Title

Narrative Explanation in the Evolution of Human Cognition

ISBN

9798291550564

Author

Farrell, Margaret Emily

Publication Date

2025-08-23

UNIVERSITY OF CALIFORNIA,
IRVINE

Narrative Explanation in the Evolution of Human Cognition

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Philosophy

by

Margaret Emily Farrell

Dissertation Committee:
Professor P. Kyle Stanford, Chair
Professor Lauren N. Ross
Professor Cailin O'Connor
Professor James Owen Weatherall

2025

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

Thank you first to my committee: to Kyle Stanford, for reminding me to pursue questions that I deeply and urgently want to know the truth about, and for believing in my ability to do so; to Lauren Ross for encouragement, support, and attention to detail in my work, both in this dissertation and other projects; to Jim Weatherall for helping me to participate in the profession through our work for *Philosophy of Science* and for reading drafts with alarming speed; and to Cailin O'Connor, for finding structure and direction even in some very rough drafts, and for encouraging me to be confident in sharing both my work and my personality with my colleagues.

Thank you also to the Department of Logic and Philosophy of Science, whose community of faculty, students, and visitors has created a home for me in Southern California. Thank you especially to my friends and colleagues who read, wrote, and commiserated with me throughout the years, among them: Jingyi Wu, Helen Meskhidze, Rebecca Korf, Adam Chin, Clara Bradley, Ben Genta, True Gibson, Ainsley May, Jaehyun Lee, Ellen Shi, David Rattray, Kino Zhao, Rena Goldstein, and Will Stafford.

I am so grateful to my family for their love and support: to Tucker and Luke, who are role models to me in their curiosity, determination, and sense of humor; to my mom, Maggie, for years of encouragement to read and learn; to my dad, Michael, for always calling to check in no matter the time zone, and to all of them for never once questioning my decision to move to California and pursue a PhD. Thank you to my found SoCal family; to True for more than I could possibly list; and to the creatures, Herschel, Lupine, and Lychee, for their antics and affection.

I am also grateful for the support of a Nevin Fellowship, Graduate Dean's Recruitment Fellowship, Associate Dean's Fellowship, and Beinecke Scholarship.

Chapter 4 of this dissertation is published under as "What would imaginary ancestors do? Thought experiments and intuitive plausibility in human cognitive evolution" in *Biology and Philosophy*, **40,** 14 (2025). It is licensed under a Creative Commons Attribution 4.0 International License. Minor changes were made for clarity in order to integrate this paper with the rest of the dissertation.

VITA

Margaret Emily Farrell

EDUCATION

Doctor of Philosophy in Philosophy **2025**
University of California, Irvine *Irvine*

Master of Arts in Philosophy **2021**
University of California, Irvine *Irvine*

Bachelor of Philosophy in History and Philosophy of Science **2018**
University of Pittsburgh *Pittsburgh*

Bachelor of Science in Ecology and Evolution **2018**
University of Pittsburgh *Pittsburgh*

RESEARCH EXPERIENCE

Managing Editor *Philosophy of Science* **2023–2025**
University of California, Irvine *Irvine*

Graduate Student Researcher **2020-2023**
University of California, Irvine *Irvine*

TEACHING EXPERIENCE

Graduate Teaching Associate **2022**
University of California, Irvine *Irvine*

Teaching Assistant **2019–2022**
University of California, Irvine *Irvine*

REFEREED JOURNAL PUBLICATIONS

What would imaginary ancestors do? Thought experiments and intuitive plausibility in human cognitive evolution **2025**
Biology and Philosophy

Genome Sequences of Mycobacteriophages AlanGrant, Baee, Corofin, Or- angeOswald, and Vincenzo, New Members of Cluster B. **2015**
Genome Announcements

INVITED CONFERENCE PRESENTATIONS

Value Influence in Causal Selection in the Social Sciences **2025**
with Rebecca Korf. Social Beliefs in Science Workshop, Irvine

Narratives, Explanation, and Cognition **2025**
Computational Cognitive Neuroscience Lab, UC Irvine

