hold differing prior beliefs about the existence of extraterrestrial intelligence is supported by the results of a survey of the astrobiology community (Vickers et al., 2025). I have expanded on Lineweaver's position within a Bayesian framework in the context of Oumuamua — specifically splitting up the confirmatory weight of the evidence and the prior probability of detecting extraterrestrial life. Although it might be expected for both sides to agree on how confirmatory the evidence is, it is expected that individuals' prior probabilities for detecting extraterrestrial life vary wildly (Vickers et al., 2025). It is this that causes the disagreement about Oumuamua's origin. This analysis extends beyond the Oumuamua case. In astrobiology more broadly, disagreements over the likelihood of biosignatures or technosignatures on exoplanets might stem from similarly divergent priors.

Finally, the problem of the priors so prevalent in Bayesian epistemology has been shown to be not problematic at all when considering the vastly unexplored epistemic landscape of astrobiology. In fact, the benefit of these differing priors in ensuring a more thorough exploration of possible hypotheses has been motivated by way of previous work on community modelling (e.g., Avin, 2019; Weisberg & Muldoon, 2009; Muldoon, 2013). It is now to a deeper dive into the promise and pitfalls of these community models that this thesis will turn.

Chapter Seven

The Promise and Pitfalls of Community Modelling in Astrobiology

The preceding chapter has exemplified how impactful differing prior probabilities are to the conclusions that scientists draw. Moreover, I have made the case for these differing priors (and hence differing conclusions) being advantageous to the field of astrobiology. This final chapter elaborates on this claim and analyses the role of agent-based models in informing us of the value of diversity in science.

I present the results of a simple agent-based model as a cautionary tale of how tempting it is to overinterpret such models when making real-world conclusions. Following this, I analyse three models whose conclusions concerning the value of diversity are widely cited as having real-world applicability (Avin, 2019; Hong & Page, 2004; Weisburg & Muldoon, 2009). The conclusions of each of these models are shown to result from working instrumental model features that fail to represent real-world features. I argue that, although diversity is likely an invaluable feature of any scientific community, the models considered do not convincingly show this.

Subsequently, I evaluate how and why computer models (both simple and complex) can be informative in science. By borrowing from the scientific structural realism literature, I argue that computer models can be accurate to their target system, even when highly idealised. This is just so long as there is a close correspondence between the functional form of the relevant working features of the target system and the functional form of the working features of the model.

7.1. Introduction

By way of Oumuamua as a case study, the previous chapter has argued that disagreement over fundamental priors in astrobiology may be at the heart of why differing conclusions are drawn by different researchers. Such disagreement may even exist when researchers completely agree on how confirmatory

the evidence is of any particular hypothesis. It was then suggested that this disagreement over fundamental priors (like the probability of abiogenesis or the probability of life becoming complex) may introduce variety into the field by spurring some researchers to take on rogue projects. A speculative suggestion was then made that such disagreement would be beneficial to the collective output of the field of astrobiology. The first aim of this chapter is to report the result of a novel agent-based model that aims to formally test this hypothesis.

§7.2 discusses the benefits that community models could have for the field of astrobiology. Following this, §7.3 reports the results of a novel agent-based model created in NetLogo (Gillen, 2025a). The objective of the model was to evaluate whether disagreement over the probabilities of projects succeeding leads to more collective utility being garnered for a community of researchers, compared to when there is agreement on an estimated prior. The model accounts for the inherent uncertainty surrounding many projects in astrobiology. Hence, unlike other models in the literature where scientists know the expected utilities, my modelled scientists work with estimates of these.

The results of the novel model show that scientists who disagree over their assignments of the expected utilities of potential projects collectively produce more total utility than scientists who agree on the expected utilities of potential projects. This conclusion could be adopted to argue that heterodox scientists who hold unpopular prior probabilities benefit science as a whole. Specifically, I show how the results of the model might be leveraged to argue that Avi Loeb's relatively large prior probability for detecting extraterrestrial life helps to maximise the output of astrobiology.

Such leveraging, however, is problematic, and this is, in part, on account of the analytic nature of models, which is discussed in §7.4. Hence, the remainder of this chapter works to undermine the informativeness and real-world applicability of my model's results and the results of three other community models (Avin, 2019; Hong & Page, 2004; Weisburg & Muldoon, 2009). The purpose of this is to exemplify how alluring community models can be and how tempting it is to overinterpret them. The conclusions drawn about my model are unpicked in §7.5 and found to be baked into the model itself. My model, therefore, stands as a motivating example of how easy it is to overinterpret community models.

Similar conclusions are drawn regarding the three popular and widely cited epistemic models I consider (Avin, 2019; Hong & Page, 2004; Weisburg & Muldoon, 2009), all of which make bold claims about the structure of science. I go through each of these models in §7.6, §7.7, and §7.8 to unpick which assumptions are actually driving the conclusions. In every model (including my own), it is found that some version of diversity is the driving factor to optimality, and this is explored in §7.9. However, I argue that the way in which success is defined within each model necessitates the type of diversity tested. Each model, therefore, is somewhat reductive and ultimately uninformative about the real world. More

troubling, though, I argue that a lack of real-world analogues to instrumental features within each model provides a critical flaw in the epistemic models considered.

Following this, I take stock of the lessons learned from my model and the three others considered. I address the question of how and why computer modelling of communities in science can be successful. §7.10 defends a space for simple models when the target system is sufficiently narrow. This is exemplified by a novel simple model, based on Condorcet's jury theorem (De Condorcet, 2014), whose assumptions map to real-world referents when they are applied in certain limited contexts. §7.11 then considers how to maximise real-world representation in more complex models. Introducing more real-world data into our models is one way to do this. Hence, mid-level models (such as Harnagel, 2019), whereby landscapes are more realistic and informed, are limitedly endorsed. Having said this, sometimes the target system is too complex to recreate, and hence simplified models are needed. §7.12 tackles how to do this by situating the discussion of model accuracy within the scientific realism debate. The adoption of a structural realist view is consistent with the claim that models need not perfectly mirror the world to be truly informative; there just needs to be a close correspondence between the functional form of the working part of the model and the functional form of the working parts of the target system. The conclusions of the chapter then follow in §7.13.

7.2. The Value of Epistemic Community Models in Astrobiology

Kitcher remarked that it would be highly surprising if the present set-up of the social structures of science were optimal (Kitcher, 1990, p.22). It is an intriguing concept to capture and replicate the structure of scientific communities in order to model and ultimately learn about them. Community models, therefore, attempt to distil the fundamental dynamics of a real-world community to see the effect of modifying different parameters.

Models of this sort were prevalent during the COVID-19 pandemic. Modelling how features like masks or lockdowns affect the spread of the virus was a central tool in informing public health policy (Eikenberry et al., 2020; Agrawal et al., 2021; McBryde et al., 2020). Such attempts to utilise these models were in part because community models are particularly good at parameter sensitivity testing, e.g., evaluating to what extent the basic reproduction number is affected by the percentage of mask wearers.

Community models, though, can be far broader in their scope. The seminal community models of Kitcher (1990, 1993) and Strevens (2003) exemplify the reach of these models well. Each of these attempts to capture the dynamics of agents (modelled as scientists) moving about and exploring a landscape. The landscape represents the wealth of scientific research projects waiting to be carried out. These models

aim to optimise how we should divide cognitive labour to maximise scientific output. A key conclusion of Kitcher (1990, 1993) and Strevens (2003) is that external rewards for being the first to discover something novel help to maximise collective and individual scientific output.

Since Kitcher and Strevens, ever more ambitious community models have emerged to advise on how to optimally construct our scientific and political communities (see, e.g., Avin, 2019; Harnagel, 2019; Weisburg & Muldoon, 2009; Hong & Page, 2004). Conclusions drawn from these models span from the advocation of more randomised funding in science (Avin, 2019) to the suggestion that diversity is more important than ability when it comes to optimising an electorate's voting decision (Hong & Page, 2004). The utility of these and other models has been the subject of much debate (e.g., Rosenstock et al., 2017; Thomson, 2014) and will be addressed in detail in §7.10 of this chapter.

Despite valid concerns, community modelling offers an exciting opportunity to learn ways to optimise societal structures for relatively little cost and time. Due to their proficiency in parameter sensitivity testing, fields with complex dynamics should be particularly well-suited for community modelling. This is because models cannot tell us anything inherently new about a system. Rather, they can illuminate how the parameters of a system affect each other over time. The more complex a system, the harder it is to spot how the parameters affect each other. And hence more unforeseen and surprising conclusions can be drawn from modelling complex systems, as opposed to simple ones.

The field of astrobiology may be modelled as one such complex system. Astrobiology is a vast field with a wealth of promise under a canopy of uncertainty. The potential payoffs of astrobiology research projects are huge, but knowing which projects will generate revolutionary results and which ones will end up in a lab bin is difficult. Moreover, *epistemic* community models that focus on the knowledge state of agents in the model are well poised to account for preferences, disagreements, communication, and other social features. Such features are no doubt consequential to the outputs of any community, including the community of astrobiologists. Community models may, therefore, shed some light on some of astrobiology's inherent unknowns and help inform on how the research efforts of astrobiologists can be maximised.

7.3. Results of a Novel Model Testing the Value of Disagreement in Astrobiology

I present here the results of a novel model in NetLogo (Gillen, 2025a) that tests whether a population of scientists who disagree over their expected utility assignments for projects outperform a population of scientists who agree on their expected utility assignments. The conclusion is that disagreement over the

probability of a project paying off leads to more total utility being harvested by the community over time. Such a conclusion might be leveraged to support the suggestions of Chapter Six: that scientists such as Avi Loeb, who ascribe unusually high probabilities to the detection of extraterrestrial life, may be beneficial to the astrobiology community.

The two-dimensional landscape consists of patches, with these patches representing research projects. For example, one patch may represent the research project of directing JWST at K2-18 b and searching for signs of life. Each patch is assigned a payoff value at random (between 0 and 100), which represents the scientific value of the project, should it provide the results we hope for. In the case of K2-18 b, if no signs of life are detected, the community would not get the payoff associated with that patch. Of course, this is an oversimplification of the multifaceted and multi-aimed nature of astrobiology projects. But, for the sake of simplicity, my model takes success to be binary; either the project succeeds, or it does not.

The patches (projects) are also assigned a risk value that captures the probability of the project paying off. This is a value between 0 and 1 and is given randomly to each patch. The payoff and the risk are multiplied together to give the expected utility of each patch.

The scientists then populate the landscape, being seeded randomly. With each tick (each update in the simulation), scientists are modelled to complete their project, represented by the patch they are standing on. Each scientist then looks at their eight surrounding neighbouring patches and moves to the patch which they perceive to have the highest expected utility. At the end of each tick, the sum of the expected utilities is recorded.

Crucially, there are two types of scientists: those who disagree on the risk assignments and those who agree. Those who disagree on the risk assignments will individually assign risk values to their eight neighbouring patches by way of random selections from normal distributions, with the actual risk assignments as the means. As such, each scientist will view each patch with a slightly different probability of paying off. Consequently, their perceived expected utility for each patch is slightly different. These scientists are taken to represent a community whereby each scientist has their own opinion about the probability of any particular project succeeding.

In contrast, the scientists who agree on the risk assignments all assign the same risk values to the patches. These risk values are not the absolute risk values, but rather decided globally by way of randomly picking from normal distributions, with the means as the actual risks for each patch. There is, therefore, a subtle but essential difference to the disagreeing scientists. Both agreeing and disagreeing scientists do not have access to the exact, actual risk, but rather an estimate of this picked from a normal distribution around

the actual risk. This is a realistic assumption of the model, as we might expect scientists to be reasonably proficient in ascertaining risks, but we should not expect them to determine the exact risk correctly.

The disagreeing scientists, on the other hand, all *individually* pick their risk from the normal distributions. This contrasts the agreeing scientists, who all use the same values picked from the normal distributions. The result is that, for disagreeing scientists, one patch may look very attractive to one scientist, whereas it looks unattractive to another. Such a result cannot happen with the agreeing scientists, who will all agree on the expected utilities of each patch.

To give a scaled-down example of this to clarify this feature of the model, let us imagine a 3x3 section of the landscape, with patches numbered 1 through to 9. An agreeing agent seeded on the central patch, patch 5, will look to its eight neighbouring patches and see the exact same expected utilities as any other agreeing agent would. However, a disagreeing agent on the central patch will assign its own risk and hence expected utilities to its neighbouring patches. Therefore, a nearby disagreeing agent will see the same patches with different expected utilities; patch 1 may have an expected utility of 50 to the first disagreeing agent and 70 to the second disagreeing agent. This difference in risk assignments is what leads to a difference in which patches (representing projects) get selected.

Each project in the landscape also has the same duration (one tic). This is in contrast to some other models (e.g., that of Avin, 2019) but in alignment with Weisburg and Muldoon's model (2009). This is not a particularly limiting assumption within the model, as the duration cost can alternatively be somewhat built into the payoffs. What this would mean is that the payoffs of patches in my model have already been moderated to subtract the time and resource cost of completing the project.

A final feature of the model is that, in line with some previous models (Avin, 2019; Pöyhönen, 2016), the payoff of a patch reduces to zero once a scientist has visited it. This is intended to represent the real-world feature of how the repetition of an already completed project has little value.

The results of my model are shown in Figures 7.1 and 7.2. We can see from these graphs that, for a community of agreeing scientists, the total utility produced drops off much more sharply than for a community of disagreeing scientists.

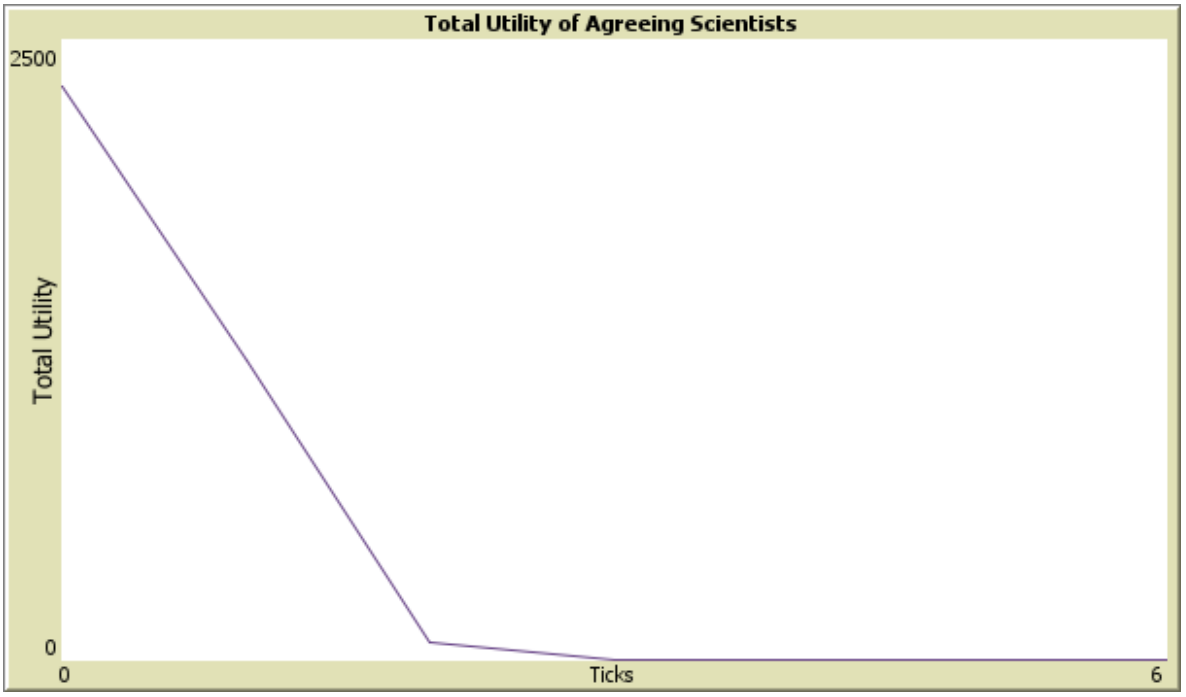


Figure 7.1. The total utility harvested in a community of agreeing scientists over time. Note that the x-axis titled "ticks" represents one update in the model, whereby scientists move about the landscape, and is hence analogous to time (Gillen, 2025a).

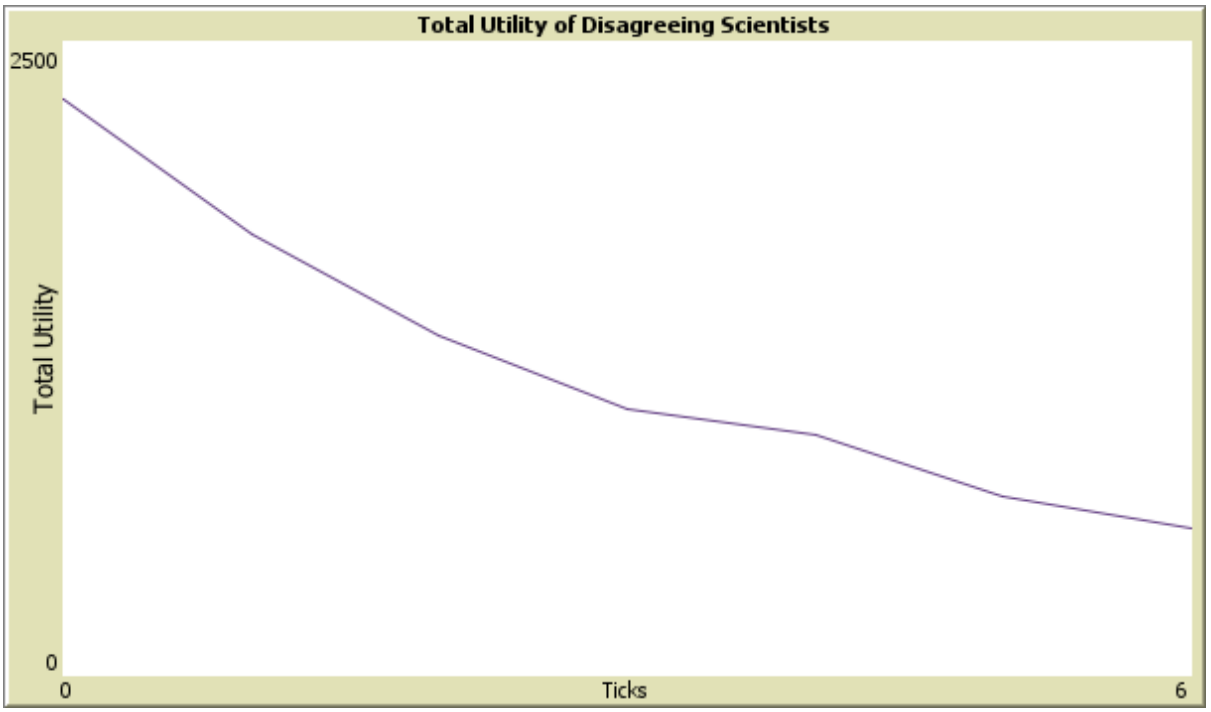


Figure 7.2. The total utility harvested in a community of disagreeing scientists over time (Gillen, 2025a).

The results of this model appear to work in favour of disagreement over prior probabilities in a field like astrobiology. The scientists who had their own opinions about the likelihood of any particular project paying off produced more collective output than the scientists who agreed on a project's probability of succeeding. It is easy to imagine such a result being presented in favour of rogue scientists in astrobiology. However, I would call for extreme caution with such an inference. The results of my model are not as they appear! They stand as an example of how to lie with statistics (to paraphrase Darrell Huff's book with the same title, 2010).

With this in mind, the present chapter will address the issue of overinterpretation within community modelling. The limitations of community models will be outlined generally before I explore what is responsible for my model's conclusions (and it is not that my model has latched onto some real-world truth about the value of disagreement, as it might appear). The internal workings of three other popular community models (Avin, 2019; Hong & Page, 2004; Weisburg & Muldoon, 2009) will be unpicked to further support my claim that often community models are overinterpreted and their real-world utility is limited.

7.4. The Analytic Nature of Community Models

Before delving into the workings of my model's conclusions and the conclusions of three other well-cited community models, a more general discussion of what we can expect to learn from these models is needed. Community models, in all their flavours, are inherently analytic regarding the knowledge they produce. They aim to represent real-world systems by puzzle-solving with the boundary conditions and rules they have been given. In short, models solve complicated mathematical problems that could, in theory, be done with pen and paper.

The reason why models are so useful, though, is that many real-world systems are highly complex, with numerous parameters that may form a chaotic system over time. Solving for such complexity would be difficult without a computer churning through the code for you, but crucially, it would not be impossible (even chaotic systems are deterministic).

Having said this, all that community models do is apply the inputted rules to the inputted boundary conditions. Hence, the results of any community model cannot tell us anything new about the world that could not already be derived from delving into the consequences of these inputted rules and boundary conditions. Consequently, knowledge gleaned from community models is analytic in nature, not synthetic; the conclusions are contingent on and fully derivable from the inputs.

This is not to say that the knowledge derived from community models is somehow useless. The knowledge is derivative in the strict sense, but it is often still not something we would have figured out without the help of a model. Community models can proficiently run through the inputted rules to take the system to an unexpected end with a speed and accuracy unmatched by humans.

Scenarios in which computer modelling is highly relevant and informative are easy to come by. One example might be the modelling of cars on a new road network to determine where a potential bottleneck might be. The model would merely provide consequences of the inputted assumptions (such as the number of cars, driver behaviour, average speed of cars, etc.). Another, perhaps more contentious, example of wide-use modelling is in climate change. It is of note that the contents of the model do not necessarily have real-world referents, and climate models wrangle with this. For there to be a close correspondence between the model's conclusions and real-world conclusions, climate models must incorporate vast amounts of information (Lupo et al., 2013; Bader et al., 2008).

So, the information taken from models is wholly dependent on the information we put into them. Nevertheless, they may be illuminating and even surprising. With this view then of what type of knowledge models can generate and what type of knowledge they cannot, we now turn to a detailed analysis of the workings of mine and three other community models.

7.5. Unpicking the Results of a Novel Model of Disagreement in Astrobiology

Let us now return to the findings of a novel model testing the value of disagreement in astrobiology (Gillen, 2025a). In §7.3, I reported that, within the model, more total utility is harvested over time for the community of disagreeing scientists than for the community of agreeing scientists (Figures 7.1 and 7.2). However, extending this conclusion to real-world communities would be a stretch. The reason for this is that the structure of the model is such that the disagreeing scientists have an unrealistic advantage due to a strange artefact of the model. This artefact is responsible for deriving the conclusion, and yet it has no real-world analogue; hence, the model does not capture any real-world phenomenon.

This working artefact is the tendency for agreeing scientists to collect on the same patch more often than disagreeing scientists. Teaming this with the fact that the total utility, each tick, is just the sum of the expected utilities of the occupied *patches* (not the sum of the total utilities associated with each *scientist*) means that the agreeing scientists are at a distinct disadvantage. To summarise, if multiple scientists are on the same patch, the expected utility only gets counted once. Hence, a community that tends to collect on patches will be at a disadvantage compared to a community that disperses.

It should be noted that the agreeing scientists do not all collect on the same patch over time, because the payoff of a patch reduces to zero once it has been visited. Rather, each tick, two or three scientists often jump onto the same high-payoff patch as they all see the same patches as attractive, and this is enough to disadvantage the agreeing scientists.

It is not a wholly unrealistic feature that the expected utility of a patch is only counted once, even when multiple scientists are on it. In the real world, having *n* independent researchers complete the same project does not result in *n* times as much payoff. However, it is an unrealistic model assumption that agreeing scientists will just blindly replicate each other's results just because they agree that the project is worthwhile. One notable exception to this was Newton and Leibniz's independent, simultaneous discovery of calculus in the 17th century. Both are credited with the discovery, but the benefit to the community is not duplicated. We might wonder what either of these might have discovered during this time if they had focussed their attention elsewhere.