So, Just Storytelling? Narrative Explanation in the Evolution of Human Cognition **2024**
Human Cognitive Evolution Workshop, Irvine

Narrative Explanation **2023**
Workshop on Topics in Causality and Explanation, Munich

REFEREED CONFERENCE PRESENTATIONS

Rights vs Knowledge: Resolving Paternalism in Iterative Decision-making Processes **2025**
with Rebecca Korf. Workshop in Science, Values, and Policy, UC Merced

Value influence in causal selection in the social sciences **2024**
with Rebecca Korf Revitalizing Science and Values, Pittsburgh

Adequacy for Purpose when the Purpose is Truth **2023**
ISHPSSB, Toronto

Causal Selection in Context: Explaining Gene Centrism **2022**
Philosophy of Science Association, Pittsburgh

What Would Imaginary Ancestors Do? **2021**
Canadian Society for History & Philosophy of Science, Remote

ABSTRACT OF THE DISSERTATION

Narrative Explanation in the Evolution of Human Cognition

By

Margaret Emily Farrell

Doctor of Philosophy in Philosophy

University of California, Irvine, 2025

Professor P. Kyle Stanford, Chair

Why do some explanations seem to take the form of a narrative? And why should evolutionary theorists in particular have recently proposed narrative explanations for the evolution of human cognition, when 'storytelling' is a commonplace pejorative for evolutionary explanation? In recent work on human cognitive evolution, several biologists and philosophers have put forth hypotheses in which they attempt to synthesize knowledge from a broad range of disciplines into comprehensive narratives. In this dissertation, I focus on one question raised by this broad and interdisciplinary approach: How does narrative structure affect the strategies and challenges for theorists of human cognitive evolution, especially compared to alternative approaches? I argue for three theses to demonstrate that the narrative structure of these hypotheses and the approach that theorists use to construct them are effects of the *instability* of the central causal relationship in question – the connection between the cognitive capacities of modern humans and those of the LCA. I first develop a novel account of narrative explanation: causal explanation takes the form of a narrative when the explanandum is a token event that stands in an unstable causal relationship with the first relevant cause in its causal history. I then use this account to make sense of two strategies that theorists use when building narratives of the evolution of human cognition in order to maximize detail under evidential scarcity: functional and historical associations. Finally, I consider some of the difficulties inherent to this approach, despite its potential advantages.

For example, narratives seem to invite speculation to fill missing causal steps. My dissertation shows that narrative hypotheses of human cognitive evolution are a fruitful case for understanding narrative explanation generally, and conversely, that once these hypotheses are characterized and evaluated *as narrative explanations*, we can better understand this recent, promising approach to human cognitive evolution.

In **Chapter 2**, **"Causal properties of narrative explanations tell us what they are and when we'll find them,"** I present my account of narrative explanation. Some scientists produce explanations that seem to have a story-like, or *narrative* form. Such explanations are especially prominent in the historical sciences, including evolutionary history. To achieve an understanding of explanatory practice in historical science as well as a fuller understanding of the diversity of explanatory practices across the sciences, philosophers of science have proposed accounts of narrative explanation that seek to identify their unique explanatory features. However, such accounts neglect or misidentify the features that distinguish mere narratives from narrative explanations, and narrative explanations from other causal explanations. In this chapter, I propose a novel account of narrative explanation that situates it within an interventionist account of causal explanation. Using my account, I propose three dimensions along which narrative explanations can be evaluated and draw several consequences for outstanding disagreements in the literature on narrative explanations. Ultimately, my account clarifies and justifies narrative explanation as a legitimate and distinct form of explanation.