Having said this, the sharper decline of the total utility for the agreeing scientists compared to the disagreeing scientists is wholly an artefact of this feature, whereby the expected utility of a patch is only counted once. Removing this feature completely removes the model's conclusions in favour of disagreement. The complete disappearance of this effect can be seen in Figures 7.3. and 7.4. The theoretical reason for the disappearance of this effect, once the feature of patches only being counted once is removed, is because disagreeing scientists assign their individual risks to patches by way of a normal distribution about the actual risk of the patch. As such, on average and over time, the expected utilities of the patches selected by the disagreeing scientists will roughly match the expected utilities of the agreeing scientists.

This novel model, therefore, serves as a cautionary example of how easy it can be to hastily extend model conclusions to scientific communities. Any model of a real-world system is bound to be a simplification. Indeed, being a simplification in and of itself does not invalidate the extension of the model's conclusions to the real world (this will be discussed in depth in §7.10). However, if the working parts of the model simply have no real-world referent, then the model's conclusions have no real-world referent either. Attention must be paid to the inner workings of these community models so as to discover any unrepresentative quirks that play a crucial role in deriving the model's conclusions. If these quirks do not accurately capture any real-world feature, the conclusions cannot be said to inform us of any real-world truth.

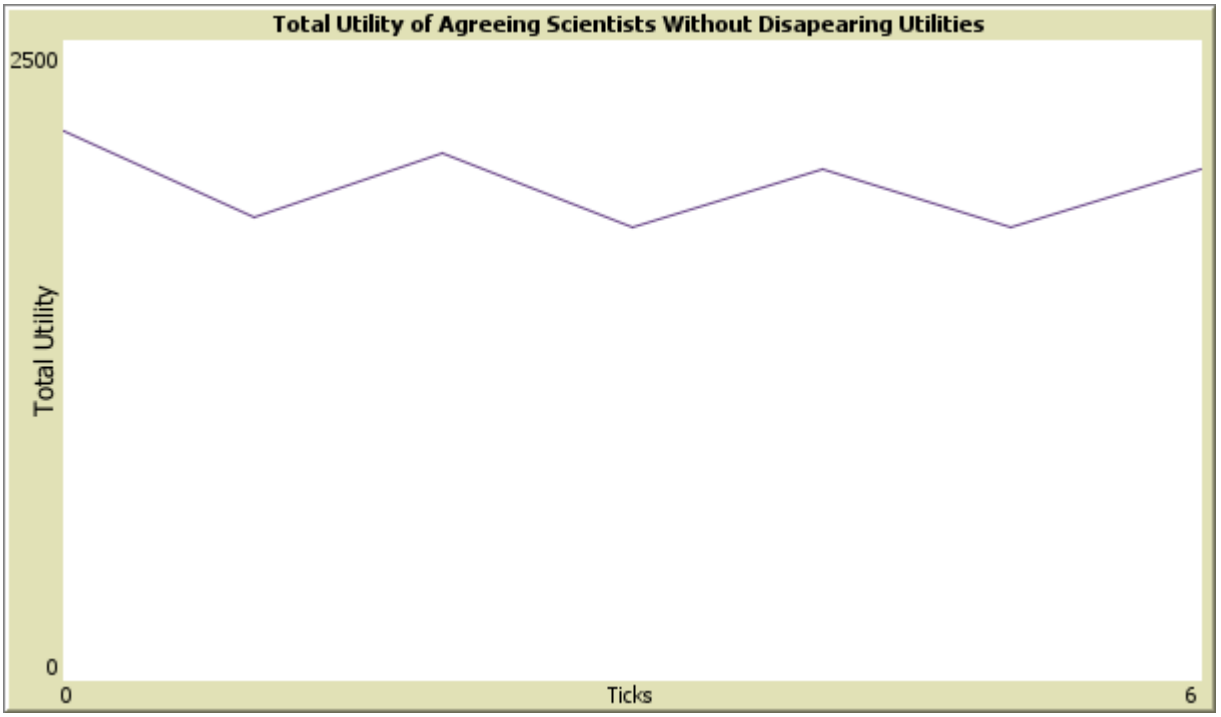


Figure 7.3. Total expected utility, over time, for agreeing scientists when patches do not lose utility after a scientist has visited it (Gillen, 2025a).

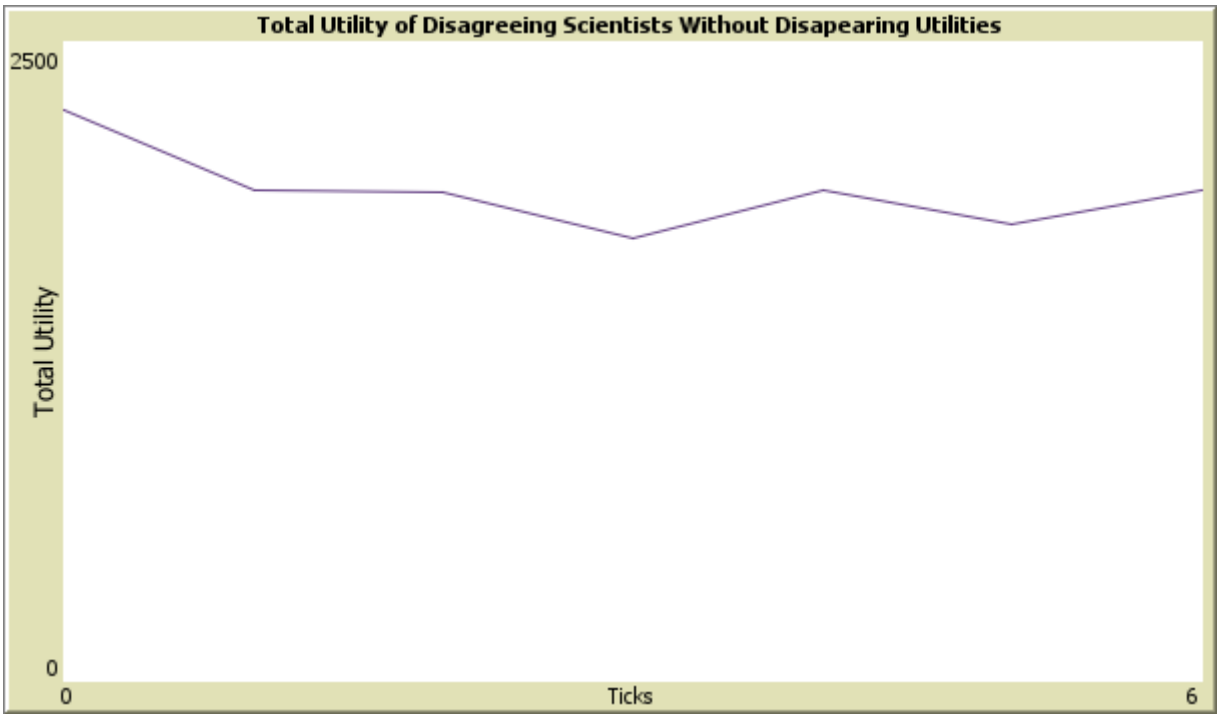


Figure 7.4. Total expected utility, over time, for disagreeing scientists when patches do not lose utility after a scientist has visited it (Gillen, 2025a).

The results of this novel model cannot be used in good faith to argue in favour of diversity in priors amongst astrobiologists. The results of this model merely suggest that scientists who tend to gather on the same projects – where these projects will only give one scientist its utility – tend to produce less overall utility than scientists who tend to spread out. This model hopefully stands as a motivating example of the need for caution over leveraging community models to inform about the real world, and the remainder of this chapter will support this claim. Presently, three popular community models will be analysed to this end.

7.6. Avin's Randomised Funding

The first model I will consider is found in Shahar Avin's "Centralized Funding and Epistemic Exploration" (2019). This model aims to investigate how different funding strategies affect the generation of scientific truths in a landscape of competing research projects and researchers. The central conclusion of Avin's model is that "On a large landscape, when a topic can be explored in many ways that could be very different from each other, random selection performs much better than selection based on past performance" (Avin, 2019, p.653).

7.6.1. Brief Outline of Avin's Model

I will here briefly summarise the structure of Avin's model for the purpose of later analysis. Avin's landscape consists of a two-dimensional space representing different research projects. The physical proximity of two tiles in the landscape is representative of their similarity in terms of research question, method, and any other relevant feature. A third dimension is added to the landscape, representing each research project's significance. As a result, Avin's landscape is a hilly one, with agents moving in the x-y axes to traverse across different projects and moving up and down along the z-axis to signify the significance of each research project.

Avin then introduces agents (representing scientists) to his landscape. These are initially distributed randomly across the landscape. The agents are programmed to sit on their tile until their tile countdown has ended (this countdown represents the duration of the research project). Once this has ended, the significance (characterised by the height of the tile) will be added to the overall accumulated significance of the model. This is representative of the scientific knowledge gained from completing that research project. Following this, the agent will look at their eight surrounding neighbourhood tiles and move to

the one with the highest significance. In this way, Avin's agents can be characterised as local maximisers, or hill climbers.

However, not all scientists who initially appear on the landscape are able to start their projects. And herein lies the crucial addition to Avin's model: the addition of different funding strategies. With each tick of the simulation, all agents who have run down their countdown will be put in a pool alongside new agents who have been randomly distributed across the landscape. Only a subset of these combined candidates will be allowed to actually populate the landscape and carry out their research project. How these agents are selected is dependent on which funding strategy is employed.

The purpose of Avin's model is to explore which funding strategy produces the highest total collected significance over time. As such, he models four different funding strategies. 1) Best: this is a god's-eye perspective where the agents on the highest tiles get funded. 2) Best_visible: agents on tiles within two tiles in any direction from another agent (past and present) are considered. Agents on the highest of these tiles are selected. 3) Lotto: agents are selected at random. 4) Oldboys: no selection is made as the existing agents just move to their highest neighbouring tile.

Of final importance is the introduction of merit dynamics. Avin includes three pivotal features in his model, which greatly influence his model's conclusions. These are, firstly, the winner-takes-it-all feature, which sets the significance of a project to zero once an agent has completed the project. Secondly, the reduced novelty feature means that the significances of the surrounding tiles of a completed project are reduced. Finally, the new avenues feature creates a new hill at a random location on the landscape after an agent has completed a project with high significance.

With the construction of his model as outlined here, Avin finds that, for sufficiently large landscapes, when only the winner-takes-it-all mechanism is employed of the three merit dynamics, the "best" funding strategy produces the most accumulated significance, followed by the "lotto", and tail-ended by "best_visible" and "oldboys", which perform similarly poorly. The results are shown in Figure 7.5. When employing all three merit dynamics, Avin found slightly different results. Most notably, "Lotto" now overtakes the "best" strategy (again, for larger landscape sizes). The results are captured in Figure 7.6.

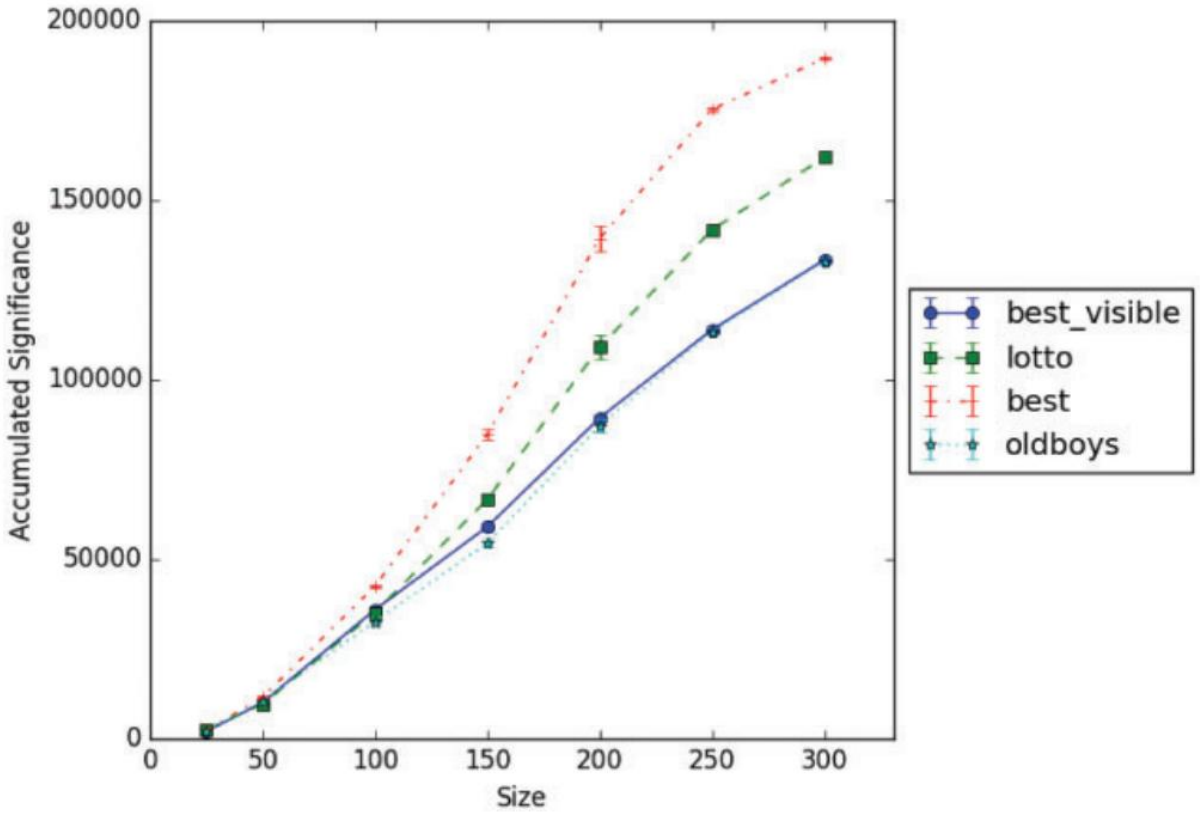


Figure 7.5. Comparison of total accumulated significance for the four funding strategies for a model employing only the winner-takes-it-all merit dynamic (Avin, 2019, p.644).

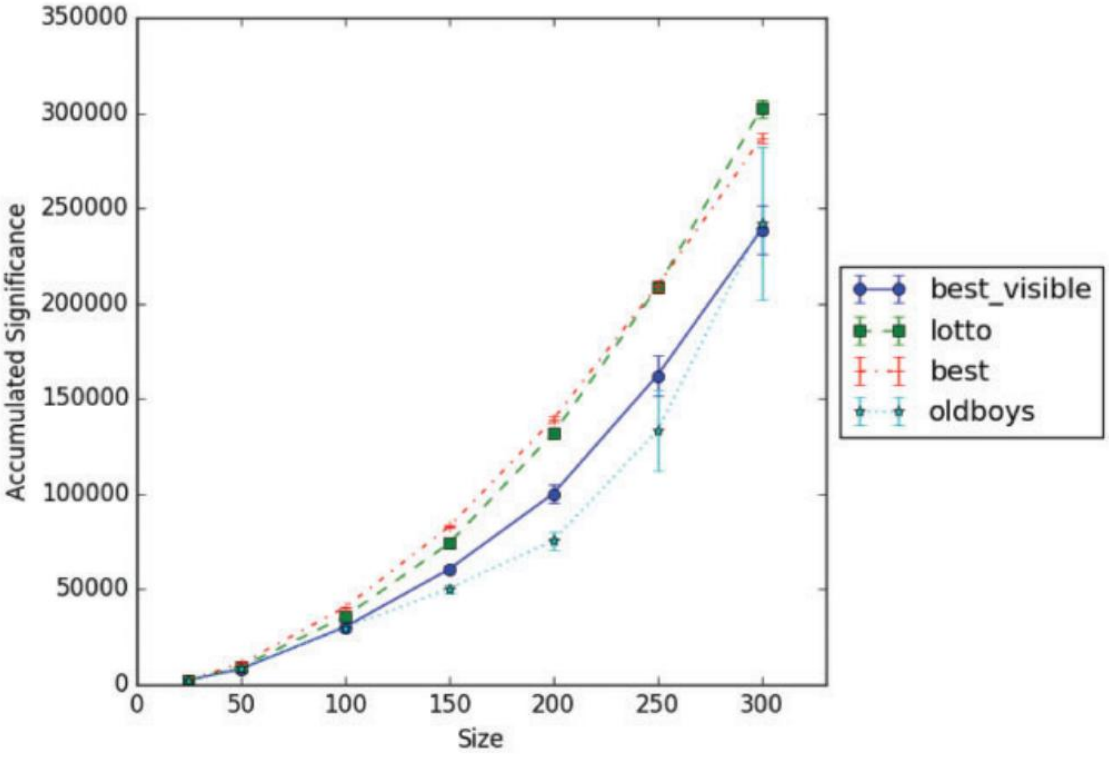


Figure 7.6. Comparison of total accumulated significance for the four funding strategies for a model employing all three merit dynamics: winner-takes-it-all, reduced novelty, and new avenues (Avin, 2019, p.643).

The upshot of these findings, Avin concludes, is that more randomness should be introduced into the funding of scientific and other epistemic projects. The traditional peer-review process is compared to the "best_visible" funding strategy (Avin, 2009, p.641), which was outperformed by the "lotto" strategy whenever the merit dynamics were introduced (strictly speaking, when only the winner-takes-it-all feature is introduced, and when all three are introduced). This conclusion, though, will presently be shown to be 1) circular and 2) overinterpreted.

7.6.2. The Short Vision of Avin's Agents

To make the case that Avin's conclusions are circular and overinterpreted, let us first dive into a pivotal feature of Avin's model: the vision of agents. Agent vision is used for the "best_visible" funding strategy, branded as representing the status quo peer review process. It is argued that "best_visible" is akin to peer review as the reviewers cannot accurately assess the significance of a research proposal that is too dissimilar to previous work (Avin, 2019). Agents in Avin's model can see up to two tiles away in any

direction, and only an agent on these tiles can be funded under "best_visible". This set of visible tiles increases over time as the "best_visible" mechanism remembers the tiles visible by previous agents. The list of visible tiles is, therefore, cumulative.

Avin reasons this vision length by noting that tiles further away from established research represent projects that are too novel to be seriously considered by a funder, and he evidences this with research on the matter. Such research finds that research proposals score the highest by funders when roughly half their key terms overlap with current literature, and half are novel (Boudreau et al., 2016). For this reason, Avin limits the distance agents can see to two tiles in any direction, thereby representing the restriction in the novelty of projects funded by peer review.

This feature is essential to deriving Avin's conclusion that "lotto" outperforms "best_visible". Note firstly that "best_visible" tends to "best" as the vision length increases, up until the point where the vision length is sufficiently large for a given landscape, such that all tiles are visible. The power, therefore, of "best_visible" being able to select the globally most significant projects is vastly determined not by the absolute visibility of the agents, but by their relative visibility compared to the size of the landscape.

Avin (2019, pp. 646-647) does explore the effect of varying the vision length, and his results are presented in Figure 7.7. Indeed, it only takes a relatively small increase in vision length for the "lotto" strategy to lose its edge over the "best_visible"; this happens at vision lengths of just over three tiles. Avin recognises that these results do challenge his conclusions that "lotto" outperforms "best_visible" but stresses that Figure 7.7 shows "triage" (which is a 50/50 divide between the "lotto" and "best_visible" strategies) still outperforms "best_visible".

However, the extent of Avin's exploration into vision length falls short of that necessary to fully understand the restricted range for which "triage" wins out over "best_visible". Essentially, the range of agent vision considered is too small for the landscape size Avin is using. Avin uses a landscape of 150 x 150 tiles, so the relative vision coming from five tiles in each direction (the maximum vision length considered) is still small. Moreover, from Figure 7.7, it is apparent that "best_visible" is rapidly catching up with "triage". We might wonder if "best_visible" would overtake "triage" if an agent vision of six tiles were tested, though this data is not provided.

Finally, and importantly, the conclusion of Avin's paper is not that "triage" outperforms "best_visible" but rather Avin finds "explicit random allocation performing significantly better than peer review on large landscapes" (Avin, 2009, p.629), where peer review is argued to be akin to "best_visible (Avin, 2009, p.641). Hence, the effect of increasing agent vision length beyond three tiles is of detriment to Avin's conclusions.

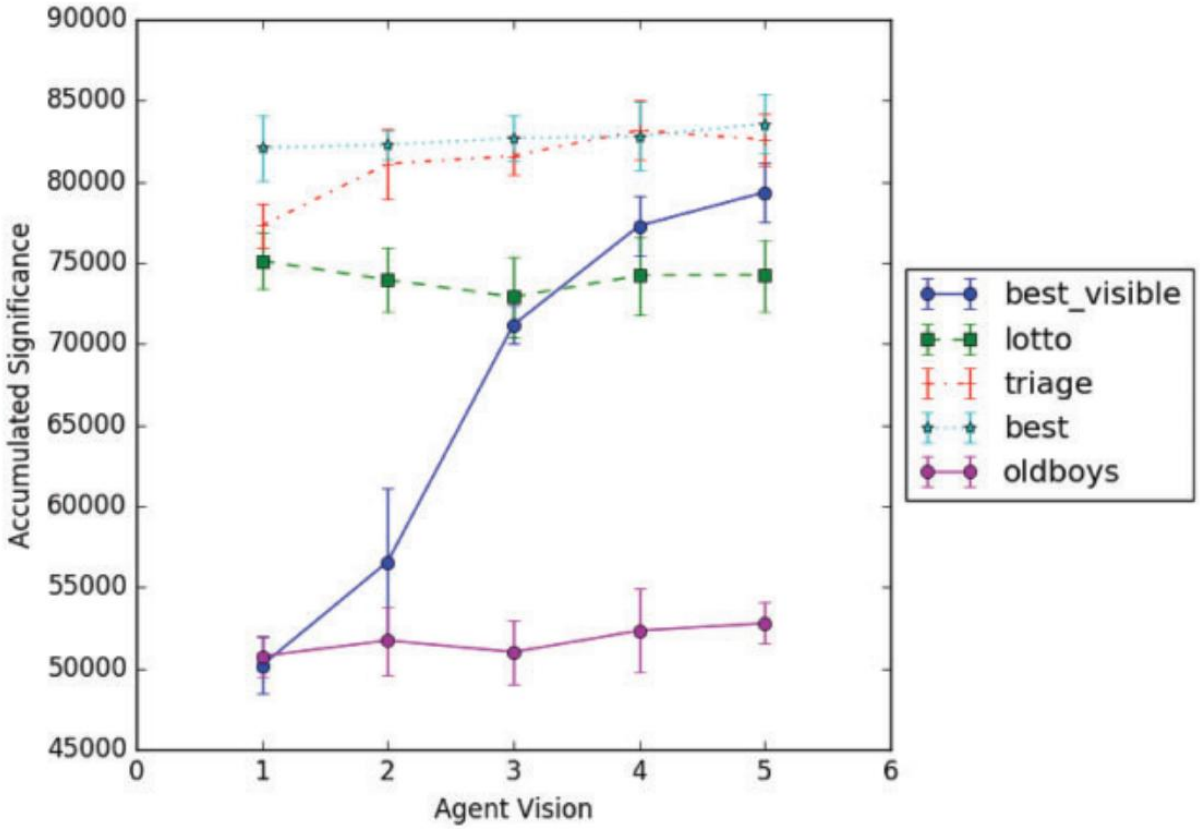


Figure 7.7. Variation of accumulated significance with agent vision for different funding strategies (Avin, 2019, p.644). Note that "triage" is a 50/50 combination of "best_visible" and "lotto".

Figures 7.5 and 7.6 also show the central role that the *relative* size of agent vision plays in the performance of "best_visible." Although difficult to see due to the scale of the y-axis, for small landscape sizes, the difference between the accumulated significance of "best_visible" and "lotto" is tiny; the difference appears to increase as landscape size increases. This is as expected, given that the power of "best_visible" is wholly determined by what proportion of the landscape agents can see.

All that can be concluded then is that, in this model, "lotto" outperforms "best_visible" *in the extreme* of small vision ranges. The real-world justifications, therefore, for enforcing these small vision ranges must be robust if the model's conclusions are to hold water.

Unfortunately, though, the real-world analogue for this vision is a stretch. It may be a genuine phenomenon that funders prefer a mixture of familiar and novel key terms in project proposals (as evidenced by Boudreau et al., 2016), and researchers may wish to exploit this in their project proposals.

However, is this effect meaningfully represented in Avin's model? By appealing to Boudreau et al.'s work on novelty and peer-review scores, Avin requires that his agent vision of two tiles in each direction in a landscape of 150 x 150, is representative of proposals having roughly half their key terms overlapping with existing literature and half their key terms being novel. Such an interpretation of two-tile-agent-vision strikes as heavily overinterpreted. Why not four tiles? Or instead adopt a vision length as a percentage of the landscape, say 10% or 50%? In this case, such vision length would be far further than the two tiles allocated by the present model. The point is that the vision length is fundamental to the performance of "best_visible" and 1) the mapping of the real-world phenomenon captured in Boudreau et al. (2019) onto agent vision length strikes as weak, and 2) the decision to set this vision length at two tiles appears convenient and poorly motivated.

The interpretational gap between agent vision in Avin's model and the real-world phenomenon it attempts to simulate is of concern. However, what makes the agent vision even more suspect is what happens to Avin's model when the agent vision feature interacts with the three merit dynamics. It is the combination and interaction of these *four* features that, I will argue, analytically derive a result which has little real-world meaning.

7.6.3. The Three Merit Dynamics

As outlined in §7.6.1, Avin introduces three merit dynamics into his model: winner-takes-it-all, reduced novelty, and new avenues. These features, especially when combined with the short agent vision, are what ultimately guarantee the conclusions of the model. It is paramount, then, that these three merit dynamics are honest representations of the real world. However, this subsection will demonstrate that: 1) these features introduce an element of circularity in that they specifically preference the conclusions drawn, and 2) these features stray too far from their real-world analogues to have significant real-world applicability.

I will now take each of the merit dynamics in turn. The winner-takes-it-all feature reduces the significance of a tile to zero after an agent has been on it for the project's duration. This idea has been modelled previously by Pöyhönen (2016), though this earlier model only proportionally reduced the significance rather than setting it to zero. The rationale for this is that, once a scientific discovery has been made, the epistemic value of rediscovering it is zero. Before delving into whether the feature truly captures its real-world analogue, let us consider the impact of the winner-takes-it-all dynamic on the model.

To do this, we must first broadly consider the value of an agent occupying a tile in Avin's landscape. Avin's landscape is a physical space (which is already an interpretation of the metaphorical closeness of

research projects regarding methods, equipment, research questions, etc.). The success of funding strategies in Avin's model comes from the ability of the funding strategy to select visible agents with the highest tile significance, and for this to remain over time. The value, therefore, of an agent being on a tile is not just limited to the tile's significance, but the significance of surrounding tiles, especially the immediate neighbouring tiles, as agents can only move one tile at a time. It is this positioning in the landscape that dictates how the accumulated significance will grow over time.

By reducing the tile significance to zero via the winner-takes-it-all feature, Avin introduces a physical hole in the landscape. Suppose the agent completing a research project moves to a neighbouring tile that is significant enough to be funded in the next round. In that case, they will now be at a positional disadvantage. This is because one of their neighbours is a hole, so their mean neighbour significance is likely less than that of other agents.

The result of this is that the winner-takes-it-all strategy reduces the average significance of the neighbouring patches of a pre-funded agent. It, therefore, introduces a bias into the model for funding agents physically far away from previously funded agents. Such funding is difficult for the "best_visible" strategy, due to the visibility being tied to where agents are or have been. However, this funding aligns with the "lotto" strategy whereby agents scattered randomly each round may be funded.

It is apparent then that the winner-takes-it-all feature drives the need for diversity in agent positioning. Having shown the mechanism by which this feature introduces a bias for diversely located agents, it remains to ask whether Avin's feature accurately captures its real-world analogue.

As mentioned previously, Avin appeals to the real-world reduction in a project's epistemic significance once it has already been carried out, and this idea does hold up. It is usually the results from the first research team to make a significant discovery that get published in scientific journals. A famous example of this from the history of science comes from Darwin's hastened publication of his *On the Origin of Species* (Darwin, 1859). Having spent over two decades after returning from HMS Beagle developing and hesitating over his theory of evolution, Darwin caught wind of his colleague's rather similar work. Upon hearing of Alfred Russel Wallace's similar ideas, Darwin quickly published his work.

Having said this, this preference for the first researcher past the post does not mean that repeats of the same experiment have zero value. The replicability of results is essential to the scientific process for verification, and it is not unusual for these verification papers to get published. A recent example of this was the detection of water vapour on the super-Earth exoplanet, K2-18 b. These results were published in *Nature Astronomy* in September 2019 (Tsiaras et al., 2019) and were quickly followed by confirmatory

papers (e.g. Benneke et al., 2019; Tsiaras, 2021). Avin's decision to reduce an explored tile to zero might, therefore, be extreme.

Moreover, the topographical manner in which Avin represents this reduced significance for previously explored projects is troublesome. The holes introduced into the landscape via the winner-takes-it-all feature mean that agents have fewer research options for projects adjacent to previous work, which seems strange. Considering an example might be helpful here. Let us imagine two colleagues working within the same laboratory on closely related projects. This would be represented by agents sitting on adjacent tiles in Avin's model. One of the researchers, Alice, makes a breakthrough and publishes her results widely. Avin's model now has the second researcher, Ben, at a positional disadvantage in the landscape regarding his next move as one of his eight neighbouring patches has become a hole.

It may indeed be true that Ben does not wish to copy Alice's work, and in this way, her tile is no longer available to him. But, in reality, more ideas should open up to Ben on account of Alice's breakthrough. Researchers talk, and knowledge is shared. Even if Ben were not working in the same laboratory as Alice, if his project is truly adjacent to hers, he would likely hear of her results through colleagues or at conferences or read about them in journals. The idea, then, of introducing a feature that disadvantages agents adjacent to completed projects is not representative of the real world.

This real-world dynamic could be better accounted for in Avin's model by having agents see past a hole in the landscape, thereby increasing their vision length in such cases. This would better account for how, once a discovery is made in the immediate vicinity of a researcher's current project, this knowledge is shared amongst those immediate researchers, who can then conceptualise and competitively put forth a research project that leapfrogs the completed one.

Such an addition to Avin's model would boost the performance of the "best_visible" funding strategy by increasing agents' relative vision. However, this is exactly the point. The winner-takes-it-all feature actively disadvantages the "best_visible" strategy, whilst favouring the "lotto" strategy in a way that has little real-world analogue. The feature favours diversity in positioning in the landscape, so it is no surprise that the model concludes that diversity is best. Such a conclusion is the result of a feature fed into the model with no real-world representation. Hence, the winner-takes-it-all feature looks problematic.

The second merit dynamic that Avin introduces is the reduced novelty feature. This feature reduces the significance of the tiles surrounding a completed tile with significance above a certain threshold. Again, this feature is central to deriving the model's conclusion that "lotto" outperforms "best_visible" as it favours the selection of new, randomly located agents (something that only happens with "lotto"). The reason for this is that having been funded for a project, that agent will subsequently find themselves

surrounded by unattractive neighbours and will hence be less likely to be refunded in the next round. By contrast, the randomly located agents that make up part of the "lotto" strategy are not disadvantaged in this way.

The result of the reduced novelty feature is that it, again, favours diversity in agent location and disadvantages agents adjacent to previously funded projects. It naturally and unsurprisingly follows that the model churns out "lotto" over "best_visible" as the preferred funding strategy. Again, there is a circularity with this feature as it feeds into the model a preference for diversity, only to conclude that diversity is preferable. Such outright circularity is not *necessarily* a problem (recall that models are analytic in nature, and so all model outputs might be regarded as circular from the inputs). But with such a crucial, impactful feature deriving the call for diversity, we should certainly want this feature to latch on to a real-world mechanism. So, does it?

Avin does not elaborate on what real-world mechanism the reduced novelty feature aims to represent beyond stating that: "when a researcher makes a significant discovery, simulated by finishing a project with associated significance above a certain threshold, the novelty of similar projects is reduced" (Avin, 2019, p.642). A generous view of this may see some rationale here. If adjacent projects in the landscape are incredibly similar to each other, there may be an element of repetition in completing adjacent projects.

However, the effect should be more than compensated for by the fact that knowledge is often cumulative, and this is something that the model does not capture. In the real world, making one discovery often elevates the significance of tangential projects. Discovering nitrogen on Enceladus boosts the significance of then discovering yet another fundamental element for life, e.g., phosphorous. According to Avin's model, the discovery of nitrogen would reduce the significance of the detection of phosphorous, and this is wrong. The reduced novelty feature again introduces a bias for diversity (and thus the conclusion favouring diversity is baked into the model here), and yet, the real-world analogue for this reduced novelty feature is lacking

The final merit dynamic introduced is the new avenues feature. What this does is, when a project with sufficient significance is completed, a new hill is created at a random location in the landscape. Such a feature, as before, introduces a bias for diversity in agent location across the landscape. This is particularly true when working in tandem with the reduced novelty feature, as pre-existing agents will see the significance of their local neighbourhood decrease, whilst agents distributed randomly may find themselves on these new hills. The significance of the local neighbourhood of a completed project will, therefore, decrease (due to the winner-takes-it-all and reduced novelty features), while the average significance of the whole landscape will not (due to the creation of a new hill). Such dynamics preference

diversity of agent position and correspondingly will preference the "lotto" strategy over the "best_visible" strategy.

With this in mind, the question remains of whether the new avenues dynamic represents anything tangible in the real world. There is indeed truth in the idea that new research projects are opened up when a significant discovery is made. For example, the characterisation of Mars' atmosphere opened up a plethora of research projects that tested extremophiles in similar conditions. It would not have been possible to model the Martian environment until we had data on it.

However, as this example shows, we should expect the new avenues to be connected to the initial discovery. More often than not, the new projects will be tangential to or build upon the previous discovery, as opposed to being completely unrelated to it. As such, it would be expected for the new hill in Avin's landscape to appear close to the initial project rather than at a random location in the landscape. On the contrary, the new hill being randomly located is analogous to the characterisation of Mars' atmosphere creating new research projects on, for example, refining methods of exoplanet detection. The fundamental point is that Avin's model assumes there is no connection between a significant discovery and the research projects that the discovery opens up. Such an assumption goes against intuition and is, again, ill-motivated.

To take stock of how Avin's conclusions are drawn: the short agent vision, winner-takes-it-all feature, reduced novelty feature, and new avenues feature all instrumentally favour diversity in agent location across the landscape. Existing agents are directly negatively impacted by the winner-takes-it-all feature and reduced novelty feature as the average significance of their neighbourhood is reduced. The new avenues feature keeps the mean significance in the landscape high, while the significance surrounding an existing agent is reduced, hence favouring new agents. Each of these introduces a bias for a funding strategy that selects new agents; only the "lotto" strategy is programmed to do this. Finally, the short agent vision reduces the power of the "best_visible" and strengthens the effect of the three merit dynamics.

Given how instrumental these four features are to deriving the result of Avin's model, I have considered whether they each have a real-world analogue. In each case, I have argued that such analogues are either heavily overinterpreted, or simply not captured by the model. Hence, Avin's extension of his conclusions to the real world is similarly overinterpreted; the model's conclusions are an analytic artefact of a model with little real-world correspondence.

7.7. Hong and Page's Diversity vs. Ability

The second model I would like to consider is the much-cited Hong and Page result. In a paper titled "Groups of diverse problem solvers can outperform groups of high-ability problem solvers" (2004), Hong and Page argue that diversity trumps ability. This conclusion, though, will again be found to lack any meaningful real-world applicability; the conclusions of the Hong and Page result are attributed to an odd quirk of the model, whereby the more appropriate conclusion would be: hostility + ability trumps an army of hyper-ability clones.

7.7.1. Brief Outline of Hong and Page's Model

Hong and Page's paper (2004) aims to evaluate the effects of functional diversity on problem-solving within a group of cooperating agents. On the matter of functional diversity, this is given to mean the diversity in how agents process problems and attempt to resolve them. Hong and Page conclude from their findings that a group selected at random from a set of diverse agents outperforms a group populated by the best-performing agents. The result is obtained from two studies: one being a mathematical theorem and the other being a computational simulation. The two approaches are intended to be connected in that the computational model produces the result that "diversity trumps ability" (Hong & Page, 2004, p.1), and the mathematical theorem explores the logic behind the result and provides the conditions necessary to derive it.

The idea of the computational model is given as follows. Agents are tasked with locating the maximum point of a function, V, and each agent is ascribed a heuristic set that determines how, and how well, they do this. This heuristic set, φ, comprises k numbers, between 1 (inclusive) and l (exclusive). For example, if $k = 3$ and $l = 12$, an agent might have the heuristic set: [1, 4, 11], with each number representing the comparison point jumps made through the solution space. With each comparison point, the agent compares their status quo position value with that of the comparison point. If the latter is larger, the agent jumps to that position; if not, the agent stays stationary. The agent then continues looking, this time in accordance with the second value in the heuristic set, and so on. The agent will continue to check for larger numbers (prescribed by the moves allowed by their heuristic values) until they are at a location where none of their heuristic value jumps locate a larger number. In Hong and Page's simulation, agents are grouped and take turns sequentially running through their search systems until they get stuck. At this point, another agent picks up the task, and so on. Figure 7.8 exemplifies how this function maximisation works.

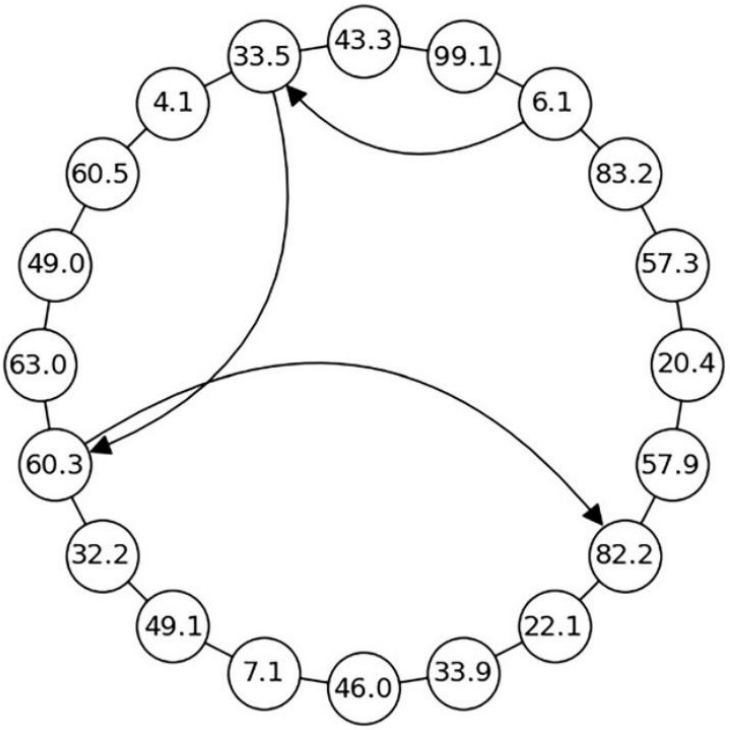


Figure 7.8. (Niesen et al., 2024, p.790) The function maximisation of Hong and Page's model (2004). The figure shows how an agent with the heuristic set [3,5,7] would move through the circle. Once they cannot move to a higher value, the next agent in their group will take over.

An agent's ability score is given by the expected value of where they get stuck (the highest value they find), considering all starting points. As given in Hong and Page (2004), this ability value is calculated as so:

$$E[V;\varphi] = \frac{1}{n}\sum_{i=1}^{n} V[\varphi(i)]$$

Here, n, represents the number of possible solutions.

Hong and Page report results for $n = 2000$, $k = 3$, $l = 12$ and choose groups of, initially, 10 agents. The results run as follows: the performance score for the group with the 10 highest ability agents was 93.2, and their diversity score was 0.72. The performance score for the randomly chosen agents was 94.7, with a diversity score of 0.92. Hong and Page thus summarise that, in this case, diversity trumps ability.

The mathematical theorem sets out four assumptions and concludes that: there exists positive integers N and N_l where $N > N_l$ such that, with probability one, the joint performance of N_l independently drawn agents will exceed the joint performance of N_l best-performing agents drawn from a pool of N agents.

If Hong and Page's results and inferences from the computational model and mathematical theorem are right, the real-world applications would be significant. Indeed, Hong and Page discuss such applications, and these include how companies choose their employees — their recommendation is to select for diversity, not just the most able applicants (Hong & Page, 2004). Moreover, Hong and Page's findings have been directly appealed to in defence of the benefits of democracy (Landemore, 2012, see especially, Chapter 4.2 and Chapter 6). If, however, the findings of the Hong and Page result can be found unjustified or the application of the result is inappropriate, the significance will be undermined.

7.7.2. Hostility Trumps Ability

Several critical flaws with the Hong and Page result can be exposed, and indeed Abigail Thomson does just this in her destructive critique "Does Diversity Trump Ability? An Example of the Misuse of Mathematics in the Social Sciences" (2014). Parts of Thomson's analysis have received some criticism (e.g., Grim et al., 2019; Kuehn, 2017; Page, 2015), and the current chapter will not address all seven arguments. However, here, I will review and develop the attacks most relevant to this chapter.

Thomson rightly unpacks the conclusion that diversity trumps ability by questioning what diversity means in this context. What has the study argued for? Hong and Page define diversity on a scale of 0 to 1. It is calculated by taking the number of places in the heuristic sets where two agents differ and dividing by the size of the heuristic set. For example, two agents with heuristics [1,2,3] and [3,1,2] would have the maximum diversity of 1. Giving this number the label of diversity and consequently extending the meaning to real-world contexts — as Hong and Page do — is a stretch.

This incredibly simplistic model fails to capture human diversity. Indeed, Thomson points out that one might better label this quantity as hostility, given that it represents disagreement between approaches (2014). It might thus be concluded that it is hostility that trumps ability. Of course, this is also not justified, but it shows that Hong and Page's argument hinges on this unwarranted equating of diversity with this value. The result fails to say much of substance about diversity, as conceived of in an everyday setting.

7.7.3. Diversity Trumps an Army of Clones

A second attack on the interpretation of the Hong and Page result and its application to the real world lies with the mathematical theorem. Specifically, this is in the strategic selection of the values k, N_l, and N. These values represent, respectively: the number of unique agents, the number of agents selected to

form the groups, and the total number of agents being selected from. Ultimately, the theory depends entirely on N and N_I being sufficiently larger than the pool of distinguishable agents.

This result might best be explained with an example inspired by Page's (2007) own recommendation for the theorem. Page (2007) advises that "when picking two hundred employees from a pool of thousands… we should keep the theorem in mind" (2007, p.164). In such a case, $N = 2000$ and $N_I = 200$. The shocking thing is that the proof of the theorem requires that k (the number of unique applicants in this example) be small enough to ensure that the group of N_I best-performing agents are identical copies of the same agent. This is stated in the final line of Hong and Page's proof: "There are more than N_I numbers of agents among the group of N agents that are the highest performing agent φ^*. Thus, the best N_I agents among the N agents are all φ^*" (2004, p.5).

It is this curious feature of Hong and Page's mathematical proof and corresponding model that derives their conclusion that diversity trumps ability. They ensure that the number of unique agents, k, is sufficiently small compared to the number of selected agents, N_I, so as to ensure that their groups of high-ability agents are identical. The result of this is that, once an agent in a group gets stuck (i.e., they have run through their heuristics, and none of them takes the agent to a higher-value spot), the next agent in the group will be useless in providing any help. All the agents in the group will have the same heuristics, so we effectively have a group comprising a single agent.

To return to the example, in order to satisfy the condition that k be suitably smaller than N_I, the employee would need to select 200 applicants from a pool of 2000, of which there are only, say, 12 unique applicants. The pool must contain multiple clones of the same applicant! This is an unrealistic requirement for a theory that Hong and Page claim has real-world implications. The result suggests that diversity can be advantageous over an army of clones, but it says little beyond this.

Of further concern is the limit on N_I required to output Hong and Page's computational result. Their results give the group performance and diversity for group sizes of 10 and 20 agents. Even within these small values, Hong and Page note that "when we enlarged the group size from 10 to 20, the random group still did better, but with a less pronounced advantage" (2004, p.3). The extent of this decreased advantage is significant. With groups of 10 agents, the performance of the best agents was 92.56, whereas that of the randomly selected group was 94.53. A percentage difference of 2.1%. With groups of 20 agents, the best agents scored 93.78, and the random agents scored 94.72. This is a percentage difference of 1.0%, less than half that of the smaller group.

Many real-world applications of Hong and Page's result involve more than 20 individuals. Examples might include selecting students for university courses, selecting research projects to receive funding, and

even debates over who in a community should be franchised to vote. Such numbers would bring about a phenomenon that Hong and Page themselves use to explain the drop in advantage that the randomly selected groups had: as group size increases, the diversity in the high-ability group increases. This is to say, as the group size increases, the likelihood that the high-ability group comprises an army of clones decreases.

The question then becomes whether the diversity would increase to the same level as the random group and if this would correspondingly reduce the random group's advantage. Hong and Page admit that, ideally, a crowd would be both diverse and smart (2008); however, they are often not, and thus, diversity should be selected over ability. This may be a false dichotomy. With sufficiently large group sizes, both diversity and ability may be achieved. As it stands, all that Hong and Page's result shows is that what is termed *diversity* trumps ability when groups are small enough to ensure multiple copies of the same individual are present in the selected group. The real-world applicability of such a result is highly limited.

7.7.4. Diversity + Ability Trumps Hyper-Ability

A final issue with extending the Hong and Page result to real-world scenarios is that it relies on excluding what are deemed unintelligent agents. This imposes a large caveat on the conclusion that diversity trumps ability. In actuality, the most the result can be taken to suggest is that diverse and smart agents trump the smartest agents.

This competence assumption is stated in the conclusion of the paper: "The main result of this paper provides conditions under which, in the limit, a random group of *intelligent* problem solvers will outperform a group of the best problem solvers" (Hong and Page, 2004, p.16389, emphasis added). The intelligence requirement for agents is needed to the extent that, given any initial starting point in the solution space, an agent can locate a weakly better solution. The inclusion of this baseline intelligence is not insignificant. It, again, restricts the scope of Hong and Page's conclusion.

It becomes inappropriate to, for example, use this result to argue for the unfiltered random selection of candidates for funding, jobs, and anything else. By introducing a baseline ability into the model, a more appropriate interpretation of the result would be an endorsement of a gated lottery system. However, considering the other interpretational and computational issues discussed above, any real-world application of the Hong and Page result is seriously undermined. The results may best be described as an interesting quirk of a mathematical model with little real-world analogue.

7.8. Weisburg and Muldoon's Mavericks and Followers

The final model I will consider here is Weisburg and Muldoon's seminal mavericks and followers model. In their "Epistemic Landscapes and the Division of Cognitive Labor" (2009), Weisburg and Muldoon report the results of a NetLogo model. By considering populations of agents who are biased toward the work of previous agents ("followers") and populations of agents who actively avoid the work of previous agents ("mavericks"), they find that pure populations of mavericks outperform pure populations of followers. However, mixed populations perform best.

As with the previous models considered in this chapter, I will briefly outline the workings of the model in order to identify what conditions are principally deriving the result. Discussions will follow as to whether such conditions sufficiently represent any real-world phenomena and, hence, whether the model can truly teach us anything about the real world.

7.8.1. Brief Outline of Weisburg and Muldoon's Model

Weisburg and Muldoon's landscape consists of two peaks, the heights of which represent the significance of each patch. By creating their model in Netlogo, they utilise the benefits of an agent-based model. They note that, in comparison to other models (e.g., Kitcher, 1990, 1993; Strevens, 2003), they can account for the epistemic state of their modelled scientists: "We can more realistically represent the actual epistemic situation of scientists who have limited knowledge about the landscape. Scientists do not see the whole landscape at the beginning of the simulation; they learn about the landscape by exploring it or observing others" (Weisburg & Muldoon, 2009, p.231).

Weisburg and Muldoon then populate their landscape with scientists. They model two types of scientists, distinguished by how they are programmed to move about the landscape, and a group of controls. The first of these groups is named the followers. These agents are biased toward patches already visited by other agents. These followers move about the landscape by asking if any of their neighbouring patches have previously been visited. They then pick the visited patch with the highest significance and move to this. If none of the neighbouring patches have been previously visited, the agent will move to an unvisited patch at random.