In **Chapter 3, "Implications of instability for the construction of narrative explanations,"** I argue that instability that underlies the structure of narrative explanation also underlies one of its characteristic features – narratives seem to be most satisfying when they include a lot of detail. This connection explains two inferential strategies that theorists use to build narratives that help them to add detail under the constraints of limited evidence. These two strategies are (1) to show that one capacity emerged at a certain point in evolutionary history by connecting it functionally with another capacity whose emergence is evidenced in the material record, and (2) to show that past traits have high evolvability with respect to their enhancement toward traits in modern humans. Both such strategies inherently invoke what I call a "close-enough judgment," wherein a theorist judges that the functional difference between one capacity and another, or that the evolutionary steps between a capacity and an enhanced version of that capacity, are *small* enough or *close* enough in an evolutionarily relevant sense. Close-enough judgments are not formally problematic but given the challenge of sparse material evidence in the hominin lineage, they constitute a potential weak point in narrative hypotheses.

In **Chapter 4**, **"What would imaginary ancestors do?"** I consider the challenges and potential pitfalls of the narrative approach. The reconstruction of the evolutionary history of human cognition is a complex, interdisciplinary science. Human evolutionary theorists generally use a wide variety of evidential sources to construct hypotheses for the emergence of novel cognitive capacities in the hominin lineage. But sometimes, they take certain associations among cognitive capacities to be intuitively plausible – so much so that they accept and include them in their hypotheses. I argue that when they make these plausibility judgments, they are actually attempting (explicitly or implicitly) to reason from the perspective of an ancestral hominin. They reason about whether some inference would be easy, natural, or obvious for a hominin with only a hypothesized set of ancestral capacities. When they do this, they invoke their intuitive sense of ease or obviousness as evidence for the association among cognitive capacities. But such intuitions are systematically misleading, because they are influenced by the very cognitive capacity whose evolution the theorist seeks to explain. Intuitions that stem from the evolved capacity in question thus make some associations among cognitive capacities appear artificially plausible. I analyze examples of this kind of reasoning from three recent hypotheses for human cognitive evolution. I argue that this is a particularly pervasive challenge for evolutionary theorists. Even though they are typically highly attuned to similar errors, theorists should expect to keep contending with this challenge, even in our best approaches to human cognitive evolution.

Chapter 1

Introduction

Why do some explanations seem to take the form of a narrative? And why should evolutionary theorists increasingly favor narrative explanations for the evolution of human cognition, when 'storytelling' is a commonplace pejorative for evolutionary explanation? In recent work on human cognitive evolution, several biologists and philosophers have put forth broad, synthetic hypotheses in which they attempt to bring together knowledge from a wide range of disciplines into comprehensive narratives. These narrative hypotheses are historically focused and phylogenetically constrained, their temporal scope makes them well-suited to tracing gradual changes in selective pressures over time, and their sequential causal structure means they need not commit to one sort of cause over another and can easily incorporate adaptive hypotheses as well as environmental events or developmental constraints. However, the narrative approach may also invite speculation more frequently than other approaches. Because they aim for a detailed causal story, authors of narrative hypotheses must make and justify a large number of constitutive causal claims, including adaptive hypotheses about the origins of particular cognitive capacities. The relatively poor evidential situation leaves theorists with gaps in the narrative that cannot be straightforwardly filled, potentially inviting speculation.

Despite differences in their content and background assumptions, these hypotheses are united by their apparent narrative structure. In this dissertation, I focus on one question raised by this broad and interdisciplinary approach: How does narrative structure affect the strategies and challenges for theorists of human cognitive evolution, especially compared to alternative approaches? I argue for three theses to demonstrate that the narrative structure of these hypotheses and the approach that theorists use to construct them are driven by the *instability* of the central causal relationship in question – the connection between the cognitive capacities of modern humans and those of the LCA. I first develop a novel account of narrative explanation to show that narrative is a distinct and important subcategory of causal explanation: narrative is the form that the causal explanation of some target must take when that target is a token event that stands in an unstable relationship with the first relevant cause in its causal history. I then use this account to argue that two major strategies employed by theorists when building narratives of the evolution of human cognition – functional and historical association – are used to address stability under evidential scarcity.

Finally, I consider some of the difficulties inherent to this approach, despite its potential advantages. Many challenges to any approach to the evolution of human cognition have to do with a scarcity of evidence – empirical evidence of ancestral hominin behavior and cognition is relatively limited, and what is available is difficult to piece together. Most obviously, we cannot go out and observe the behavior of extinct hominin ancestors. Fossils of our extinct relatives are rare, the information those fossils provides degrades over time, and even our best fossil specimens preserve information about behavior and cognition only indirectly (Turner, 2005). By the standards of evolutionary reconstruction for other species, our closest living relatives are not very close, and our clade appears to be sparse overall – there are few species, extant or extinct, with which to compare our behavioral and cognitive traits (Kaplan, 2002).

The evolution of human cognition and behavior is an old, ever popular, and difficult explanatory target. The unique features of human cognition and behavior call for an explanation of their origins and evolution, but this has been notoriously difficult to achieve. The somewhat broader forms of this target, the human mind or human nature, have been subjects of philosophical thought more or less since antiquity. Early evolutionary theories for the origin of human psychology and behavior developed alongside general theories of biological evolution, e.g., in Darwin's *Descent of Man* (1879), but the roots of contemporary approaches emerge in the 20$^{\text{th}}$ century. Paul Griffiths separates four distinct phases: instinct theory, classical ethology, human sociobiology, and Evolutionary Psychology (Griffiths, 2011, 2006).[1]

The sociobiology of Edward O. Wilson's *Sociobiology: The New Synthesis* was extensively critiqued[2] and largely replaced by the Evolutionary Psychology of Jerome Barkow, Leda Cosmides, and John Tooby, which remains a prominent approach to the evolution of human cognition (Griffiths, 2011). Griffiths emphasizes that there are many points of continuity in the lineage he traces through ethology, sociobiology, and Evolutionary Psychology and that there are many, often overlapping contemporary approaches to human cognitive evolution.[3] Aside from the contemporary practice of human sociobiology and Evolutionary Psychology, these include human behavioral ecology, cultural evolution, gene-culture co-evolution, and what Griffiths calls a 'developmentalist tradition' (Griffiths 2011, 410). Additionally, evolutionary game theorists pursue related questions using the formal apparatus of games to model behavior, although this should not be construed as a wholly different approach compared to the others; work on altruistic behavior in (not just human) sociobiology is an early example, e.g., Smith and Price (1973), Trivers (1971), Williams and Williams (1957). Griffiths also points out that Evolutionary Psychologists incorporate the same population genetic

[1] Following convention from Buller (2005; 2000),I refer to the framework of Barkow, Tooby, and Cosmides as the capitalized "Evolutionary Psychology" to contrast it with the more general "evolutionary psychology," which refers to the subject area of the evolutionary origins of human psychology and not to a specific framework. Evolutionary Psychology has also been called the "Santa Barbara church of psychology" (Lala and Brown, 2011) and "narrow evolutionary psychology" (Scher and Rauscher, 2003).

[2] See also Kitcher (1989), Lewontin (1979), Gould (1978), Wilson (1975), Barkow et al. (1992), Crawford et al. (1987), Sterelny (1992).

[3] Downes (2016, 2001), Scher and Rauscher (2003), and Lala and Brown (2011) also survey and provide examples of these different approaches.

and evolutionary game theoretic models developed in the context of sociobiology into their own theory and methodology (Griffiths 2011, 402). This formal approach has been developed extensively toward understanding the evolution of social behaviors, especially through signaling games, see, e.g., O'Connor (2020) for review.