The second group of scientists in the model is called the mavericks. These are so-called because they are biased *against* patches that have been previously visited. Conversely to the followers, mavericks ask if any of their neighbouring patches have been *unvisited*. If so, they move to this patch (if there are multiple unvisited patches, the agent will choose randomly between them). If there are no unvisited patches, the

maverick agent will choose randomly between the visited patches which have higher significance than the agent's previous patch. If none of the visited patches have higher significance, the agent will go back one patch and repeat the process until either an unvisited patch is found, or a visited patch with higher significance than the agent's previous approach is found.

Unlike followers and mavericks, agents in the control group do not account for whether a patch has been previously visited. Instead, they move about the landscape as follows. Firstly, agents are given a random heading, essentially the direction they face in the landscape. For each tick, the agent will move forward one patch and ask whether it is more significant than their previous patch. If so, they will remain on this patch. If it has equal significance, with a 2% probability, they will move to another patch with a random heading. Otherwise, they will stay put. If the patch has less significance, the control agent will return to their previous patch, set a new heading at random and try again until they find a new patch.

Unlike Avin's later model (2019), Weisburg and Muldoon's patches do not lose their significance once an agent has visited them. Similarly, though, Weisburg and Muldoon's agents are hill climbers in that they are local maximisers. The speed at which the controls, followers, and mavericks reach the two peaks in the landscape is an important output of the model. A second important point of comparison between the groups is what Weisburg and Muldoon call *epistemic progress*. This is defined as "the percentage of patches with significance greater than zero that have been visited by the community of scientists" (Weisburg & Muldoon, 2009, p.237).

The key takeaways of Weisburg and Muldoon's model are as follows: 1) for pure populations, mavericks find the peaks fastest, followed by controls, then followers; 2) for pure populations, mavericks produce the highest epistemic progress, followed by controls, then followers; 3) adding mavericks into a population of followers increases the epistemic production of pure populations of followers, hence "making polymorphic populations of mavericks and followers ideal in many research domains" (ibid., p.225).

The authors then further extend their model's conclusions to draw analogies with Kuhn's normal vs. revolutionary science. The followers represent the puzzle-solving of normal science, whereas the mavericks engage in revolutionary, novel science that often results in the most significant discoveries. Finally, Weisburg and Muldoon comment on how mavericks may be more expensive to society (on account of them avoiding preexisting work, equipment, methods, etc.). As such, there exists an optimal ratio of followers to mavericks that would result in an optimal division of cognitive labour (ibid., p.251).

As with the models of Avin (2019) and Hong and Page (2004), the following will unpick the workings of Weisburg and Muldoon's model to ascertain what assumptions and mechanisms are responsible for the

model's conclusions. Whether or not these assumptions and mechanisms are true representations of the real world or merely artefacts of an unrepresentative model will be discussed.

7.8.2. Unrealistic Landscape Shape

The first important feature of the model to discuss here is the shape of Weisburg and Muldoon's landscape. The landscape consists of two Gaussian peaks atop an otherwise flat terrain. The authors note, "This epistemic landscape is not meant to model any particular target scientific domain; however, we believe that it has several features which are common to many kinds of domains we wish to study" (Weisburg & Muldoon, 2009, p.234). There are two key features provided. These are, firstly, that domains often have multiple approaches that will yield scientifically significant results. This reasoning motivates having two peaks rather than one. Secondly, approaches associated with scientific significance tend to cluster together. This motivates having *only* two peaks rather than numerous peaks scattered throughout the landscape.

The shape of the landscape is significant as a more homogenous landscape with more local maxima, as opposed to only two large peaks, would result in different conclusions. As it stands, Weisburg and Muldoon's model has an equilibrium state of all the agents converging on one of the two peaks. This is because all three groups of agents (controls, followers and mavericks) are local maximisers. Hence, if all the agents happen to arrive at the base of just one of the peaks, they will likely move up that peak and never make it onto the other one.

If, instead, the landscape consisted of many small peaks, agents would get stuck sooner but would be more distributed throughout; a higher proportion of the landscape would be explored. This would result in a higher epistemic progress score (defined as the percentage of non-zero significance patches explored) and a quicker speed of ascent. There should, therefore, be a landscape dependence on the results of the model, and the unrealistic two-peaked landscape of Weisburg and Muldoon's model cannot be taken without due consideration; the landscape is not a neutral feature of the model.

Alexander et al. (2015, especially §7) also pick up on the limitations that an epistemic landscape imposes on a model's conclusions. Specifically, the authors highlight how the ruggedness of the landscape would heavily determine the effectiveness of social learning. If the landscape is rugged (as opposed to smooth), this would limit the benefit of another agent moving to a nearby project — that nearby project just might be at the bottom of a cliff. Moreover, real-world scientists will possess biases about certain projects, which leaves them unable to objectively assess the potential success of a candidate project (Alexander et al., 2015, p.450).

Another limitation of epistemic landscapes is that funding is not distributed equally across all projects in science. Some areas of the real-world "landscape", which might contain high-yielding results, may occupy an area of science with very little funding. Additionally, the real-world "landscape" that the model landscape is attempting to simulate is rarely static or well-mapped. Real scientists may lack the knowledge of which parts of the landscape are richer, and they may lose track of this as the landscape changes from underneath them. The epistemic landscape models, which assume a static landscape occupied by unbiased and unconditionally funded scientists, are therefore unrealistic. Further discussion of the importance of landscape shape will follow in §7.11 of this chapter. For now, though, we turn to another problematic feature of Weisburg and Muldoon's model.

7.8.3. The Problematic Definition of Epistemic Progress

Perhaps the most instrumental features in deriving the result of Weisburg and Muldoon's model are the circularity in how the mavericks' movement is coded and how epistemic progress is defined. The authors conclude that a population of mavericks far outperforms a population of followers or controls. This result is shown in Figure 7.9. Crucially, though, performance is with regard to average epistemic progress.

As mentioned above, epistemic progress is defined as the percentage of patches with significance greater than zero that have been visited by agents (Weisburg & Muldoon, 2009, p.237). Interestingly, this measure of progress does not track the total significance harvested by the community but rather the number of non-zero patches that have been visited. An agent repeatedly visiting the same highest significance patch in the landscape will, therefore, result in the same epistemic progress as an agent repeatedly visiting the same near-zero significance patch. Perhaps even stranger, agents occupying ten patches at a peak will produce the same epistemic progress as ten agents near the bottom of the peak.

Measuring progress in this way explicitly favours mavericks over any other group. This is because epistemic progress has nothing to do with the amount of significance but everything to do with the proportion of non-zero patches in the landscape visited. What this means is that agents who strive to explore new patches will do better than agents who either strive to return to old patches (followers) or simply strive to move to high-significance patches (controls). The distinguishing feature of mavericks is that they prioritise unvisited patches. Hence, it is unsurprising and uninformative that mavericks outperform followers and controls when it comes to epistemic progress, as so defined. It is apparent that a preference for diversity in agent positioning has been baked into the model via this definition of progress, and hence, the conclusion that diversity is beneficial is somewhat circular.

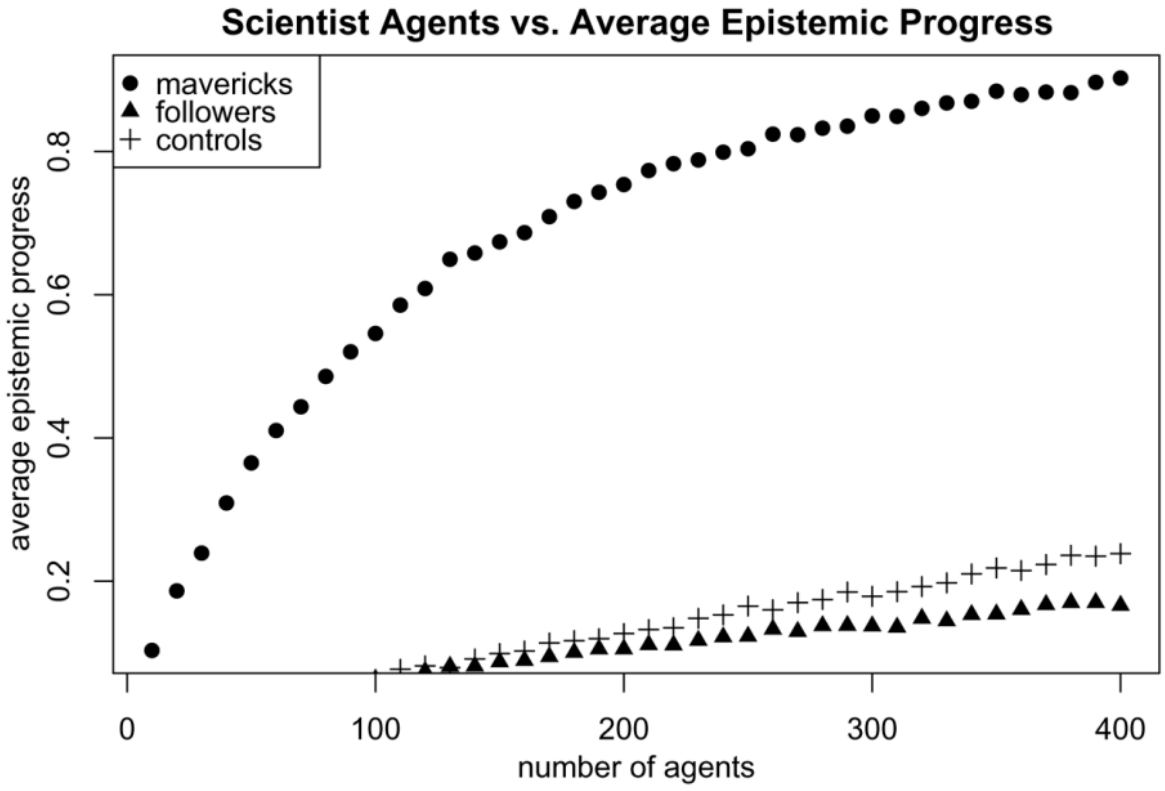


Figure 7.9. Average epistemic progress against number of agents in the landscape for pure populations of mavericks, followers, and controls (Weisburg & Muldoon, 2019, p.245).

Does, then, this way of defining epistemic progress latch on to anything useful in the real world? Knowing the degree to which all non-zero approaches are explored in a landscape is certainly not inconsequential. However, the more relevant measure should be the absolute, total significance garnered by the community. Therefore, the adoption of Weisburg and Muldoon's measure of epistemic progress that specifically favours mavericks leaves the conclusion that their mavericks are important to real scientific communities unsubstantiated.

7.8.4. Circularity Between Mavericks and Speed of Ascent

A further area of circularity, or baking-in of the conclusion, concerns Weisburg and Muldoon's conclusion about the speed of ascent. The authors conclude that "mavericks are far more efficient at finding the peaks than controls" (2009, p.244). By efficiency, the authors refer to both the speed and reliability of finding both peaks in the landscape. The followers performed worse on these measures, and

"populations of controls are pretty good at finding peaks; given enough time, they will always find at least one of the peaks", though not as quickly as the mavericks (ibid., p.245).

The real-world interpretation of these results is a commentary on how scientists who actively think outside of the box and engage in novel research will help the community find the most significant scientific discoveries. However, by again delving into the mechanism by which the mavericks find the peaks most efficiently, we will see that such a real-world interpretation is a stretch.

Given how the followers are programmed to move, it is not surprising that they are the slowest and least successful when finding the two peaks. This stands as an example of circularity within the model — the conclusions are arbitrarily baked into the model. As a reminder, each tick, followers will ask if any of their neighbouring patches have already been visited. They will then move to the visited patch with the highest significance. If no patches have been visited, they will move to a patch at random. There are two distinct reasons why this strategy reduced the followers' efficiency at finding and climbing peaks.

Firstly, the followers prioritise pre-visited patches over the highest-significance neighbouring patch. This slows them down at ascending a peak. By contrast, the controls will move to a higher significance patch at least 98% of the time; the followers can often move downhill under this movement function. Secondly, the followers can easily get stuck in preexisting loops. This is again a consequence of them prioritising already-visited patches. If a loop exists which consists of the highest-significance patches in a local neighbourhood, followers can fall into this loop and will quite literally follow each other in circles. This will reduce the reliability of followers in finding both peaks.

There may indeed be a real-world analogue of the first aspect of follower behaviour. It is reasonable to suppose that researchers often select familiar methods and projects over ones that will produce the most scientific significance, which would result in less scientific output overall. Though it might be extreme to suppose that followers always explicitly ignore a high-significance patch in preference for a low-significance, pre-visited patch.

More problematic is the second aspect of follower behaviour — the modelling of researchers perpetually moving in circles and rehashing colleagues' work. Although a degree of this undoubtedly happens, the reward structures in science discourage this. Specifically, this is due to the first past the post element of publishing, whereby the group of scientists first to discover something is most likely to have their work published. This is accounted for in Kitcher's (1990, 1993) and Strevens' (2003) marginal contribution/reward models, where rewards are given to researchers who are the first to discover something. Avin's (2019) model also accounts for this by reducing the significance of a patch to zero once a researcher has visited it. The feature of Weisburg and Muldoon's followers running in circles is,

therefore, a quirk of the model whereby the significance of a patch does not reduce as researchers repeatedly visit it.

The finding that mavericks are the quickest at hill climbing, and the extension of this to the real world, is similarly problematic. The way in which mavericks are programmed to move, in comparison to the controls, guarantees their success in an uninformative way. The implication is that the mavericks' out-of-the-box approach is what makes them quicker at ascending the peaks than the controls. However, their success is again better attributed to an unrepresentative artefact of the model.

With each tick, mavericks could potentially move to one of eight neighbouring patches, whereas the controls have just one option. The mavericks, therefore, have eight times as many options as the controls. However, this in and of itself does not give them an advantage. This is because the mavericks do not get to choose the highest significance patch but rather choose an unvisited patch at random.

What essentially guarantees the mavericks' success compared to the controls is that, each tick, controls are far less likely to move at all. What constitutes one tick (and thus the measure of time) is given in a footnote: "One cycle corresponds to each scientist agent following its rule set one time" (Weisburg & Muldoon, 2009, p.235). This means that agents might move through multiple cycles before they move, and the mean number of cycles per move for controls is inevitably higher than for mavericks.

For controls, if their singular forward-facing patch is not more significant than the patch they are on, they will not move. Even if the patch has equal significance, they will not move with 98% probability. By contrast, mavericks will move if either 1) there is an unvisited patch in their eight neighbours or 2) there is a patch with higher significance than their current patch. What this means is that mavericks have far more options each tick, and so are far more likely to move each tick compared to the controls. Hence, it is no surprise that they climb the peaks faster.

It is truly difficult to conceive of any real-world analogue for this instrumental mechanism. We would need something akin to a tendency for fringe scientists to carry out research quicker than other scientists. It is unclear whether such a tendency exists, and even if so, evidence must be provided to support this assumption. Furthermore, regardless of whether this assumption is substantiated, the conclusion that mavericks ascend peaks quicker than controls can largely be attributed to this assumption, not to any hidden and unforeseen nature of maverick research approaches.

7.8.5. Speculative Extrapolation to Mixed Populations

A final consideration of the Weisburg and Muldoon model is their conclusion that "In mixed populations, mavericks stimulate followers to greater levels of epistemic production, making polymorphic populations of mavericks and followers ideal in many research domains" (Weisberg & Muldoon, 2009, p.225). I will briefly discuss here how this conclusion does not follow from the results of the model but rather from a speculative comment born out of the model's limitations.

Granting the efficiency measures used by Weisburg and Muldoon (speed of ascent and epistemic progress), the model reliably shows that pure populations of mavericks outperform pure populations of followers. Weisburg and Muldoon then present the results for mixed populations of mavericks and followers. They investigate how increasing the number of mavericks present in a population of followers affects the total epistemic progress. Note here that *total* epistemic progress is the number of patches visited, regardless of whether they have significance or not (Weisburg & Muldoon, 2009, p.248). The total number of scientists is kept fixed at 400; they begin with a population of 400 followers and incrementally swap out followers for mavericks until they have a population of 400 mavericks. The results are given in Figure 7.10.

As Figure 7.10 shows, total epistemic progress increases as the proportion of mavericks in a population increases until the limit of a pure population of mavericks. A significant reason why mavericks boost the performance of a population of followers is that mavericks help followers get unstuck from their loops; a passing maverick may intercept a follower's loop and thereby give the follower a way out. The results of Figure 7.10 would suggest that a pure population of mavericks should be optimal over a mixed population.

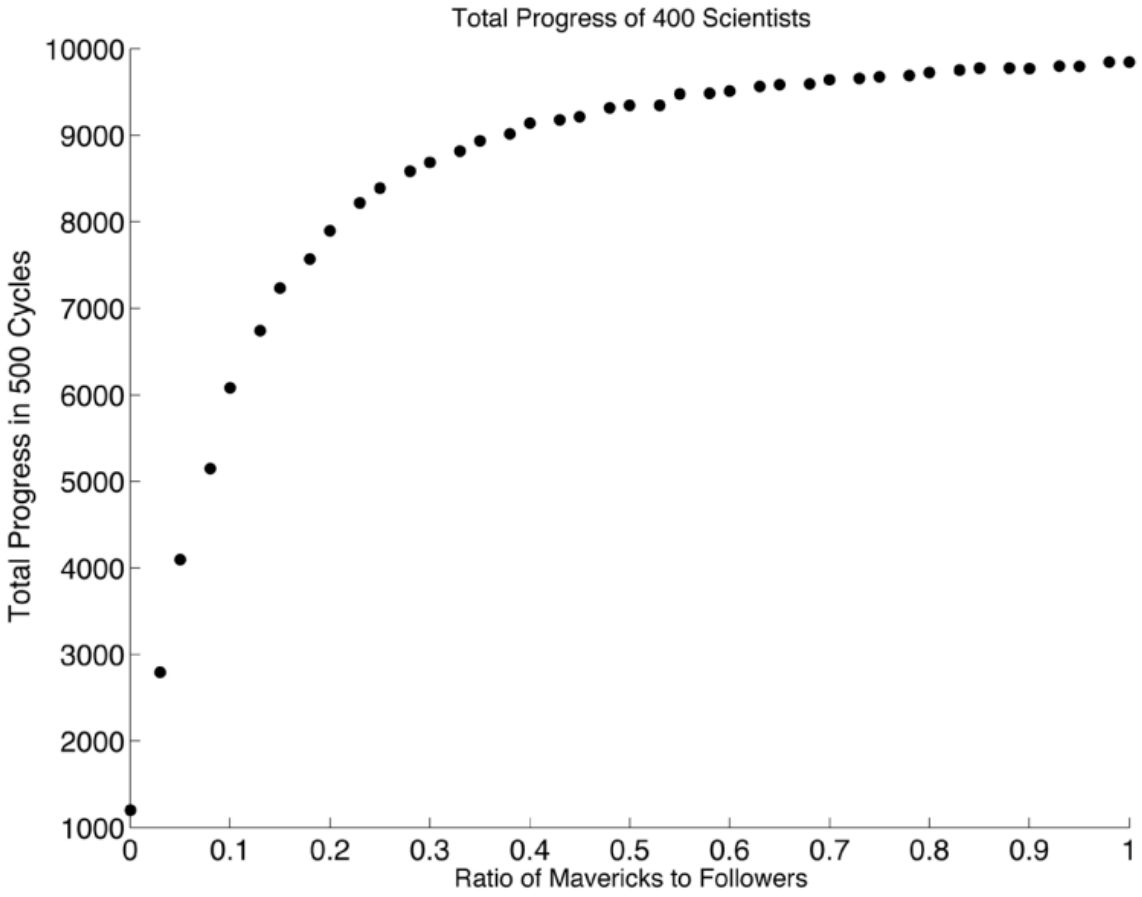


Figure 7.10. Effect of increasing the proportion of mavericks to followers in a population of 400 scientists. Note that "1" on the x-axis corresponds to 100% mavericks. (Weisburg & Muldoon, 2009, p.248).

The authors, however, arrive at this latter conclusion by suggesting that mavericks carry out more costly research than followers. This is a reasonable assumption that the authors justify: mavericks may require more exotic and costly equipment, it might take longer for unorthodox projects to be designed and set up, and background research may be lacking. With this in mind, the authors conclude that a mixture of followers and mavericks is optimal for a scientific community; however, "Without considerably more detail added to our models, it is hard to say exactly what the optimum balance should be" (ibid., p.251).

This is a fair caveat to the final conclusion. However, it should be noted that such a conclusion does not actually follow from the results of the model. Feasibly, due to the mavericks' high efficiency, a small additional cost per maverick may still result in pure populations of mavericks winning out. Follow-up research would indeed be needed to evaluate how much more costly mavericks would need to be compared to followers to result in a mixture being preferable. As things stand, though, the results of

Weisberg and Muldoon's model show that pure populations of mavericks outperform (as defined and discussed above) any other combination of scientists.

To summarise, then, this section has reviewed Weisburg and Muldoon's seminal model of the optimal distribution of followers and mavericks in a scientific community. I have unpicked the working assumptions within the model that are responsible for deriving the conclusions that 1) pure populations of mavericks outperform pure populations of controls and followers and 2) mixed populations of mavericks and followers are optimal for many research domains.

Considering the first of these conclusions, three key features within the model are found to be responsible for deriving this conclusion. These features are the landscape shape, the definition of epistemic progress, and how the movement of mavericks is coded such that it ensures they ascend peaks faster than followers and controls. All these features have been found to lack real-world analogues, and hence, the extension of the model conclusions to any real-world system is unjustified. Considering the second conclusion, this has been found to not follow from the model but rather from speculation and hence should be portrayed as such.

7.9. The Circularity Between Diversity and Success

This subsection will take stock of the three models discussed above and highlight the trend that diversity often derives success. More than this, though, the way in which success is defined necessitates diversity. This first observation, that diversity is a key feature in models of this kind, has been acknowledged elsewhere. Muldoon notes that "Many of the benefits of the division of cognitive labor flow from leveraging agent diversity" (2013, p.117). Similarly, Avin remarks that "Diversity in the community trumps individual pursuit of excellence as a way of making communal epistemic progress" (2019, p.653), which, of course, mirrors the language of Hong & Page's "diversity trumps ability" conclusion (2004, p.16385).

At the heart of each of the three models considered, agent diversity has proven essential to the chosen measures of success. Whether this diversity is in agent positioning to avoid new holes and find new peaks (Avin, 2019), sufficient diversity in problem-solving approaches to avoid groups of identical clones (Hong & Page, 2004), or diversity in terms of avoiding previous agent behaviour to prevent agents getting stuck in circles (Weisburg & Muldoon, 2009), it is diversity simpliciter that is essential.

Diversity, however construed, can largely account for the successes in many epistemic models. But, as the previous subsections have argued, this does not necessarily mean that the models considered make a

convincing case for diversity leading to success in any real-world system. Of course, real-world evidence favouring the epistemic benefits of diversity exists — e.g., in the contexts of science (Fehr, 2011), politics (Bohman, 2006), and business (Steel & Bolduc, 2020), to name a few. And the requirement for diversity alongside consensus amongst a scientific community for scientific claims to be considered "future proof", has been convincingly defended (Vickers, 2022). However, the question here is whether the models considered can add credence to these arguments.

The problem has been that, in many cases, diversity has been inherently baked into how the models define success. Weisburg and Muldoon's (2009) definition of epistemic progress is one notable example of this. As a reminder, epistemic progress is a count of the percentage of non-zero significance patches explored rather than the total amount of significance. As such, a strategy whereby agents deliberately occupy unexplored patches will outperform a strategy whereby agents strive to explore the highest-significant patches. This conclusion comes directly from the odd way in which epistemic progress is defined, and not from any real-world feature. Another such example of diversity being baked into a model's definition of success is the disappearing and reappearing peaks of Avin's (2019) model.

Finally, additional features of the three models discussed that are instrumental to the conclusions have been found to have no real-world analogues. All these features considered work to undermine any extrapolation of the models' results to any real-world system. The models have been found to suffer from circularity and a lack of real-world representation. The remainder of this chapter will suggest how these downfalls might be avoided in the utilisation of community models in science, with consideration of simple models, mid-level models, and an appeal to scientific realism literature.

7.10. Where to Go from Here: A Space for the Simple, Humble Model?

This chapter has so far shown that, for each of the models considered, there are features responsible for deriving the models' conclusions which do not have real-world referents. The question might then arise: what can be done so that we can better leverage community modelling in astrobiology and other sciences?

It might appear as if the only way to get accurate results from our model that map onto the real world is to add more and more complexity into our model. We might ask whether the limit of maximum transferability of model conclusions to the real world only comes when our models perfectly mirror their target system. When target systems constitute entire social-scientific networks, such a feat is presently computationally impossible.

However, if our target system is much smaller, and the assumptions of the model are benign and explicitly stated, there is indeed a space for simple models in science and elsewhere. The crucial requirement is that the working parts of the model (that is, any feature which leads to the model's conclusions) must be true or approximately true to the target system. This requirement was not satisfied by the four models evaluated so far in this chapter. But satisfying this requirement can be done, and the model need not be overly complicated to do so.

One example of a simple model whose conclusions are transferable to its target system is a novel model of Jury accuracy (Gillen, 2025b). In this model, I group agents into a group of "jurors", who must come to a majority vote on the innocence or guilt of a defendant. I test whether adding less accurate jurors to a highly accurate group will increase or decrease the group's overall accuracy. The only assumption I bring into the model is that all jurors are competent, where competence is defined as an above 50% chance of individual jurors voting correctly.

To give a concrete example, I test whether adding a juror who is only 51% competent (this meaning that they correctly identify innocence or guilt 51% of the time and get it wrong 49% of the time) to a group of jurors who are, on average, 95% accurate, increases or decreases the accuracy of the group.

This setup is a reproduction of Condorcet's Jury Theorem, developed by the Marquis de Condorcet in 1785 (De Condorcet, 2014). This mathematical theorem is quite straightforward, but its conclusions are surprising. To summarise, the theorem states that, so long as individual jurors are competent, adding more jurors to a group that decides by majority rule will increase the average accuracy of the group. This means that, in the example above, adding a juror whose accuracy is only slightly better than the flip of a coin to a highly accurate group is actually beneficial to the group; the group's average accuracy will go up.

The mathematical proof of this result is as follows. Given a set of voters, N, voting by majority rule between two outcomes where each individual, X_i , has the same probability p of selecting the correct outcome and probability $1 - p$ of selecting the incorrect outcome, the probability of the majority vote, V_m , being the correct one, V_C , increases with N and tends to 1 as $N \rightarrow infinity$. This, so long as the following conditions are met:

1. The voters must be independent, given the correct outcome
2. The voters must be competent, and this in equal measure. That is: all voters must have the same probability, p, of selecting the correct outcome, and $p > ½$

The proof of the theory stems from the law of large numbers, as given below (Nitzan, 2009, p.206).

$$\lim_{n\to\infty} P(V_m = V_C) = \lim_{n\to\infty} \sum_{i=\frac{N+1}{2}}^{N} \binom{N}{i} p^i (1-p)^{N-i} = 1$$

The unity result of Condorcet's jury theorem rests upon the limit of an infinite number of jurors. However, given finite *N*, and the assumptions, the theory states that the probability of the majority vote being correct will tend to increase with increasing jury size. Moreover, and significantly, Nitzan (2009) has shown that the requirement that all voters must have the *same* competence can be relaxed, so long as *on average* the jurors are competent.

My model gives a spread of competencies to jurors, all of which are above 50%, and hence Condorcet's jury theorem will hold: the more jurors I add to a group, the better the accuracy of the group (Nitzan, 2009). And, indeed, this result is found by the model (Gillen, 2025b).

To summarise, the conclusion of the model is this: assuming jurors are competent (defined as above 50% accuracy), and the decision is made on a majority rule, adding more jurors will, on average, increase the competency of the group. This is even the case when the new jury has a lower competence than the existing group's competence, or even when this is lower than the lowest-competence member of the group.

This model is relatively simple. It does not account for nuances in human behaviour such as the impact of deliberation. The effect of deliberation on jury accuracy is discussed in detail in Hedden (2017) and, correspondingly, the effects of relaxing the independence assumption are also explored in the novel Condorcet model (Gillen, 2025b). Yet, despite the model's simplicity, its conclusions are surprising and have far-reaching implications. Debates, especially with regards to democracy, are rife (see, for example, Brennan, 2016; Reiss, 2019). The theorem holds, but the content of these discussions often centres on whether the competency and independence assumptions have been met in real-world systems. With these assumptions clearly stated, those wishing to apply the theory to real-world contexts can begin the task of deciding whether their system satisfies the assumptions.

In contrast, the four models I have discussed earlier in this chapter have hidden assumptions that just cannot be easily replicated in real-world systems. For example, Avin's (2019) vision length of two tiles is a hidden, working assumption within his model that is simply too abstract to accurately replicate in the real world. So too, is Hong and Page's (2004) three-number heuristic set which strives to represent how people problem-solve. These working assumptions are not benign in the way that the assumptions of Condorcet's jury theorem model are; assuming competence, we can confidently apply the model to real-world systems where individuals vote for the correct outcome more than 50% of the time.

The point is that simple models can be informative, so long as the assumptions that derive the conclusions are not wholly unrealistic or unrepresentative of any real-world feature. The question then becomes whether these assumptions can be satisfied in the real-world. The model of Condorcet's jury theorem can be applied to some real-world juries (controlling for deliberation, as in Hedden, 2017). And the successful application of this simple model in this context is partly due to the target system being specific and quite simple itself (e.g., the act of voting).

However, the application of the Condorcet jury theorem model to, say, democracy is more problematic. The target system might be considered too broad for such a simple model, and hence, our model misses many relevant features of the real world that would influence outcomes. Moreover, the application of Condorcet's jury theory to democracy assumes that there is a *correct* outcome, which becomes difficult to define in the context of politics. Consequently, attempts have been made to accommodate this voter-specific truth into Condorcet's Jury Theorem (List & Spiekermann, 2016; List & Goodin, 2001; Berg, 1996).

So, simple models have a place in science and elsewhere. But only in so far as the relevant, working features of their target system are simple. What of more complex systems? Many of the models considered in this chapter (especially Avin, 2019; Weisburg & Muldoon, 2009; and the novel model on disagreement in astrobiology) have complex target systems whereby it is expected for nuanced social features to play a working role in the outcome in question (e.g., which funding strategy is optimal, or how unorthodox scientists should behave).

We might still reasonably want to utilise models to inform us of these complex systems. However, we risk our simple models omitting crucial real-world features which do impinge on the outcomes under question. In such scenarios, simple models may not be appropriate, but we have seen throughout this chapter that great care needs to be taken when features are introduced into models to account for real-world phenomena. Adding more features *might* result in a closer correspondence between models and their target system, but equally, the introduction of unrepresentative model features may act to invalidate the model completely. The next section of this chapter will appeal to literature on mid-level models as a guide for utilising complex models that avoid introducing unrepresentative model features.

7.11. Mid-Level Models

As discussed in §7.1.1, epistemic community modelling has the power to highlight parameter dependencies that are otherwise obscured in the complexities of multi-parameter fields like astrobiology. To emphasise again the points made in §7.2.1: models cannot tell us anything synthetically new about our

target system; rather, they can illuminate unforeseen, yet analytically derivable, relationships. This still allows models to be useful and informative (though in a weakened sense) tools.

However, to maximise their utility, these tools must match their relevant target system as closely as possible. Sometimes this relevant target system can be quite simple, as is the case with a group of individuals voting by majority rule. However, sometimes the relevant target system is complex, and models of such need to capture this complexity in an accurate way. Mid-level models are one attempt at this. Harnagel (2019) proposes and develops an approach to such mid-level modelling.

Harnagel specifically proposes a mid-level model of scientific funding allocation which uses bibliometric data to create a dynamic and more realistic landscape. In this way, Harnagal's mid-level model takes the complex and relevant feature of landscape shape and strives to achieve a closer correspondence between this and the real-world feature it attempts to emulate: the significance of projects and the similarities between projects.

Harnagel's model is broadly based on Avin's (2019) model. However, Avin's landscape consists of randomly located peaks, whereas the location and height of Harnagel's peaks are informed by empirical data about citation counts. These citation counts are taken to broadly track the significance of that project; hence, citation counts correspond to patch height in the landscape. Following this, Harnagel (2019) incorporated data on the similarity in references and keywords in published papers. This was used to inform how closely related projects are and, hence, where these projects should be positioned relative to each other in the landscape; papers which share many references and keywords are positioned close together. Finally, Harnagel's model is dynamic, and so the landscape evolves over time. This is achieved by the landscape being updated each round with the real-world citation data from the subsequent year.

Each of these additions ensures a more realistic and representative landscape. This chapter has already discussed the effects of Weisburg and Muldoon's very smooth, two-peak landscape. The small number of local maxima (two) results in an equilibrium state where agents can easily get stuck on one peak and miss out on half of the landscape. Such a landscape shape, therefore, favours an element of randomness. The shape of the landscape is, therefore, a non-neutral feature of the model that is capable of introducing a bias for particular agents. There exists a need, then, to carefully and intentionally choose the model's landscape and the Harnagel's mid-level model (2019) does this well.

7.12. Scientific Realism and Community Modelling

The discussions thus far concerning the ability of simple and complex computer models to tell us accurate truths about the real world have a lot of cross-over with the scientific realism debate. This final section will pool on this literature to offer an answer to the question: how can we ensure that our imperfect and approximated models tell us real-world truths?

Before we address this question, let us briefly summarise the relevant features of the realism debate, before highlighting how the debate might aid us in this context. The development of science throughout history might resemble less of the gradual putting together of a single puzzle, and more of a continual throwing out of the puzzle to start a higher resolution one. Such are these jumps between theoretical foundations that some would render them paradigm shifts within scientific revolutions (Kuhn, 1970). Despite these radical theory changes, empirical success can often be traced across an array of different, and sometimes contradictory, theories in the history of science. Hence it is against this backdrop that the contentious debate of scientific realism is incident.

Scientific realism is an epistemically rich position, whereby the realist believes that our most successful scientific theories are so because they contain true statements about the world. At least some facets of the theory do genuinely refer to the world in the way represented by the theory. Such an example might be that the success of predicting planetary motion is, in part, because the planets genuinely orbit the sun. The scientific realist has a more epistemically optimistic outlook than the instrumentalist, whose pragmatic view of scientific theories does not concern any deeper commitment to the existence of the terms in said theory (Duhem, [1906] 1954; Rowbottom, 2011). The realist is also more committed to theory tenets than the constructive empiricist (an offshoot of instrumentalism), who does not commit to the existence of unobservables within theories, i.e., electrons (Van Fraassen, 1980 is a notable advocate of constructive empiricism). However, what counts as an unobservable has been the focus of much debate, with attacks (e.g., Sober, 1985; Musgrave, 1985) and defences (e.g., Muller, 2008; 2004). Conversely, the realist wishes to assert a causal link between scientific success and truth and would deem success born out of falsity to be an undesirable and absurd circumstance.

The pivot of the realism debate has, on one side, the no-miracles argument, and on the other side sits the pessimistic induction. The no-miracles argument was advanced by Putnam (1975) and has a strong intuitive appeal. It stands as the flagship for the general realist position and might be summarised as so: if our best scientific theories are not a genuine representation of the world, i.e., their contents do not comprise genuine truth, then the astonishing empirical success of such theories must be attributed to a mere miracle. Such a position aligns with intuition and is hard to shake off — so long as one does not dig too deep into the subsequent implications.

However, through this further analysis, the pessimistic induction raises its head. This counter-realist argument takes a comb through the history of scientific theories and identifies the myriad of highly successful yet false theories (at least by the standards of today's theories). Laudan's (1985) infamous list compiles a selection of such apparently false yet successful theories that stand to make a mockery of the no-miracles argument. These historical cases enjoyed empirical successes and were consequently held as containing elements of truth during their heyday. However, time and time again, these successful theories were replaced with different and contradicting subsequent theories.

The inductive step, and the crux of the attack, is this: if past successful and convincing theories have been continually disproved and replaced with other successful and convincing theories, how is one justified in believing our best current theories will not succumb to the same fate? Allowing for a privileged and exceptional view of current scientific theories would be biased and unjustified. It is necessary to note that we need not require any subsequent theory to actually be true in order for this pessimistic induction to render belief in multiple theories irrational. Indeed, as is the case with many examples throughout history, if two theories are contradictory, it is impossible for any more than one of the theories to be correct, at least when taking the theories in their entirety. Hence, any serious realist who is convinced by the no miracles argument and aims to invalidate the pessimistic induction is saddled with the task of individually tracing through the history of science to identify specific reasons why abandoned yet successful theories were accurate in their cases, with an array of strategies depending on the brand of realism invoked.

Nuanced and more sophisticated forms of realism have emerged over time to navigate this difficult space between the no-miracles argument and the pessimistic induction. The various versions are best grouped broadly under two categories: selective realism and structural realism. Both are more discerning with which specific theory elements are responsible for the success, and these elements alone warrant a realist commitment: the realist would claim that they truly reflect the way the world is.

The difference between selective realism and structural realism is the kind of constituents believed to be responsible for the theory's success. Selective realist approaches (as pioneered by Psillos, 1999, 1996 and Kitcher, 1993) aim to identify and isolate the working parts of a theory from the idle parts, hence otherwise known as the *divide et impera* strategy (Psillos, 1999, 1996; Kitcher, 1993). The working parts alone are necessarily true, and the idle parts can be false, and these parts usually take the form of a concept or details of the relata in the theory. Alternatively, structural realism — as propagated by John Worrall in his powerful "Structural Realism: The Best of Both Worlds?" (1989) but proposed earlier by Poincare [1905] (1952) — mandates that, so long as the structure or syntax is maintained throughout theory change, the details of the relata may change. This structure may concern the mathematical relations within the theory or, as advanced by Votsis (2011), the minimally interpreted mathematical parts. A particular

example might be the continuity of Fresnel's equations for the propagation of light in the ether through to Maxwell's equations of electromagnetism. The semantics of each theory differ, but the structure (or minimally interpreted structure a la Votsis, 2011) is maintained (Worrall, 1994).

So, how does the literature on scientific realism provide a helpful perspective on how to skilfully use computer models to inform us of the real world? If one wishes to be a naïve realist (that is, believing all or almost all parts of successful scientific theories are true), the ability of simple computer models to make accurate predictions about the real world should be troubling. The success of the simple model of jury deliberation shows how, given that jurors are competent, it will benefit a jury to bring in more jurors, even if the new juror is less competent than any existing juror. This model does not perfectly mimic real-world juries; humans are reduced to a for-loop in a computer simulation, for one thing. Yet, so long as the competence assumption is met, the results of the computer simulation extend accurately to the real world.

The naïve realist must reconcile with the fact that a simple model, whose lack of complexity does not accurately mirror the real world, successfully predicts a more complex real-world system. As has been discussed at length within the scientific realism debate, the error in the naïve realist's position is their determination to ascribe realist belief to *every* part of a successful theory/model. If the realist is willing to give up parts of a theory or model and instead only commit to the working parts (in alignment with selective realism), then the success of computer models can be explained.

To recap, the selective realist searches for the working parts of a theory or model and makes an ontological commitment to these parts alone (Psillos, 1999, 1996; Kitcher, 1993). The idle parts of a theory or model can be disregarded as superfluous. Adopting a selective realist view of computer simulations would therefore allow substantial differences between the model and the real-world target system. Humans may be unproblematically reduced to agents in a for-loop, just so long as the essential, working features that lead to the real-world conclusions are captured by the model. Such essential working features might feasibly be the binary choice and the majority rule function.

To elaborate, if your preferred flavour of realism is indeed selective realism, what lessons can be taken regarding how best to utilise computational community modelling? The selective realist should not be concerned by the success of a simple model. Just because the model makes accurate predictions does not mean they must believe every part and only parts within the model are true; the selective realist does not find themselves in the inevitable position of having to believe that jurors are literally bits of code sorted into two bins.

What selective realism does require though is that the parts responsible for the real-world conclusions are present in the model (e.g., the binary choice and the majority rule). If we wish to successfully model real-world systems, selective realism emphasises the need for a close correspondence between the working parts of the real-world system and the model features. However, although this may be possible with some simple models, it becomes difficult as the complexity of the model increases. This is largely on account of how complex models attempt to represent complex real-world features in abstract ways.

An example of this is Avin's (2019) and Weisburg and Muldoon's (2009) representations of project similarity. Avin argues that project similarity is an important feature in how real-world scientists choose their next research proposal (Avin, 2019, §4.3). Hence, we should expect project similarity to be a working feature of funding selection. However, both Avin and Weisburg and Muldoon model project similarity in the abstracted way of physical proximity on a two-dimensional landscape (Avin, 2019, pp.635-636; Weisburg & Muldoon, 2009, §2.1). This should be concerning to the selective realist if they wish to believe the results of both models. The selective realist finds themselves trapped in the following reasoning: 1) I am convinced of the real-world conclusion of these models, 2) a working feature of these models is the physical proximity of projects in a two-dimensional landscape, 3) I am convinced that, in the real world, similar projects physically sit close to each other!

Such reasoning appears misguided due to the requirement that working posits in a successful model must be true features of the real world. Instead, computer models of community structures attempt to recreate the *functional form* of real-world features with model features. To expand on this, let us return to the two-dimensional landscape representation of project similarity (Avin, 2019, pp.635-636; Weisburg & Muldoon, 2009, §2.1). The authors attempt to capture the function that project similarity has in the real world concerning the proposal of research projects. They then attempt to recreate this same function in a different format within the model. The relevant working real-world feature that Avin and Weisburg and Muldoon need to capture is the tendency for scientists to pursue only projects with significant cross-over with their existing research project. By modelling project similarity as physical proximity, it is hoped that the function of the former is replicated in the latter.

This view of computer modelling strikes parallels with structural scientific realism. As discussed above, structural realism acknowledges that a theory or model can derive correct conclusions, even if the details of the theory are false. This is just so long as the structure or syntax of the theory or model is correct. In the case of the models discussed in this chapter (especially Avin, 2019; Hong & Page, 2004; Weisberg & Muldoon, 2019), attempts have been made to represent real-world features by model features that function the same way as their real-world counterparts. We need not ascribe realist commitment to the representations themselves (e.g., physical proximity of projects, physical height of projects, or appearing

and disappearing hills). But, if we are to believe the model's conclusions are transferable to the real world, we should believe that the model's features capture the same functional form as their real-world counterparts.

The arguments presented in §7.6, §7.7, and §7.8 of this chapter have attempted to undermine the transferability in functional form from working features of the models in question to working features of their target systems. I have unpicked how these working features within the models lack functional representation in the real world. Structural realism about models allows diversity in how working structures are dressed up. But their functional form must be true to their target system.

The methods proposed in the structural realism literature are therefore helpful guides for checking the validity of our computer models of scientific communities. Structural realism embraces the possibility that target systems can be modelled differently, with different models producing the same accurate predictions. However, great care is needed to ensure that the abstract representations of real-world features are functionally equivalent to their real-world counterparts.

In summary, under a structural realist lens, models of scientific community structures can have a place in science. If they succeed in capturing the working structures of their target system, and they do not add additional unrepresentative working features, their conclusions can be said to extend to the real world. However, the models considered in this chapter have failed to adequately mirror the functional form of the features in question. More focus on how to capture and represent the structure of real-world features is needed if the conclusions of these models are to extend to real-world communities.

7.13. Conclusions

This chapter had three primary objectives. First, I presented the results of a simple model concluding the benefits of disagreement in priors within astrobiology (Gillen, 2025a). Second, I presented a review of three seminal epistemic community models. By examining the instrumental assumptions and mechanisms within each model, I hope to have shown that the conclusions lack real-world representation. Finally, I have considered how we might guard against unrepresentative models and have pooled literature within the scientific realism debate to provide theoretical support for my claims.

It can be tempting to take the results of simulated community models and extend them to our real-world systems. The novel model of disagreement within astrobiology presented in §7.3 stands as a cautionary

tale of such over-extension from the model to the real-world target system. In this model, the conclusion in favour of disagreement derives from a quirk in the model that has no real-world analogue.

Similarly, the models of Avin (2019), Hong and Page (2004), and Weisburg and Muldoon (2009) have been found to be similarly unrepresentative, and this discussion was had in §7.6, §7.7., and §7.8. Each of these models contains features that are essential to the conclusions in favour of diversity, and yet these features fail to capture the functions of their real-world analogues. Specific examples of this include Avin's disappearing and reappearing peaks (2019, see especially §4.5), Hong and Page's three-number heuristic set (2004, p.16386), and Weisberg and Muldoon's twin-peak landscape (2009, p.234).

For model conclusions to be valid, I have argued that the features responsible for deriving the conclusions must be functionally equivalent to features in their target system. Hence, the question may arise as to whether the accuracy of models only increases with increasing correspondence to the target system, up to the limit of the model being a perfect clone of the target system. If the answer to this is yes, then the role of simple computer models in science looks precarious.

However, §7.10 has defended a space for simple computer models of scientific communities. This was done by taking a narrower view of the target system. If these relevant working features in the target system are simple, then the results of a simple model can be representative. The novel model of Condorcet's jury theorem presented in §7.10 is one example of such a simple model of a tight target system (Gillen, 2025b).

However, target systems in science cannot often be reduced to a handful of working features. In such scenarios, more complex models are required, whereby more assumptions and data must be fed into the model. §7.11 has addressed how representation can be maintained between complex models and their target systems. As complexity increases, it is easy for working features to stray from being the functional equivalents of their real-world features. One way to partially mitigate this is to embed more real-world data into the model, as is done with mid-level models (Harnegal, 2019).

Finally, I have appealed to scientific realism literature in §7.12 to provide a helpful perspective on both how and why computer models can be informative in science. The endorsement of a structural realist view of community modelling explains how it is possible to get real-world truths out of an imperfect model. Models need not contain superfluous content or even dress up the structural content as it is in the real world. Rather, models can produce accurate results, just so long as the working model features have the same functional form as the working features in the real world.

To conclude, computer models are analytic in nature, and hence we cannot get more certainty out than the data we put in. Having said this, computer modelling of scientific communities can still be highly

informative; these community models pose an exciting and potentially fruitful arena in which to test hard-to-see parameter sensitivity in complex systems such as astrobiology. However, their successful use is dependent on the close correspondence between the functional form of working model features and the functional form of the real-world working features in the target system. Such correspondence has been missing in many popular community models (e.g., Avin, 2019; Hong & Page, 2004; Weisburg & Muldoon, 2009). Structural realism endorses the use of computer modelling of scientific communities (both simple and complex). It is possible to recreate the form of real-world systems in abstract, computational ways. However, modellers and reviewers must closely analyse the working tenets of these models to ensure that 1) conclusions have not been baked in, and 2) the functional forms of key tenets are representative of the functional forms of the real-world systems they strive to represent.

Chapter Eight

Conclusions

8.1. Statement of Conclusions

In a field with content as rich as the study of life in the universe, the need for optimal decision making cannot be understated. Mathematical and computational community models have great promise in aiding with optimising decision making. However, deep-rooted uncertainty, arising from the newness and broadness of astrobiology, acts to undermine the leveraging of these models. This thesis comes as a response to the need for a thorough analysis of the effects that uncertainty has on utilising mathematical and computational community models within astrobiology.

The following five primary conclusions have been presented:

> **C1:** I proposed and defended a new definition of biosignature: *any phenomenon for which biological processes are a known possible explanation and whose potential abiotic causes have been reasonably explored and ruled out* (Gillen et al., 2023, p.1228). This is strong enough to be meaningful but leaves room for uncertainty over the list of possible explanations captured by the problem of unconceived alternatives (Stanford 2001, 2006a). This conclusion was discussed in Chapter Two and Chapter Three of this thesis.
>
> **C2:** I proposed a new corresponding definition of potential biosignature: *any phenomenon for which biological processes are a known possible explanation but whose potential abiotic causes have not yet been reasonably explored and ruled out* (Gillen et al., 2023, p.1238). This was argued for in Chapter Two and Chapter Three.
>
> **C3:** In light of the vast and unexplored potential research areas constituting the field of astrobiology, theoretical arguments exist for funding high-uncertainty, high-payoff research. These

arguments do not apply to high-risk, high-payoff research. This conclusion was discussed in Chapter Four and Chapter Five of this thesis.