Cognitive paleoanthropology is another approach that comes from a somewhat different starting point; its roots in (paleo)anthropology, archaeology, and the philosophies thereof. Archaeologists, anthropologists, and philosophers have continuously confronted problems of evidence and inference in fields that rely on material evidence from the deep past (Wylie and Chapman, 2016; Wylie, 2011, 1985). Another cluster of contemporary philosophers of the historical sciences, anthropology, and cultural evolution have brought work like Wylie's into explicit contact with recent hypotheses for human cognitive evolution. They see cultural and gene-culture evolution as coalescing into the discipline of "cognitive paleoanthropology" (Currie et al., 2024).[4]

But several recent hypotheses from across these disciplines seem to be united in virtue of having a *narrative* structure. And these narrative hypotheses are the focus of this dissertation. Broadly, theorists proceed in these narratives by empirically establishing the 'endpoints' of the relevant portion of evolutionary history – the Last Common Ancestor (LCA) shared among humans, chimpanzees, and bonobos on the earlier end, and modern human cognition on the later end – and then reconstructing, in as much detail as possible, the events and conditions that served as selective pressures and the gradual changes in cognition along the hominin lineage. Theorists endeavor to synthesize a wide variety of evidence to do so. This results, ideally, in a comprehensive causal sequence capturing the evolutionary history of uniquely human cognition.

Exemplars of what I'll call the 'narrative approach' include Michael Tomasello's Shared

[4] See, e.g., Meneganzin and Killin (2024), Currie and Meneganzin (2024), Pain (2023), Killin and Pain (2023), Meneganzin and Currie (2022), Killin and Pain (2021), Pain (2021), Currie and Killin (2019), and Currie (2018, 2015).

Intentionality Hypothesis (2019; 2016; 2014), Sarah Blaffer Hrdy's alloparenting hypothesis (2009), Cecelia Heyes's cognitive gadgets hypothesis (2018), Kim Sterelny's apprentice model for early social evolution (2012), his hypothesis for later social evolution (2021), and, with Ronald Planer, their hypothesis for the evolution of language (2021). The development of these hypotheses and the narrative approach they represent is exciting, as their historical narrative structure provides a potentially promising strategy for incorporating the directives that have come out of debates on adaptationism and criticisms of Evolutionary Psychology.

Typical classifications place these hypotheses among the different approaches listed above. Heyes's *Cognitive Gadgets* is typically categorized with cultural evolution, *Mothers and Others,* where Hrdy presents her alloparenting hypothesis, is one of the earliest examples of the narrative approach and is usually classified in anthropology, primatology, and/or sociobiology. Sterelny and Planer's work is often classified as gene-culture co-evolution, joining, e.g., Boyd and Richerson's *Culture and The Evolutionary Process* (1988) and Richard Joyce's *The Evolution of Morality* (2006). Sterelny's work is also included under the umbrella of cognitive paleoanthropology, and Tomasello's work is typically categorized in social and developmental psychology, comparative psychology, or evolutionary psychology. These classifications reflect the content of these hypotheses and the disciplinary background of the theorist more so than the structure of the resulting hypothesis. For the purposes of understanding their individual contexts, emphases, and assumptions, these classifications are useful. My purpose is different here; I group them together for the purposes of studying their explanatory structure, as they share a narrative structure.

The origins of human cognition and behavior, and the related question about whether there is a real 'human nature' has deep, important implications on how we view ourselves and one another, making it both popular and extremely high stakes. Popular media is full of "nature vs nurture" questions and "scientific" bases for human nature. And Evolutionary Psychology in particular is extremely popular in trade books, e.g., Steven Pinker's *How the*

Mind Works was a bestseller and Pulitzer Prize finalist.[5] People seem to care very much what science has to say about human nature, mind, and behavior. Part of this popularity is intertwined with critiques of the approach. Human sociobiology and Evolutionary Psychology have been thoroughly criticized in and out of academia. In *Adapting Minds*, Buller (2005) summarizes the controversy over Evolutionary Psychology beginning in the 90s that, to some extent, continues today. All of the following critics of Sociobiology and Evolutionary Psychology acknowledge the popularity the subject has and the social importance it seems to hold. Some criticism of Evolutionary Psychology that came from popular or political bodies criticized it for its tendency to produce evolutionary justification for pre-existing biases, political and social structures, as well as racism, sexism, and other forms of discrimination (Buller 2005, 3–6). Evolutionary Psychologists tended to respond by emphasizing the well-confirmed consensus that humans are an evolved species, and maintaining that their critics, in rejecting Evolutionary Psychology, are rejecting the tenets of evolution, a blatant commitment to ideology over scientific truth (Buller 2005, 11–12). This response is likely motivated in part by the way they saw themselves as breaking from mainstream academic disciplines of psychology and sociology, which, they argued, stubbornly ignored evolutionary theory for unscientific reasons (Barkow et al., 1992). By framing the scientific and public debate over Evolutionary Psychology in this false dichotomy — acceptance or rejection of evolutionary theory writ large — Evolutionary Psychology asserted itself in the public eye as the primary, if not only, truly scientific approach to understanding the human mind (Buller 2005, 10–12).