C4: High uncertainty surrounds key fundamental probabilities in astrobiology, such as the probability of abiogenesis. This uncertainty means that astrobiologists disagreeing about fundamental probabilities might be beneficial. This conclusion is found in Chapter Six.

C5: The deployment of computer models of scientific communities, such as astrobiology, requires a close correspondence between the functional form of working real-world features and the functional form of working model features. This is consistent with a structural realist view of computer modelling. This conclusion was discussed in Chapter Seven.

8.2. Summary of Chapters

Having outlined the foundational motivation for this thesis and its situation in current literature in Chapter One, Chapter Two then considered a central yet problematic term in astrobiology, this being the term *biosignature*. I proposed a novel definition, whereby a biosignature is *any phenomenon for which biological processes are a known possible explanation and whose potential abiotic causes have been reasonably explored and ruled out* (Gillen et al., 2023, p.1228). This definition stands between existing strong definitions (e.g., Des Marais et al., 2003) and existing weak definitions of biosignature (e.g., Pohorille and Sokolowska, 2020; Catling et al., 2018; Schwieterman et al., 2018). In doing so, it avoids the limitations levelled on these existing definitions. Specifically, the new definition provides a practical response to the problem of unconceived alternatives (Stanford, 2006a, 2001) that is so prevalent in astrobiology.

A complementary definition of *potential biosignature* is also provided in Chapter 2. I define this as: *any phenomenon for which biological processes are a known possible explanation but whose potential abiotic causes have not yet been reasonably explored and ruled out* (Gillen et al., 2023, p.1238). This definition works in tandem with the novel definition of biosignature and captures the active science that needs to be done to arrive at a biosignature claim.

Both the definition of biosignature and of potential biosignature are epistemic in nature. Subsequently, they have come under attack in recent literature, and these challenges are addressed and responded to in Chapter Three. Cowie (2023a), in particular, argues in favour of an objective (akin to an ontic) definition of biosignature as opposed to a subjective (epistemic) one. Indeed, ontic definitions are far more common and, in some senses, intuitive than epistemic ones. On the contrary, epistemic definitions are fewer and farther between. One such example of an epistemic definition discussed at length in Chapter Three is

that of legal guilt. A defendant is legally guilty just if a judge and/or jury has deemed them to be. In the absence of this epistemic (or, perhaps more fittingly, doxastic) position, the defendant does not meet the criteria of legal guilt. I have argued in Chapter Three that biosignatures should be regarded analogously to legal guilt due to the simultaneous high degree of uncertainty surrounding life detection and the need for a term representing the limit of our available information. Hence, by framing biosignatures epistemically rather than ontically, we ensure that the term remains flexible enough to accommodate future discoveries while still maintaining its scientific rigour.

Following this, Chapters Four and Five considered the value of high-uncertainty, high-payoff projects in astrobiology. Chapter Four began this with a critique of how risk and uncertainty are being used in astrobiology funding calls. NASA's Research Opportunities in Space and Earth Science (ROSES) have specifically called for more high-*risk* projects (NASEM, 2018). I have argued that such language is misleading and may result in the irrational funding of projects with a known low probability of success. Instead, there may be a reason to fund high-uncertainty, high-payoff projects, and I provided the theoretical motivation for this in Chapter Four: true uncertainty (as opposed to risk), that arises from a rich and unexplored research area, can point us to fascinating and presently obscured discoveries.

Having argued for the funding of high-uncertainty, high-payoff research, Chapter Five assessed whether SETI research, which is privately funded, would fall under the category of high-uncertainty, high-payoff research. By using the Drake equation as a framework by which the prevalence of accessible extraterrestrial intelligence can be estimated, I concluded that SETI research is indeed high-uncertainty. An overview of the state-of-the-art for pinning down the seven parameters in the Drake equation was provided, and it was found that uncertainty is rife, especially in the final four parameters. However, due to a lack of consensus over whether the potential payoffs of SETI research would be positive, I found that it is problematic to call SETI research high-payoff. As such, SETI does not classify as high-uncertainty, high-payoff research in the same way that many alternative astrobiology projects do.

Chapters Six and Seven then considered how disagreement (on account of uncertainty) amongst individual scientists might be beneficial to astrobiology. Chapter Six took the case study of Oumuamua to provide an illustrative real-world example of how disagreement can arise from uncertainty over the probability of detecting extraterrestrial intelligence. Disagreement over Oumuamua's origin was rampant in the years immediately following its detection (e.g., Bernger & Seligman, 2023; Bannister et al., 2019; The 'Oumuamua ISSI Team, 2019; Loeb, 2018). This chapter attempted to diagnose what, precisely, the disagreement was over when considering Oumuamua's origin. I evaluated two previous viewpoints on this (Cowie, 2021, 2023a; Matarese, 2022). However, I argued in favour of Lineweaver's (2022b) diagnosis that differing priors were responsible for the divergence in opinion.

Having endorsed Lineweaver's (2022b) view, I expanded on this in Chapter Six within a Bayesian framework. Within this framework, it is straightforward to split up the prior probability of detecting extraterrestrial life and the confirmatory weight of the evidence for that prior. In this way, I argued that individuals on each side of the debate may well be acting rationally (by performing a Bayesian update), and even weighting the evidence equally. However, on account of beginning with different priors, the conclusions have diverged. I concluded this chapter by appealing to community modelling literature (Weisberg & Muldoon, 2009; Muldoon, 2013) to suggest the beneficial role that differing priors may have in astrobiology.

The hypothesis that disagreement over fundamental priors in astrobiology may benefit the field was then tested in Chapter Seven. Here, I considered the conclusions of a simple novel computer model which could be interpreted to argue in favour of diversity of priors within astrobiology. However, this model is then used as a cautionary tale on how easy it is to overinterpret computational community models to make real-world conclusions. I argued that the conclusions were baked into my model and that this is a common feature of community models.

To further warn of the dangers of overinterpretation and the baking-in of conclusions in computational community models in science, I took three other community models commonly used to argue in favour of diversity in science (Avin, 2019; Hong & Page, 2004; Weisburg & Muldoon, 2009). By unpicking the inner workings of these three community models (alongside my own model), I found the conclusions to be over-interpreted and attributed to an element of circularity. Although there may be good reason to expect diversity to be beneficial to science, we must take care when employing community models to argue this.

Chapter Seven then concluded with a discussion on how computer community modelling can be informative in science, and this was done through the lens of structural realism about models. The analytic nature of computer models indeed precludes them from telling us anything synthetically new about the world, thus, we cannot get certain data out of these models by inputting uncertain data. However, I concluded that computer modelling of scientific communities can still be highly informative; these community models pose an exciting and potentially fruitful arena in which to test parameter sensitivity in complex systems. Their successful use, though, is contingent on the close correspondence between the functional form of working model features and the functional form of the real-world working features in the target system. Such correspondence has been missing in many popular community models (e.g., Avin, 2019; Hong & Page, 2004; Weisburg & Muldoon, 2009). Structural realism endorses the use of computer modelling of scientific communities (both simple and complex). However, if these models are to be

valuable tools in astrobiology, a concerted effort is needed on the part of modellers and reviewers to ensure that this functional form has been accurately recreated in the model.

8.3. Final Takeaway

Uncertainty is indeed a pervasive feature of astrobiology. It challenges the terms we use (e.g., biosignatures, as in Chapters Two and Three); undermines the use of maximising expected utility to aid in funding decisions (e.g., as in Chapters Four and Five); leads to disagreement within Bayesian frameworks (e.g., as in Chapter Six); and, limits the precision of information gleaned from computational community models (e.g., as in Chapter Seven).

However, the uncertainty in astrobiology need not paralyse us. C1 and C2 respond to how uncertainty has undermined the key term of biosignature and propose a useful and accurate alternative. C3 makes a practical suggestion on how funding should be allocated to maximise utility within an uncertain context. C4 states that disagreement might be prevalent in astrobiology but may benefit the community. C5 proposes how computational community models can maximise their usefulness, even within complex fields.

The uncertainty in astrobiology, while presenting challenges, also offers immense promise. It is a reflection of the vast, unexplored frontiers that lie ahead, and with this uncertainty comes the opportunity for groundbreaking discoveries. While the uncertainties within our mathematical and computational tools may seem restrictive, they also serve as a guiding tool to uncharted territories and new avenues of exploration. Uncertainty may leave us operating in the dark, but ultimately, it is within this uncertainty that the true potential of astrobiology resides, waiting to illuminate the unknown. Future astrobiological research must embrace this uncertainty not as a hindrance, but as a pathway to discovery, integrating robust mathematical and computational methodologies while remaining vigilant to their epistemic limitations.

References

Agrawal, M., Kanitkar, M., & Vidyasagar, M. (2021). Modelling the spread of SARS-CoV-2 pandemic-Impact of lockdowns & interventions. *Indian Journal of Medical Research*, *153*(1-2), 175-181.

Akins, A. B., Lincowski, A. P., Meadows, V. S., & Steffes, P. G. (2021). Complications in the ALMA detection of phosphine at Venus. *The Astrophysical Journal Letters*, *907*(2), L27.

Alexander, J. M., Himmelreich, J., & Thompson, C. (2015). Epistemic landscapes, optimal search, and the division of cognitive labor. *Philosophy of Science*, *82*(3), 424-453.

Aller, L. H., Fernie, J. D., Chaisson, E. J. and Brecher, K. "star". *Encyclopedia Britannica*, (2024), Available at: https://www.britannica.com/science/star-astronomy. [Accessed 20 March 2024].

Almár, I. (2001, September). How the Rio Scale should be improved. In *52nd International Astronautical Congress Preprints, IAF, Toulouse*.

Almár, I., & Tarter, J. (2011). The discovery of ETI as a high-consequence, low-probability event. *Acta Astronautica*, *68*(3-4), 358-361.

Allwood AC, Rosing MT, Flannery DT, Hurowitz JA, Heirwegh CM. (2018). Reassessing evidence of life in 3,700-million-year-old rocks of Greenland. *Nature*. 2018 Nov; 563(7730):241-244. doi: 10.1038/s41586-018-0610-4.

Archibald, J. D. (2014). *Aristotle's ladder, Darwin's tree: the evolution of visual metaphors for biological order*. Columbia University Press.

Ashworth, A., & Player, E. (2005). Criminal Justice Act 2003: the sentencing provisions. *The Modern Law Review*, *68*(5), 822-838.

Averner, M. M., & MacElroy, R. D. (1976). *On the habitability of Mars: An approach to planetary ecosynthesis* (No. NASA-SP-414).

Avin, S. (2019). Centralized funding and epistemic exploration. *The British Journal for the Philosophy of Science*.

Ayer, A. J. (1956). The Problem of Knowledge.

Bader, D., Covey, C., Gutowski, W., Held, I., Kunkel, K., Miller, R., ... & Zhang, M. (2008). Climate models: an assessment of strengths and limitations.

Balbi, A., & Lingam, M. (2023). Beyond mediocrity: how common is life?. *Monthly Notices of the Royal Astronomical Society*, *522*(2), 3117-3123.

Bannister, M. T., Bhandare, A., Dybczyński, P. A., Fitzsimmons, A., Guilbert-Lepoutre, A., Jedicke, R., ... & Ye, Q. (2019). The natural history of 'Oumuamua. *Nature astronomy*, *3*(7), 594-602.

Barrett, J. A. (2014). Description and the Problem of Priors. Erkenntnis 79 (6):1343-1353.

Bartal, I. B. A., Decety, J., & Mason, P. (2011). Empathy and pro-social behavior in rats. *Science*, *334*(6061), 1427-1430.

Basalla, G. (2006). *Civilized life in the universe: Scientists on intelligent extraterrestrials*. Oxford University Press.

Baum, S. D., Haqq-Misra, J. D., & Domagal-Goldman, S. D. (2011). Would contact with extraterrestrials benefit or harm humanity? A scenario analysis. *Acta Astronautica*, *68*(11-12), 2114-2129.

Baumgartner, R. J., Van Kranendonk, M. J., Wacey, D., Fiorentini, M. L., Saunders, M., Caruso, S., ... & Guagliardo, P. (2019). Nano− porous pyrite and organic matter in 3.5-billion-year-old stromatolites record primordial life. *Geology*, *47*(11), 1039-1043.

Bayes T. (1764). An essay towards solving a problem in the doctrine of chances. *Philos Trans R Soc Lond* 1764; 53: 370–418.

Bell, M. S. (2007). Experimental shock decomposition of siderite and the origin of magnetite in Martian meteorite ALH 84001. *Meteoritics & Planetary Science*, *42*(6), 935-949.

Belton, M. J. S. et al. (2018). The excited spin state of 1I/2017 U1 'Oumuamua. *Astrophys. J. Lett.* **856**, L21.

Benneke, B., Wong, I., Piaulet, C., Knutson, H. A., Lothringer, J., Morley, C. V., ... & Fraine, J. (2019). Water vapor and clouds on the habitable-zone sub-Neptune exoplanet K2-18b. *The Astrophysical Journal*, *887*(1), L14.

Benner, S. A. (2010). Defining life. *Astrobiology*, *10*(10), 1021-1030.

Berg, S. (1996). Condorcet's Jury Theorem and the reliability of majority voting. *Group Decision and Negotiation*, *5*, 229-238.

Bergner, J. B., & Seligman, D. Z. (2023). Acceleration of 1I/'Oumuamua from radiolytically produced H2 in H2O ice. *Nature*, *615*(7953), 610-613.

Bialy, S., & Loeb, A. (2018). Could solar radiation pressure explain 'Oumuamua's peculiar acceleration?. *The Astrophysical Journal Letters*, *868*(1), L1.

Bianciardi, G., Miller, J. D., Straat, P. A., & Levin, G. V. (2012). Complexity analysis of the Viking labeled release experiments. *International Journal of Aeronautical and Space Sciences*, *13*(1), 14-26.

Billingham, J. (Ed.). (1999). *Social Implications of the Detection of an Extraterrestrial Civilization: A Report of the Workshops on the Cultural Aspects of SETI Held in Three Sessions, October 1991, May 1992, and September 1992 at Santa Cruz, California*. SETI Press.

Binmore, K. (2008). Rational decisions. In *Rational Decisions*. Princeton university press.

Binzel, R. P. (1997). A Near-Earth Object Hazard Index. *Annals of the New York Academy of Sciences*, *822*(1), 545-551.

Bitten, R. E., Shinn, S. A., & Emmons, D. L. (2019). Challenges and potential solutions to develop and fund NASA flagship missions. In *2019 IEEE Aerospace Conference* (pp. 1-13). IEEE.

Bohman, J. (2006). Deliberative democracy and the epistemic benefits of diversity. *Episteme*, *3*(3), 175-191.

Bohr, N. (1913). 'On the Constitution of Atoms and Molecules', Philosophical Magazine (6), 26, pp. 1-25, 476-502 and 857-875.

Bolton, S. J., & Juno Science Team. (2010). The juno mission. *Proceedings of the International Astronomical Union*, *6*(S269), 92-100.

Bostrom N., (2002). *Anthropic Bias: Observation Selection Effects*. New York: Routledge. Chap. 1, 1.

Boudreau, K. J., Guinan, E. C., Lakhani, K. R. and Riedl, C. (2016). "Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science", *Management Science,* 62, pp. 2765–83.

Buckland, W. W. (1963). A Text-book of Roman Law from Augustus to Justinian (3 ed.). Cambridge: Cambridge UP. pp. 695–6.

Buick, R. (1990). Microfossil recognition in Archean rocks; an appraisal of spheroids and filaments from a 3500 my old chert-barite unit at North Pole, Western Australia. *Palaios*, **5**(5), 441–459.

Butler, D. A. (2017). Who owns the moon, mars, and other celestial bodies: lunar jurisprudence in corpus juris Spatialis. *J. Air L. & Com.*, *82*, 505.

Campion, N. (2016). The moral philosophy of space travel: A Historical Review. In *Commercial Space Exploration* (pp. 9-22). Routledge.

Carnot, S. (1824). Reflections on the motive power of fire, and on machines fitted to develop that power. *Paris: Bachelier*, *108*(1824), 1824.

Carter B., (1974). Large number coincidences and the anthropic principle in cosmology. In Longair MS (ed.), *Confrontation of Cosmological Theories with Data*. Dordrecht: Reidel, pp. 291–298,.

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

Cassan, A., Kubas, D., Beaulieu, J. P., Dominik, M., Horne, K., Greenhill, J., ... & Wyrzykowski, Ł. (2012). One or more bound planets per Milky Way star from microlensing observations. *Nature, 481(7380),* 167-169.

Catling D., Kiang N., Robinson T., Rushby A., & Genio A. (2018). "Exoplanet biosignatures: a framework for their assessment". *Astrobiology*.

Chamberlin Law Firm (2018). *Difference Between Legal & Factual Guilt | Tampa, FL*. [online] Chamberlin Law Firm, PA. Available at: https://chamberlinlawfirm.com/courtroom-difference-legal-factual-guilt/.

Chapman, B. B., Brönmark, C., Nilsson, J. Å., & Hansson, L. A. (2011). The ecology and evolution of partial migration. *Oikos*, *120*(12), 1764-1775.

Chopra, A., & Lineweaver, C. H. (2016). The case for a Gaian bottleneck: the biology of habitability. *Astrobiology*, *16*(1), 7-22.

Ćirković, M. M. (2012). *The astrobiological landscape: Philosophical foundations of the study of cosmic life* (Vol. 7). Cambridge University Press.

Clarke, A., Morris, G. J., Fonseca, F., Murray, B. J., Acton, E., & Price, H. C. (2013). A low temperature limit for life on Earth. *PLoS One*, *8*(6), e66207.

Cleland, C. E. (2019). *The Quest for a Universal Theory of Life: Searching for Life as we don't know it* (Vol. 11). Cambridge University Press.

Cowie, C. (2021) 'The 'Oumuamua Controversy: A Philosophical Perspective.', Nature astronomy., 5. 526-527.

Cowie, C. (2023a). Arguing about extraterrestrial intelligence. *The Philosophical Quarterly*, *73*(1), 64-83.

Cowie, C. (2023b). New Work on Biosignatures. *Mind*.

Cronin, J., Pizzarello, S., Cruikshank, D.P. (1988). Organic matter in carbonaceous chondrites, planetary satellites, asteroids, and comets. In: *Meteorites and the Early Solar System*, ed. by J.F. Kerridge, M.S. Mathews. University of Arizona Press, Tucson, pp. 819–857.

Crown Office & Procurator Fiscal Service, (2024). "Words and meanings". Available at: https://www.copfs.gov.uk/resources/words-and-meanings/ [Accessed on 07/09/24]

Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray.

Darwin, C. (1874). *The Descent of Man, and Selection in Relation to Sex,* 2nd ed., John Murray, London.

Davies, P., (1993). *The Accidental Universe*, Cambridge University Press, pp. 70–71.

Dawid, R. (2006). Underdetermination and theory succession from the perspective of string theory. *Philos. Sci.* **73**(3), pp.298–322.

Dawid, R. (2007): Scientific Realism in the Age of String Theory, Physics and Philosophy 11: 1-32.

Dawid, R. (2009): On the Conflicting Assessments of the Current Status of String Theory, Philosophy of Science,76/5, 984-996.

Dawid, R. (2013). *String Theory and the Scientific Method*. Cambridge University Press, Cambridge.

Dawid, R. (2016): Modelling Non-Empirical Confirmation, in Ippoliti, E. T. Nickles and F. Sterpetti (eds.), Models and Inferences in Science, pp 191-205, Springer 2016.

Dawid, R. (2018). Delimiting the unconceived. *Found. Phys.* **48**(5), 492–506.

Dawid, R. (2019). The significance of non-empirical confirmation in fundamental physics. In: Dardashti, R., Dawid, R., Thebault, K. (eds.) *Why Trust a Theory?—Reconsidering Scientific Methodology in Light of Modern Physics*. Cambridge University Press, Cambridge. arXiv:1702.01133.

Dawid, R. (2019): The Significance of Non-Empirical Confirmation in Fundamental Physics. In Dardashti, R., R. Dawid and K. Thebault (eds) *Why Trust a Theory? -Epistemology of Fundamental Physics*, Cambridge University Press.

Dawid, R. (2022). Meta-empirical confirmation: Addressing three points of criticism. *Studies in history and philosophy of science*, *93*, 66-71.

Dawid, R. Hartmann, S., and Sprenger, J. (2015). The No Alternatives Argument, *The British Journal for the Philosophy of Science* 66(1), 213-234.

De Condorcet, N. (2014). *Essay on the application of analysis to the probability of decisions rendered by plurality of votes*. Cambridge University Press.

De Finetti, B. (2017). *Theory of probability: A critical introductory treatment* (Vol. 6). John Wiley & Sons.

Des Marais, D. J., Allamandola, L. J., Benner, S. A., Boss, A. P., Deamer, D., Falkowski, P. G., ... & Yorke, H. W. (2003). The NASA astrobiology roadmap. *Astrobiology*, *3*(2), 219-235.

Dick, S. J. (1984). *Plurality of Words: The Extraterrestrial Life Debate from Democritus to Kant*. CUP Archive.

Dick, S. J. (2001). Life on other worlds. In *Encyclopedia of Astronomy & Astrophysics* (pp. 1-4). CRC Press.

Dick, S. J. (2013). *Discovery and classification in astronomy: Controversy and consensus*. Cambridge University Press.

Dick, S. J. (Ed.). (2015). *The impact of discovering life beyond earth*. Cambridge University Press.

Dick, S. J. (2018). *Astrobiology, discovery, and societal impact* (Vol. 9). Cambridge University Press.

Dick, S. J. (2020). *Space, time, and aliens*. Springer International Publishing.

Dick, S. J., & Strick, J. E. (2005). The living universe: NASA and the development of astrobiology. *Journal of the History of Biology*, *38*(2).

Dicke, R.H., (1957). Principle of Equivalence and Weak Interactions. *Rev. Mod Phys.,* **29,** 355.

do Nascimento-Dias, B. L., & Martinez-Frias, J. (2023). Brief review about history of astrobiology. *International Journal of Astrobiology*, *22*(1), 67-78.

Drake, F. (1961). Project Ozma. *Physics Today*. 14 (4): 40–46.

Drake, F. (2006). On-line Debate Astrobiology Magazine. [Online] Available at: http://www.astrobio.net/news/article239.html [Accessed on 8/12/2023].

Duhem, Pierre Maurice Marie, [1906] (1954). *The Aim and Structure of Physical Theory*, Philip P. Wiener (tr.), Princeton: Princeton University Press.

Dunér, D., Holmberg, G., & Persson, E. (Eds.). (2013). *The history and philosophy of astrobiology: Perspectives on extraterrestrial life and the human mind*. Cambridge Scholars Publishing.

Edwards, D. A. (1981). Mathematical foundations of quantum field theory: Fermions, gauge fields, and supersymmetry part I: Lattice field theories. *International Journal of Theoretical Physics*, *20*, 503-517.

Ehman, J., (2019). The Wow! Signal with Discoverer Dr. Jerry Ehman. *Event Horizon* (Interview). Interviewed by John Michael Godier. [Online] Available as: "The Wow! Signal with Discoverer Dr. Jerry Ehman". [Accessed 12/01/2023].

Eikenberry, S. E., Mancuso, M., Iboi, E., Phan, T., Eikenberry, K., Kuang, Y., ... & Gumel, A. B. (2020). To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic. *Infectious disease modelling*, *5*, 293-308.

Elia, D., Molinari, S., Schisano, E., Soler, J. D., Merello, M., Russeil, D., ... & Liu, S. J. (2022). The Star Formation Rate of the Milky Way as Seen by Herschel. *The Astrophysical Journal*, *941*(2), 162.

Emanuel, K. A. (1991). The theory of hurricanes. *Annual Review of Fluid Mechanics*, *23*(1), 179-196.

England J., (2013). Statistical physics of self-replication. *The Journal of Chemical Physics* 139, 121923.

Evans, J. E., & Maunder, E. W. (1903). Experiments as to the actuality of the "Canals" observed on Mars. *Monthly Notices of the Royal Astronomical Society*, *63*, 488-499.

Feehly, C., (2023). SETI Institute gets $200 million to seek out evidence of alien life. [Online] Accessed at: https://www.space.com/searth-extraterrestrial-life-major-funding-boost-seti [Accessed on 10/12/2023].

Fehr, C. (2011). *What is in it for me? The benefits of diversity in scientific communities* (pp. 133-155). Springer Netherlands.

Forgan, D., Wright, J., Tarter, J., Korpela, E., Siemion, A., Almár, I., & Piotelat, E., (2019). Rio 2.0: revising the Rio scale for SETI detections. *International Journal of Astrobiology*, *18*(4), 336-344.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., ... & Barabási, A. L. (2018). Science of science. *Science*, *359*(6379), eaao0185.

Fox, S. W. (1991). Synthesis of life in the lab? Defining a protoliving system. *The Quarterly Review of Biology*, *66*(2), 181-185.

Furnes, H., Banerjee, N. R., Muehlenbachs, K., Staudigel, H., & de Wit, M., (2004). Early life recorded in Archean pillow lavas. *Science*, *304*(5670), 578-581.

Gallup, G. G. (1970). "Chimpanzees: Self-Recognition", *Science*, 167(3914): 86–87. doi:10.1126/science.167.3914.86

Garber, S. J., (1999). Searching for good science-the cancellation of NASA's SETI program. *Journal of the British Interplanetary Society*, *52*(1), 3-12.

Gardner, J. P., Mather, J. C., Clampin, M., Doyon, R., Greenhouse, M. A., Hammel, H. B., ... & Wright, G. S. (2006). The james webb space telescope. *Space Science Reviews*, *123*, 485-606.

Gargaud, M., Mustin, C., & Reisse, J. (2009). Traces of past or present life: biosignatures and potential life indicators?. *Comptes rendus. Palévol*, *8*(7).

Gell-Mann, M. (2018). The Eightfold Way: A Theory of strong interaction symmetry. In *The Eightfold Way* (pp. 11-57). CRC Press.

Gettier, E., (1963). "Is Justified True Belief Knowledge?", *Analysis*, 23(6): 121–123. doi:10.2307/3326922

Gillen, C. (2025a). Bayesian Astrobiologists. Retrieved from osf.io/rz4xj. DOI 10.17605/OSF.IO/RZ4XJ

Gillen, C. (2025b). Condorcet's Jury Theorem for Real Juries. Retrieved from osf.io/ukdtv. DOI 10.17605/OSF.IO/UKDTV

Gillen, C., Jeancolas, C., McMahon, S., & Vickers, P. (2023). The Call for a New Definition of Biosignature. *Astrobiology*, *23*(11), 1228-1237.

Godfrey-Smith, P. (2013). Cephalopods and the evolution of the mind. *Pacific Conservation Biology*, *19*(1), 4-9.

Gold, T. (1985). The origin of natural gas and petroleum and the prognosis for future supplies, Annu.

Golden, D. C., Ming, D. W., Schwandt, C. S., Lauer Jr, H. V., Socki, R. A., Morris, R. V., ... & McKay, G. A. (2001). A simple inorganic process for formation of carbonates, magnetite, and sulfides in Martian meteorite ALH84001. *American Mineralogist*, *86*(3), 370-375.

Golden, D. C., Ming, D. W., Morris, R. V., Brearley, A. J., Lauer Jr, H. V., Treiman, A. H., ... & McKay, G. A. (2004). Evidence for exclusively inorganic formation of magnetite in Martian meteorite ALH84001. *American Mineralogist*, *89*(5-6), 681-695.

Golden, D. C., Ming, D. W., Lauer Jr, H. V., Morris, R. V., Trieman, A. H., & McKay, G. A. (2006). Formation of "Chemically Pure" Magnetite from Mg-Fe-Carbonates Implications for the Exclusively Inorganic Origin of Magnetite and Sulfides in Martian Meteorite ALH84001. In *Lunar and Planetary Science Conference*.

Goldman, A. (1979). What is justified belief. *Justification and Knowledge/Reidel Publishing Company*.

Goldman, A. I., & Shaked, M. (1991). An economic model of scientific activity and truth acquisition. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *63*(1), 31-55.

Goodall, J., (1964). Tool-using and aimed throwing in a community of free-living chimpanzees. *Nature*, *201*(4926), 1264-1266.

Gott, 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*. Norton, New York.

Grantham, P. J., & Wakefield, L. L. (1988). Variations in the sterane carbon number distributions of marine source rock derived crude oils through geological time. *Organic geochemistry*, *12*(1), 61-73.Greaves, J. S., Richards, A., Bains, W., Rimmer, P. B., Sagawa, H., Clements, D. L., ... & Hoge,

Grasset, O., Dougherty, M. K., Coustenis, A., Bunce, E. J., Erd, C., Titov, D., ... & Van Hoolst, T., (2013). JUpiter ICy moons Explorer (JUICE): An ESA mission to orbit Ganymede and to characterise the Jupiter system. *Planetary and Space Science*, *78*, 1-21.

Greaves, J. S., Richards, A., Bains, W., Rimmer, P. B., Sagawa, H., Clements, D. L., ... & Hoge, J. (2021). Phosphine gas in the cloud decks of Venus. *Nature Astronomy*, *5*(7), 655-664.

Green, J., Hoehler, T., Neveu, M., Domagal-Goldman, S., Scalice, D. and Voytek, M., (2021). "Call for a framework for reporting evidence for life beyond Earth". *Nature*, 598(7882), pp.575-579.

Grim, P., Singer, D. J., Bramson, A., Holman, B., McGeehan, S., & Berger, W. J. (2019). Diversity, ability, and expertise in epistemic communities. *Philosophy of Science*, *86*(1), 98-123.

Harrison, A. A. (1997). After contact: The human response to extraterrestrial life. *After Contact: The Human Response to Extraterrestrial Life*.

Haqq-Misra, J. (2012). An ecological compass for planetary engineering. *Astrobiology*, *12*(10), 985-997.

Harnagel, A. (2019). A mid-level approach to modeling scientific communities. *Studies in History and Philosophy of Science Part A*, *76*, 49-59.

Harvey, R., McSween, H. (1996). A possible high-temperature origin for the carbonates in the martian meteorite ALH84001. *Nature* 382, 49–51.

Hawking, S., (1988). *A Brief History of Time,* Bantam Books, ISBN 0-553-05340-X, pp. 7, 125.

He, M., Moldowan, M. J., & Peters, K. E. (2018). Biomarkers: petroleum. *Encyclopedia of Geochemistry. Encyclopedia of Earth Sciences Series. Springer, Cham*, *10*, 978-3.

Hecht, M. H., Kounaves, S. P., Quinn, R. C., West, S. J., Young, S. M., Ming, D. W., ... & Smith, P. H. (2009). Detection of perchlorate and the soluble chemistry of martian soil at the Phoenix lander site. *Science*, *325*(5936), 64-67.

Hedden, B., 2017. "Should juries deliberate?", Social Epistemology. 31(1): 1-19.

Heesen, R. (2019). The credit incentive to be a maverick. *Studies in History and Philosophy of Science Part A*, *76*, 5-12.

Hein, A. M., Perakis, N., Eubanks, T. M., Hibberd, A., Crowl, A., Hayward, K., ... & Osborne, R. (2019). Project Lyra: Sending a spacecraft to 1I/'Oumuamua (former A/2017 U1), the interstellar asteroid. *Acta Astronautica*, *161*, 552-561.

Heisenberg, W. (1927). Heisenberg Uncertainty Principle.

Heuer, V. B., Inagaki, F., Morono, Y., Kubo, Y., Spivack, A. J., Viehweger, B., ... & Hinrichs, K. U., (2020). Temperature limits to deep subseafloor life in the Nankai Trough subduction zone. *Science*, *370*(6521), 1230-1234.

Hofmann, H. J. (1972). Precambrian remains in Canada: Fossils, dubiofossils, and pseudofossils. In *Proceedings of the 24th International Geological Congress, Section* (Vol. **1**, pp. 20-30).

Hong, L., and Page, S., (2004). "Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers," *Proceedings of the National Academy of Sciences* 101 (46): 16385– 89.

Hood, L., & Rowen, L. (2013). *The human genome project: big science transforms biology and medicine. Genome Medicine, 5 (9), 79*.

Horowitz, A. (2017). Smelling themselves: Dogs investigate their own odours longer when modified in an "olfactory mirror" test. *Behavioural processes*, *143*, 17-24.

Howard, S. R., Avarguès-Weber, A., Garcia, J. E., Greentree, A. D., & Dyer, A. G. (2019). Numerical cognition in honeybees enables addition and subtraction. *Science advances*, *5*(2), eaav0961.

Hoyle, F. (1955). *Frontiers of Astronomy*. Heineman, London.

Hubbard, G. S. (2022). "Astrobiology: Its Origins and Development". *NASA*.

Huff, D. (2010). *How to lie with statistics*. WW Norton & company.

Hume, D. (1757). "Of the Standard of Taste," *Essays Moral and Political*, London: George Routledge and Sons, 1894.

Hunt, T. S. (1863). Report on the Geology of Canada. *Canadian Geological Survey Report: Progress to 1863*. Canadian Geological Survey; Calgary.

Jammer, M. (1974). Philosophy of Quantum Mechanics. the interpretations of quantum mechanics in historical perspective.

Jeancolas, C., Gillen, C., McMahon, S., Ward, M., & Vickers, P. J. (2024a). Breakthrough results in astrobiology: is 'high risk' research needed?. *International Journal of Astrobiology*, *23*, e1.

Jeancolas, C., Gillen, C., McMahon, S., & Vickers, P. (2024b). Is astrobiology serious science?. *Nature Astronomy*, *8*(1), 5-7.

Jellema, H. (2024). Reasonable doubt from unconceived alternatives. *Erkenntnis*, *89*(3), 971-996.

Jerison, H., (1955). Brain to body ratios and the evolution of intelligence, *Science* 121, 447–449.

Jerison, H., (1973). *Evolution of the Brain and Intelligence*, Academic, New York.

Jha, A., (2010). "Is Stephen Hawking right about aliens?". *The Guardian.* [Online] Available at: https://www.theguardian.com/science/2010/apr/30/stephen-hawking-right-aliens [Accessed on 10/02/2024].

Joyce G. F. (1994) Forward. In Deamer D.W. Fleischaker G. *Origins of Life: The Central Concepts*

Kahneman, D., and Tversky, A., (1979). "Prospect Theory: An Analysis of Decision under Risk", Econometrica, Vol. 47, No. 2, pp. 263-292.

Kahneman, D. (2011). Thinking, fast and slow. New York: Farrar, Straus and Giroux.

Kaitala, A., Kaitala, V., & Lundberg, P. (1993). A theory of partial migration. *The American Naturalist*, *142*(1), 59-81.

Kashyap, A. (2023). General Relativity, MOND, and the problem of unconceived alternatives. *European Journal for Philosophy of Science*, *13*(3), 30.

Kennicutt Jr, R. C., & Evans, N. J., (2012). Star formation in the Milky Way and nearby galaxies. *Annual Review of Astronomy and Astrophysics*, *50*, 531-608.

Kepes, S., Banks, G. C., & Oh, I. S. (2014). Avoiding bias in publication bias research: The value of "null" findings. *Journal of Business and Psychology*, *29*, 183-203.

Kilston, S. D., Drummond, R. R., & Sagan, C., (1966). A search for life on Earth at kilometer resolution. *Icarus*, *5*(1–6), 79–98.

Kitcher, P. (1990). The division of cognitive labor. *The journal of philosophy*, *87*(1), 5-22.

Kitcher, P. (1993). *The Advancement of Science*, New York: Oxford University Press.

Klein, H. P. (1978). The Viking biological experiments on Mars. *Icarus*, *34*(3), 666-674.

Kocevski, D. D., Onoue, M., Inayoshi, K., Trump, J. R., Haro, P. A., Grazian, A., ... & Yung, L. Y., (2023). Hidden Little Monsters: Spectroscopic Identification of Low-Mass, Broad-Line AGN at $ z> 5$ with CEERS. *arXiv preprint arXiv:2302.00012*.

Koch, K. R., & Koch, K. R. (1990). Bayes' theorem. *Bayesian Inference with Geodetic Applications*, 4-8.

Kochenderfer, M. J. (2015). *Decision making under uncertainty: theory and application*. MIT press.

Kohda, M., Hotta, T., Takeyama, T., Awata, S., Tanaka, H., Asai, J. Y., & Jordan, A. L. (2019). If a fish can pass the mark test, what are the implications for consciousness and self-awareness testing in animals?. *PLoS biology*, *17*(2), e3000021.

Kolmogorov, A. N., & Bharucha-Reid, A. T. (2018). *Foundations of the theory of probability: Second English Edition*. Courier Dover Publications.

Krissansen-Totton, J., Bergsman, D. S., & Catling, D. C. (2016). On detecting biospheres from chemical thermodynamic disequilibrium in planetary atmospheres. *Astrobiology*, *16*(1), 39-67.

Kudryavtsev, N. A. (1951). Against the organic hypothesis of the origin of petroleum, Petroleum Econ. (Neft. Khoz.) 9, 17–29.

Kuehn, D. (2017). Diversity, Ability, and Democracy: A Note on Thompson's Challenge to Hong and Page. *Critical Review*, *29*(1), 72–87. https://doi.org/10.1080/08913811.2017.1288455

Kuhn, T. S. (1977). Objectivity, value judgment, and theory choice. *Thomas Kuhn, The Essential Tension: Selected Studies in Scientific Tradition and Change*, 320-339.

Kuhn, T. S., (1997). *The structure of scientific revolutions* (Vol. 962). Chicago: University of Chicago press.

Kukla, Q. R. (2022). What counts as a disease, and why does it matter?. *The Journal of Philosophy of Disability*.

Kummer, T. A., Whipple, C. J., & Jensen, J. L. (2016). Prevalence and persistence of misconceptions in tree thinking. *Journal of microbiology & biology education*, *17*(3), 389-398.

Landemore, H. (2012). Democratic reason: Politics, collective intelligence, and the rule of the many.

Langston, S. M. (2016). Space travel: risk, ethics, and governance in commercial human spaceflight. *New Space*, *4*(2), 83-97.

Lederberg, J., (1960). Exobiology: Approaches to Life beyond the Earth. *Science*, *132*(3424), 393–400. https://doi.org/10.1126/science.132.3424.393

Legal Information Institute (n.d.). "Guilty", Cornell Law School. Available at: https://www.law.cornell.edu/wex/guilty [Accessed on 20 March 2024].

Lehman, C., (2017) How Many Bacteria Live on Earth? *Sciencing*. [Online] Available at: *https://sciencing.com/how-many-bacteria-live-earth-4674401.html* [Accessed on 02/12/2023].

Lepland, A., van Zuilen, M. A., Arrhenius, G., Whitehouse, M. J., & Fedo, C. M. (2005). Questioning the evidence for Earth's earliest life—Akilia revisited. *Geology*, *33*(1), 77-79.Lesquereux, L. (1866). Report on the fossil plants of Illinois, Ill. Geol. Surv. 2, 425–470.

Lewis, C. I. (1946). An Analysis of Knowledge and Valuation (La Salle, Illinois. *EG," what would happen" if everyone acted on a certain maxim*, 263.

Licquia, T. C., & Newman, J. A., (2015). Improved estimates of the Milky Way's stellar mass and star formation rate from hierarchical Bayesian meta-analysis. *The Astrophysical Journal*, *806*(1), 96.

Lin, D. N., Bodenheimer, P., & Richardson, D. C. (1996). Orbital migration of the planetary companion of 51 Pegasi to its present location. *Nature*, *380*(6575), 606-607.

Lin, H. (2022). "Bayesian Epistemology", *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2022/entries/epistemology-bayesian/>. (DOA: 24/01/2023).

Lincowski, A. P., Meadows, V. S., Crisp, D., Akins, A. B., Schwieterman, E. W., Arney, G. N., ... & Domagal-Goldman, S. (2021). Claimed detection of PH3 in the clouds of Venus is consistent with mesospheric SO2. *The Astrophysical Journal Letters*, *908*(2), L44.

Lineweaver, C. H. (2006). 6 Cosmological and Biological Reproducibility: Limits on the Maximum Entropy Production Principle. In *Non-equilibrium Thermodynamics and the Production of Entropy: Life, Earth, and Beyond* (pp. 67-77). Berlin, Heidelberg: Springer Berlin Heidelberg.

Lineweaver, C. H., (2007). Paleontological tests: Human-like intelligence is not a convergent feature of evolution. *arXiv preprint arXiv:0711.1751*.

Lineweaver, C. H. (2022a). A Lonely Universe. Inference: International Review of Science. 6. 10.37282/991819.22.11.

Lineweaver, C. H. (2022b). The 'Oumuamua Controversy: Bayesian priors and the evolution of technological intelligence. *Astrobiology*, *22*(12), 1419-1428.

Lineweaver, C. H., & 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., & Grether, D., (2003). What fraction of sun-like stars have planets? *The Astrophysical Journal*, *598*(2), 1350.

Lingam, M., & Loeb, A. (2017). The Astrophysical Journal Letters, 837, L23 2018, The Astronomical Journal, 156-193.

List, C., & Goodin, R. E. (2001). Epistemic democracy: Generalizing the Condorcet jury theorem.

List, C., & Spiekermann, K. (2016). The Condorcet Jury Theorem and Voter-Specific Truth. *Goldman and his critics*, 219-233.

Loeb, A. (2018). Six strange facts about our first interstellar guest, 'Oumuamua'. *arXiv preprint arXiv:1811.08832*.

Loeb, A. (2018). Six strange facts about our first interstellar guest, 'Oumuamua'. *arXiv preprint arXiv:1811.08832*.

Loeb, A. (2021). *Extraterrestrial: The First Sign of Intelligent Life Beyond Earth*. John Murray.

Loeb, A. (2022). On the Possibility of an Artificial Origin for 'Oumuamua. Astrobiology. Dec;22(12):1392-1399. doi: 10.1089/ast.2021.0193. PMID: 36475965.

Lorenz, R. D., Turtle, E. P., Barnes, J. W., Trainer, M. G., Adams, D. S., Hibbard, K. E., ... & Bedini, P. D. (2018). Dragonfly: A rotorcraft lander concept for scientific exploration at Titan. *Johns Hopkins APL Technical Digest*, *34*(3), 14.

Lovell, R. (1977). Letters to L-5. *L-5 News*, *2*(1).

Lowell, P. (1895). *Mars*. Houghton, Mittlin.

Lowell, P. (1906). *Mars and its Canals*. New York: The Macmillan Company; London: Macmillan & Company, Limited.

Lowell, P. (1980). Mars as the Abode of Life (1908). *The Quest for Extraterrestrial Life*, 76.

Lupo, A., Kininmonth, W., Armstrong, J. S., & Green, K. (2013). Global climate models and their limitations. *Climate change reconsidered II: Physical science*, *9*, 148.

Maák, I., Lőrinczi, G., Le Quinquis, P., Módra, G., Bovet, D., Call, J., & d'Ettorre, P. (2017). Tool selection during foraging in two species of funnel ants. *Animal Behaviour*, *123*, 207-216.

Machery, E. (2012). Why I stopped worrying about the definition of life... and why you should as well. *Synthese*, *185*, 145-164.

MacKenzie, S. M., Kirby, K. W., Greenauer, P. J., Neveu, M., Gold, R., Davila, A., ... & Crowley, D. (2020). Encealdus Orbilander: A flagship mission concept for astrobiology.

MacKenzie, S. M., Neveu, M., Davila, A. F., Lunine, J. I., Craft, K. L., Cable, M. L., ... & Spilker, L. J., (2021). The Enceladus Orbilander mission concept: Balancing return and resources in the search for life. *The Planetary Science Journal*, *2*(2), 77.

Macklem P. T., Seely A. (2010). Towards a definition of life. Perspect Biol Med. Summer;53(3):330-40. doi: 10.1353/pbm.0.0167. PMID: 20639603.

Malaterre, C. (2024). Is Life Binary or Gradual?. *Life*, *14*(5), 564.

Malaterre, C., Jeancolas, C., & Nghe, P. (2022). The origin of life: what is the question?. *Astrobiology*, *22*(7), 851-862.

Malaterre, C., & Lareau, F. (2024). Visualizing hidden communities of interest: A case-study analysis of topic-based social networks in astrobiology. *Scientometrics*, 1-15.

Malaterre, C., ten Kate, I.L., Baque, M., Debaille, V., Grenfell, J.L., Javaux, E.J., Khawaja, N., Klenner, F., Lara, Y., McMahon, S., Moore, K., Noack, L., Lucas Patty, C.H., Postberg, F. (2023) Is there such a thing as a biosignature? *Astrobiology.*

Manin, Y. (1980). Computable and uncomputable. *Sovetskoye Radio, Moscow*, *128*, 28.

Mann, S. (2012). Systems of creation: the emergence of life from nonliving matter. *Accounts of chemical research*, *45*(12), 2131-2141.

Marino, L., (2015). Fraction of life-bearing planets on which intelligent life emerges, f1, 1961 to the present. In The Drake Equation, edited by D.A. Vakoch and M.F. Dowd, Cambridge University Press, Cambridge, UK, pp 181–204.

Mashchenko, S. (2019). Modelling the light curve of 'Oumuamua: evidence for torque and disc-like shape. *Monthly Notices of the Royal Astronomical Society*, *489*(3), 3003-3021.

Mastrandrea, M.D., Field, C.B., Stocker, T.F., Edenhofer, O., Ebi, K.L., Frame, D.J., Held, H., Kriegler, E., Mach, K.J., Matschoss, P.R., Plattner, G.K., Yohe, G.W., and Zwiers, F.W., (2010). Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. Intergovernmental Panel on Climate Change (IPCC). Available at

Matarese, V. (2022). 'Oumuamua and meta-empirical confirmation. *Foundations of physics*, *52*(4), 83.

Maxwell, J. C. (1861). LI. On physical lines of force. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *21*(141), 338-348.

Mayor, M., Marmier, M., Lovis, C., Udry, S., Ségransan, D., Pepe, F., ... & Santos, N. C. (2011). The HARPS search for southern extra-solar planets XXXIV. Occurrence, mass distribution and orbital properties of super-Earths and Neptune-mass planets. *arXiv preprint arXiv:1109.2497*.

Mayor, M., Queloz, D., (1995). "A Jupiter-mass companion to a solar-type star". *Nature*. **378** (6555): 355–359.

Mayr, E. (1982). The growth of biological thought: Diversity, evolution and inheritance. *Harvard Univ Pr*.

Mayr, E. (1995). Space topics: Search for extraterrestrial intelligence. Available at: *https://web. archive. org/web/20081115225902/http://www. planetary. org/explore/topics/search_for_life/seti/mayr. html*.

McBryde, E. S., Meehan, M. T., Adegboye, O. A., Adekunle, A. I., Caldwell, J. M., Pak, A., ... & Trauer, J. M. (2020). Role of modelling in COVID-19 policy development. *Paediatric respiratory reviews*, *35*, 57-60.

McGrew, W. C. (2013). Is primate tool use special? Chimpanzee and New Caledonian crow compared. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1630), 20120422.

McKay, C. P. (2020). What is life—and when do we search for it on other worlds. *Astrobiology*, *20*(2), 163-166.

McKay, C. P., Anbar, A. D., Porco, C., & Tsou, P. (2014). Follow the plume: the habitability of Enceladus. *Astrobiology*, *14*(4), 352-355.

McKay, D.S., Gibson Jr, E.K., Thomas-Keprta, K.L., Vali, H., Romanek, C.S., Clemett, S.J., Chillier, X.D., Maechling, C.R. and Zare, R.N. (1996). Search for past life on Mars: possible relic biogenic activity in Martian meteorite ALH84001. *Science*, *273*(5277), pp.924-930.

McMahon, S., & Cosmidis, J. (2022). False biosignatures on Mars: anticipating ambiguity. *Journal of the Geological Society*.

McMahon, S., Ivarsson, M., Wacey, D., Saunders, M., Belivanova, V., Muirhead, D., ... & Frost, D. A. (2021). Dubiofossils from a Mars-analogue subsurface palaeoenvironment: The limits of biogenicity criteria. *Geobiology*, *19*(5), 473-488.

McNamara, J. M., Green, R. F., & Olsson, O. (2006). Bayes' theorem and its applications in animal behaviour. *Oikos*, *112*(2), 243-251.

Meadows, Graham, H., Abrahamsson, V., Adam, Z., Amador-French, E., Arney, G., Barge, L., Barlow, E., Berea, A., Bose, M., Bower, D., Chan, M., Cleaves, J., Corpolongo, A., Currie, M., Domagal Goldman, S., Dong, C., Eigenbrode, J., Enright, A., … Young, L. (2022). Community Report from the Biosignatures Standards of Evidence Workshop. http://arxiv.org/abs/2210.14293

Miles, B. E., Biller, B. A., Patapis, P., Worthen, K., Rickman, E., Hoch, K. K., ... & Zhang, Z. (2023). The JWST Early-release Science Program for Direct Observations of Exoplanetary Systems II: A 1 to 20 μm Spectrum of the Planetary-mass Companion VHS 1256–1257 b. *The Astrophysical journal letters*, *946*(1), L6.

Miller, K. B. (2012). Countering common misconceptions of evolution in the paleontology classroom. *The Paleontological Society Special Publications*, *12*, 109-122.

Minnesota House of Representatives (2009). "Permanent disqualification". *Bill Summary: House Research Department*. Minnesota. Available at: https://web.archive.org/web/20091204055107/http:/www.house.leg.state.mn.us/hrd/bs/86/hf1750.html [Accessed on 20 March 2024].

Mix L. J. (2015). Defending definitions of life. *Astrobiology*. Jan(1), pp.15-19.

Mosby Jr, G., Rauscher, B. J., Bennett, C., Cheng, E. S., Cheung, S., Cillis, A., ... & Wen, Y. (2020). Properties and characteristics of the nancy grace roman space telescope h4rg-10 detectors. *Journal of Astronomical Telescopes, Instruments, and Systems*, *6*(4), 046001-046001.

Muldoon, R. (2013). Diversity and the division of cognitive labor. *Philosophy Compass*, *8*(2), 117-125.

Muller, F. A. (2004). Can a constructive empiricist adopt the concept of observability?. *Philosophy of Science*, *71*(1), 80-97.

Muller, F. A. (2008). In defence of constructive empiricism: Maxwell's master argument and aberrant theories. *Journal for General Philosophy of Science*, *39*(1), 131-156.

Musgrave, A. (1985). Realism versus constructive empiricism. *Images of science*, 197-221.

NASA astrobiology, (n.d.). Retrieved from https://astrobiology.nasa.gov/about/faq/

NASEM, (2017). *Review of the Restructured Research and Analysis Programs of NASA's Planetary Science Division*. National Academies Press.

NASEM (2018). *Exoplanet Science Strategy* (September 2018, National Academies of Science, Engineering, and Medicine). Washington DC: The National Academies Press; https://sites.nationalacademies.org/SSB/CompletedProjects/SSB_180659.

NASEM, (2019). *An Astrobiology Strategy for the Search for Life in the Universe*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25252.

Navarro-González, R., Navarro, K. F., Rosa, J. D. L., Iñiguez, E., Molina, P., Miranda, L. D., ... & McKay, C. P. (2006). The limitations on organic detection in Mars-like soils by thermal volatilization–gas chromatography–MS and their implications for the Viking results. *Proceedings of the National Academy of Sciences*, *103*(44), 16089-16094.

Neumann, J., Morgenstern, O., (1953). *Theory of Games and Economic Behavior*. Princeton, NJ. Princeton University Press.

Neveu, M., Hays, L. E., Voytek, M. A., New, M. H., & Schulte, M. D. (2018). "The ladder of life detection". *Astrobiology*, *18*(11), 1375-1402.

Newberry J. S. (1873). The general geological relations and structure of Ohio, Ohio Geological Survey Report 1, Pt. 1 p. 222

Niesen, P., Spiekermann, K., Herzog, L., Girard, C., & Vogelmann, F. (2024). Does Diversity Trump Ability?. *Politische Vierteljahresschrift*, *65*(4), 785-805.

Nitzan, S., (2009). Collective Preference and Choice, Cambridge University Press.

Nutman, A. P., Bennett, V. C., Friend, C. R., Van Kranendonk, M. J., & Chivas, A. R., (2016). Rapid emergence of life shown by discovery of 3,700-million-year-old microbial structures. *Nature, 537*(7621), 535-538.

Office for Budget Responsibility, (2021). *Uncertainties around key epidemiological assumptions* (2021) Available at: https://obr.uk/box/uncertainties-around-key-epidemiological-assumptions/#:~:text=Vaccine%20rollout.&text=The%20Government%20plans%20to%20have,53%20million)%20by%2031%20July. (Accessed: 21 September 2024).

Page, S., (2007). *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies.* Princeton, NJ: Princeton University Press.

Pappalardo, R. T., Buratti, B. J., Korth, H., Senske, D. A., Blaney, D. L., Blankenship, D. D., ... & Niebur, C. (2024). Science Overview of the Europa Clipper Mission. *Space Science Reviews*, *220*(4), 1-58.

Penny, A. J., (2013). The SETI episode in the 1967 discovery of pulsars. *The European Physical Journal H*, *38*(4), 535-547.

Persson, E. (2012). The moral status of extraterrestrial life. *Astrobiology*, *12*(10), 976-984.

Persson, E. (2017). Ethics and the potential conflicts between astrobiology, planetary protection, and commercial use of space. *Challenges*, *8*(1), 12.

Persson, E., Lehmann Imfeld, Z., & Losch, A. (2018). A philosophical outlook on potential conflicts between planetary protection, astrobiology and commercial use of space. *Our common cosmos: Exploring the future of theology, human culture and space sciences*, 141-160.

Petigura, E. A., Howard, A. W., & Marcy, G. W., (2013). Prevalence of Earth-size planets orbiting Sun-like stars. *Proceedings of the National Academy of Sciences*, *110*(48), 19273-19278.

Pettigrew, R. G. (2016). Accuracy, Risk, and the Principle of Indifference. Philosophy and Phenomenological Research 92 (1):35-59.

Planck, M. (1900). The theory of heat radiation. *Entropie*, *144*(190), 164.

Plantinga, A. (1993). *Warrant: The current debate*. Oxford University Press, USA.

Plaut, S., Barabash, S., Bruzzone, L., Dougherty, M. K., Erd, C., Fletcher, L., ... & Wahlund, J. E., (2014). Jupiter icy moons explorer (JUICE): science objectives, mission and instruments.

Poincaré, H., [1905] (1952). *Science and Hypothesis*, New York: Dover.

Ponnamperuma, C., & Klein, H. P. (1970). The coming search for life on Mars. *The Quarterly Review of Biology*, *45*(3), 235-258.

Pop, V. (2008). *Who owns the moon?: Extraterrestrial aspects of land and mineral resources ownership* (Vol. 4). Springer Science & Business Media.

Popper, K. (2014). *Conjectures and refutations: The growth of scientific knowledge*. routledge.

Postberg, F., Sekine, Y., Klenner, F. *et al.*, (2023). Detection of phosphates originating from Enceladus's ocean. *Nature* **618**, 489–493. https://doi.org/10.1038/s41586-023-05987-9

Pöyhönen, S. (2016). "Value of Cognitive Diversity in Science", Synthese, 194, pp. 4519– 40.

Press, S. J. (2009). *Subjective and objective Bayesian statistics: principles, models, and applications*. John Wiley & Sons.

Psillos, S., (1996). "Scientific Realism and the 'Pessimistic Induction'". Philosophy of Science, 63, pp.S306-S314.

Psillos, S., (1999). *Scientific Realism: How Science Tracks Truth*, London: Routledge.

Public Health England, (2021). JCVI Advises on COVID-19 vaccine for people under 40. https://www.gov.uk/government/news/jcvi-advises-on-covid-19-vaccine-for-people-aged-under-40

Putnam, H., (1975). "Mathematics, Matter and Method". Philosophical Papers, Vol.1. Cambridge: Cambridge University Press.

Rampelotto, P. H., (2010). Panspermia: A promising field of research. In *Astrobiology Science Conference 2010: Evolution and Life: Surviving Catastrophes and Extremes on Earth and Beyond* Vol. 1538, p. 5224.

Rigotti, L., & Shannon, C. (2005). Uncertainty and risk in financial markets. *Econometrica*, *73*(1), 203-243.

Robitaille, T. P., & Whitney, B. A., (2010). The present-day star formation rate of the milky way determined from spitzer-detected young stellar objects. *The Astrophysical Journal Letters*, *710*(1), L11.

Rosenstock, S., Bruner, J., & O'Connor, C. (2017). In epistemic networks, is less really more?. *Philosophy of Science*, *84*(2), 234-252.

Rospars, J.P., (2013). Trends in the evolution of life, brains and intelligence. Int J Astrobiol 12:186–207.

Rothschild, L. J., & Mancinelli, R. L. (2001). Life in extreme environments. *Nature*, *409*(6823), 1092-1101.

Rowbottom, D. P., (2011). "The Instrumentalist's New Clothes", *Philosophy of Science*, 78(5): 1200–1211. doi:10.1086/662267

Ruhmkorff, S. (2011). Some difficulties for the problem of unconceived alternatives. *Philosophy of Science*, *78*(5), 875-886.

Russell D.A., (1983). Exponential evolution: implications for intelligent extraterrestrial life. Adv Space Res. 3(9):95-103.

Sagan, C. (1970). The definition of life, in *Encyclopaedia Britannica*, 14th edition.

Sagan, C., (1995). Carl Sagan responds [Online] Available at: http://www.planetary.org/html/UPDATES/seti/Contact/debate/Sagan2.htm [Accessed on 10/02/2024].

Sagan, C. (2010). Definitions of life. *The Nature of Life: Classical and Contemporary Perspectives from Philosophy and Science*, 303-306.

Sagan, C., Thompson, W. R., Carlson, R., Gurnett, D., & Hord, C., (1993). A search for life on Earth from the Galileo spacecraft. *Nature*, *365*(6448), 715–721.

Sartwell, C. (1991). Knowledge is merely true belief. *American philosophical quarterly*, *28*(2), 157-165.

Saslow, W. M. (2020). A history of thermodynamics: the missing manual. *Entropy*, *22*(1), 77.

Savage, L. J. (1972). *The foundations of statistics*. Courier Corporation.

Schindler, S. (2018). *Theoretical virtues in science: Uncovering reality through theory*. Cambridge University Press.

Schingler R., Marshall W., MacDonald A., Lupisella M., Lewis B., (2009). ROSI – Return on Science Investment: A system for mission evaluation based on maximizing science, a white paper in support of the *Planetary Science Decadal Survey 2013-2022*, a report of the National Research Council. Washington DC: National Academies Press.

Schneider, E. D., & Kay, J. J. (1994). Life as a manifestation of the second law of thermodynamics. *Mathematical and computer modelling*, *19*(6-8), 25-48.

Schrödinger, E. (1926). SCHRÖDINGER 1926C. *Annalen der Physik*, *79*, 734.

Schrödinger, E. (1944). What is life? The physical aspect of the living cell.

Schulze-Makuch, D., & Bains, W., (2017). *The Cosmic Zoo: Complex Life on Many Worlds*. Springer.

Schwartz, J. S. (2013). On the moral permissibility of terraforming. *Ethics & the Environment*, *18*(2), 1-31.

Schwartz, J. S. (2019). Where no planetary protection policy has gone before. *International Journal of Astrobiology*, *18*(4), 353-361.

Šekrst, K. (2024). Astrobiology in philosophy or philosophy in astrobiology? *Cosmos and History* 20 (1):405-415.

SETI Institute, (n.d.). Financials. [Online] Available at: https://www.seti.org/about-us/financials [Accessed on 10/12/2023].

Shapley, H., (1959). "Life On Other Planets?". *Sydney Morning Herald*

Shermer, M., (2002). "Why ET Hasn't Called". *Scientific American*. **287** (2): 21. doi:10.1038/scientificamerican0802-33.

Shillito, D. J., Gallup Jr, G. G., & Beck, B. B. (1999). Factors affecting mirror behaviour in western lowland gorillas, Gorilla gorilla. *Animal Behaviour*, *57*(5), 999-1004.

Shostak S., (2015). Fraction of civilizations that develop a technology that releases detectable signs of their existence into space, fc, 1961 to the present. In: Vakoch DA, Dowd MF, eds. *The Drake Equation: Estimating the Prevalence of Extraterrestrial Life through the Ages*. Cambridge Astrobiology. Cambridge: Cambridge University Press; 2015:227-240.

Simoncini, E., Virgo, N., & Kleidon, A. (2013). Quantifying drivers of chemical disequilibrium: Theory and application to methane in the Earth's atmosphere. Earth System Dynamics, 4(2), 317 331. https://doi.org/10.5194/esd-4-317-2013.

Siraj, A., & Loeb, A., (2020). Possible transfer of life by Earth-grazing objects to exoplanetary systems. *Life*, *10*(4), 44.

Smith, K. C. (2014). Manifest complexity: A foundational ethic for astrobiology?. *Space Policy*, *30*(4), 209-214.

Snellen, I. A. G., Guzman-Ramirez, L., Hogerheijde, M. R., Hygate, A. P. S., & Van der Tak, F. F. S. (2020). Re-analysis of the 267 GHz ALMA observations of Venus-No statistically significant detection of phosphine. *Astronomy & Astrophysics*, *644*, L2.

Sober, E. (1985). Constructive empiricism and the problem of aboutness. *The British journal for the philosophy of science*, *36*(1), 11-18.

Sommerfeld, A. (1916). Zur quantentheorie der spektrallinien. *Annalen der Physik*, *356*(17), 1-94.

Sparrow, R. (1999). The ethics of terraforming. *Environmental Ethics*, *21*, 227-246.

Spergel, D., Berea, A., Bianco, F., Brothers, R., Bontempi, P., Buss, J., Drake, N., Gold, M., Grinspoon, D., Evans, D., Kelly, S., Mountain, M., Randolph, W., Scott, W., Semeter, J., Toner, K., Write, S. (2023). *Unidentified Anomalous Phenomena Independent Study Team Report.* Available at: https://science.nasa.gov/wp-content/uploads/2023/09/uap-independent-study-team-final-report.pdf [Accessed on 05//01/2025].

Spiegel, D. S., & 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.

Stanford, P. K. (2006a). *Exceeding our grasp: Science, history, and the problem of unconceived alternatives* (Vol. 1). Oxford University Press.

Stanford, P. K. (2006b). Darwin's Pangenesis and the Problem of Unconceived Alternatives. *The British journal for the philosophy of science*.

Steel, D., & Bolduc, N. (2020). A closer look at the business case for diversity: The tangled web of equity and epistemic benefits. *Philosophy of the Social Sciences*, *50*(5), 418-443.

Stephens, J. E. R. (1896). Growth of trial by jury in England. *Harv. L. Rev.*, *10*, 150.

Stoner, I. (2021). The ethics of terraforming: A critical survey of six arguments. *Terraforming Mars*, 99-115.

Strevens, M. (2003). "The Role of the Priority Rule in Science", *The Journal of Philosophy*, 100, pp. 55–79.

Tarter, D. E. (1992). Interpreting and reporting on a SETI discovery: we should be prepared. *Space Policy*, *8*(2), 137-148.

Tennen, L., & Forgan, D., (2018). A Report on The IAA Permanent SETI Committee's Review of the SETI Post-Detection and Reply Protocols. *42nd COSPAR Scientific Assembly*, *42*, F3-8.

The 'Oumuamua ISSI Team (2019). "The natural history of 'Oumuamua." *Nature Astronomy* 3, no. 7: 594-602.

The Planetary Society, (n.d.). *NASA's FY 2022 Budget.* [Online} Available at: https://www.planetary.org/space-policy/nasas-fy-2022-budget [Accessed on 10/12/2023].

Thompson, A., (2014). "Does Diversity Trump Ability? An Example of the Misuse of Mathematics in the Social Sciences." *Notices of the American Mathematical Society* 61:1024– 30.

Thompson, M. A., Krissansen-Totton, J., Wogan, N., Telus, M., & Fortney, J. J. (2022). The case and context for atmospheric methane as an exoplanet biosignature. Proceedings of the National Academy of Sciences, 119(14), e2117933119. https://doi.org/10.1073/pnas.2117933119.

Thompson, D. R., & Martin, C. R. (2017). Bayes' theorem and its application to cardiovascular nursing. *European Journal of Cardiovascular Nursing*, *16*(8), 659-661.

Thomson, W. (1853). XV.—On the Dynamical Theory of Heat, with numerical results deduced from Mr Joule's Equivalent of a Thermal Unit, and M. Regnault's Observations on Steam. *Earth and Environmental Science Transactions of The Royal Society of Edinburgh*, *20*(2), 261-288.

Titelbaum, M. G. (2022). *Fundamentals of Bayesian epistemology 2: Arguments, challenges, alternatives*. Oxford University Press.

Totani T., (2020). Emergence of life in an inflationary universe. *Scientific Reports* 10, 1671.

Trilling, D. E., Mommert, M., Hora, J. L., Farnocchia, D., Chodas, P., Giorgini, J., ... & Micheli, M. (2018). Spitzer observations of interstellar object 1I/'Oumuamua. *The Astronomical Journal*, *156*(6), 261.

Trompet, L., Robert, S., Mahieux, A., Schmidt, F., Erwin, J., & Vandaele, A. C. (2021). Phosphine in Venus' atmosphere: Detection attempts and upper limits above the cloud top assessed from the SOIR/VEx spectra. *Astronomy & Astrophysics*, *645*, L4.

Tsiaras, A. (2021). K2-18b: first water vapour detection in a habitable-zone planet. *Bulletin of the American Astronomical Society*, *53*(3), 1145.

Tsiaras, A., Waldmann, I. P., Tinetti, G., Tennyson, J., & Yurchenko, S. N. (2019). Water vapour in the atmosphere of the habitable-zone eight-Earth-mass planet K2-18 b. *Nature Astronomy*, *3*(12), 1086-1091.

Vakoch, D. A. (2013). *Astrobiology, History, and Society*.

Vakoch, D. A., & Lee, Y. S. (2000). Reactions to receipt of a message from extraterrestrial intelligence: A cross-cultural empirical study. *Acta Astronautica*, *46*(10-12), 737-744.

van Fraassen, Bas C., (1980). *The Scientific Image*, Oxford: Oxford University Press.

Vanleynseele, E., (2022). *New record in Government Space Defense spendings driven by investments in Space Security and Early Warning.* [online] Euroconsult. Available at: https://www.euroconsult-ec.com/press-release/new-record-in-government-space-defense-spendings-driven-by-investments-in-space-security-and-early-warning/ [Accessed on 10/12/2023]

Vickers, P., (2020). Expecting the unexpected in the search for extraterrestrial life. *International Journal of Astrobiology*, *19*(6), 482-491.

Vickers, P. (2022). *Identifying future-proof science*. Oxford University Press.

Vickers, P., Cowie, C., Dick, S. J., Gillen, C., Jeancolas, C., Rothschild, L. J., & McMahon, S. (2023). Confidence of life detection: The problem of unconceived alternatives. *Astrobiology*, *23*(11), 1202-1212.

Vickers, P., Gardiner, E., Gillen, C., Hyde, B., Jeancolas, C., Mitchell Finnigan, S., ... & McMahon, S. (2025). Surveys of the scientific community on the existence of extraterrestrial life. *Nature Astronomy*, 1-3.

Villanueva, G. L., Cordiner, M., Irwin, P. G., de Pater, I., Butler, B., Gurwell, M., ... & Kopparapu, R. (2021). No evidence of phosphine in the atmosphere of Venus from independent analyses. *Nature Astronomy*, *5*(7), 631-635.

von Hegner, I. (2019). Astrobiology and astrophilosophy: subsuming or bifurcating disciplines?. *Philosophy and Cosmology*, *23*(23), 62-79.

Von Hoerner, S., (1961). The search for signals from other civilizations. *Science* 134:1839–1843.

von Neumann, J., & Morgenstern, O. (1947). Theory of games and economic behavior, 2nd rev.

Votsis, I., (2011). "The Prospective Stance in Realism". Philosophy of Science, 78(5), pp.1223 1234.

Wade, N., (1975). Discovery of pulsars: a graduate student's story. *Science*, *189*(4200), 358-364.

Walker, S. I., Bains, W., Cronin, L., DasSarma, S., Danielache, S., Domagal-Goldman, S., ... & Smith, H. B. (2018). "Exoplanet biosignatures: future directions". *Astrobiology*, *18*(6), 779-824.

Wang, J., Veugelers, R., & Stephan, P., (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, *46*(8), 1416-1436.

Weisberg, M., & Muldoon, R. (2009). Epistemic landscapes and the division of cognitive labor. *Philosophy of science*, *76*(2), 225-252.

Werth, A., (2022). The Problems of Evolution as a "March of Progress," https://www.sapiens.org/biology/evolution-march-of-progress/. [Accesses on 20/11/2024].

Westall, F., Loizeau, D., Foucher, F., Bost, N., Betrand, M., Vago, J., & Kminek, G. (2013). Habitability on Mars from a microbial point of view. *Astrobiology*, *13*(9), 887-897.

Wickramasinghe C., (2011). Bacterial morphologies supporting cometary panspermia: a reappraisal. *International Journal of Astrobiology*. 10(1):25-30.

Wilson, T. L. (2001). The search for extraterrestrial intelligence. *Nature*, *409*(6823), 1110-1114.

Witze, A., (2023). 'It's a dream': JWST spies more black holes than astronomers predicted. *Nature*.

Wogan, N. F., & Catling, D. C. (2020). When is chemical disequilibrium in Earth-like planetary atmospheres a biosignature versus an anti-biosignature? Disequilibria from dead to living worlds. *The Astrophysical Journal*, *892*(2), 127.

Wolszczan, A., & Frail, D. A. (1992). A planetary system around the millisecond pulsar PSR1257+ 12. *Nature*, *355*(6356), 145-147.

Wood, J. A., Dickey Jr, J. S., Marvin, U. B., & Powell, B. N. (1970). Lunar anorthosites. *Science*, *167*(3918), 602-604.

Wordsworth, R. D. (2016). The climate of early Mars. *Annual Review of Earth and Planetary Sciences*, *44*(1), 381-408.

Worrall, J., (1989). "Structural Realism: The Best of Both Worlds?". dialectica, 43(1-2), pp.99-124.

Worrall, J. (1994). How to remain (reasonably) optimistic: scientific realism and the "luminiferous ether". In *PSA: Proceedings of the biennial meeting of the Philosophy of Science Association* (Vol. 1994, No. 1, pp. 334-342). Cambridge University Press.

Wright, J. T., Marcy, G. W., Howard, A. W., Johnson, J. A., Morton, T. D., & Fischer, D. A., (2012). The frequency of hot Jupiters orbiting nearby solar-type stars. *The Astrophysical Journal*, *753*(2), 160.

Yan, H., Ma, Z., Ling, C., Cheng, C., & Huang, J. S. (2022). First batch of candidate galaxies at redshifts 11 to 20 revealed by the James Webb Space Telescope early release observations. *arXiv preprint arXiv:2207.11558*.

Young, A. V., Robinson, T. D., Krissansen-Totton, J., Schwieterman, E. W., Wogan, N. F., Way, M. J., ... & Windsor, J. D. (2024). Inferring chemical disequilibrium biosignatures for Proterozoic Earth-like exoplanets. *Nature Astronomy*, *8*(1), 101-110.

Zagzebski, L. (1994). The inescapability of Gettier problems. *The Philosophical Quarterly (1950-)*, *44*(174), 65-73.

Zollman, K. J. (2007). The communication structure of epistemic communities. *Philosophy of science*, *74*(5), 574-587.

Zuckerman, B. (2022). 'Oumuamua Is Not a Probe Sent to Our Solar System by an Alien Civilization. *Astrobiology*, *22*(12), 1414-1418.