Other critics of Evolutionary Psychology, especially from within philosophy of science, take Evolutionary Psychology to be a failed project on the basis of evidence as well as a misinterpretation and/or misuse of the basics of biological evolution. Buller takes up this project throughout the rest of *Adapting Minds*. His approach is similar to Philip Kitcher's in *Vaulting Ambition* (1989), in which Kitcher criticizes human sociobiology on basically the same charges. Elisabeth Lloyd (2010; 2003; 2002; 1999) focuses her criticism on the evidence

[5] https://www.pulitzer.org/finalists/steven-pinker

that Evolutionary Psychologists invoke, their persistent reliance on radically modular and adaptationist assumptions in spite of their verbal acknowledge of the limitations on those assumptions, and on the connections between Evolutionary Psychology's assertions and political and social beliefs.

While these authors focus on the evidential and conceptual problems with Evolutionary Psychology, they also highlight the fact that the deeply human subject matter of human evolution and the popularity of evolutionary origins for human behavior makes it an exceptionally high-stakes science. Kitcher emphasizes an inductive risk argument as motivation for his critique: because human sociobiologists (and later Evolutionary Psychologists) and the public audience alike take their conclusions to inform social beliefs and policies about race, gender, sex, egalitarianism, and discrimination, the stakes are high. The consequences of asserting, publicizing, and acting upon faulty beliefs about, e.g., evolved sex differences in aggression would be, as he says, "grave indeed" (Kitcher 1989, 9).

These concerns, too, remain present in Evolutionary Psychology as they were in sociobiology. Gillian Barker, in a more recent discussion of the relationship between human evolutionary biology and social change, writes that "It has become plain that substantial social change of some kind will inevitably occur over the next few human generations, driven by new conditions of population density, resource scarcity, and climate change – conditions that are themselves in turn strongly influence by what humans do and how we interact with one another. We face imminent and consequential choices about which social change to pursue and which to resist, and by what means" (Barker (2015), 1). She goes on to show that Evolutionary Psychologists have tended to interpret their own findings with undue pessimism, indeed, what she calls *biofatalism* about the potential for human social change. Her focus is different from the critiques mentioned above in that she critiques the relationship between the work and conclusions of Evolutionary Psychologists and the social views they profess in their popular work, rather than only the problems with the connection between evidence and the conclusions themselves. She also aims to show a positive view of what sound scientific work on the evolution of human psychology can tell us about social change. Thus she both shows the impact that Evolutionary Psychologists have had on the public as well as herself exemplifies the idea that the conclusions of an evolutionary psychology ought to have an impact on human societies. Taking such a stance is not my aim here, but this helps to show that philosophers, biologists, and the public are and often believe we ought to be invested in truths about the origins of human cognition and behavior.

What of narrative, then? When Stephen J. Gould and Richard Lewontin criticized sociobiology for "telling just-so stories," they were, like the critics above, criticizing the tendency for sociobiologists to accept hypotheses on insufficient evidence.[6] Accepting claims on the basis of insufficient evidence is bad practice. But it is, as it were, *normal* bad practice. There is nothing in this criticism to suggest that there is something inherently suspicious about narrative structure. Moreover, it is not clear that explanations in Evolutionary Psychology and sociobiology actually did have the explanatory structure of a narrative even if they were sometimes communicated as stories. So while at first it may seem puzzling that so many recent hypotheses seem to take the form of narratives, given that *storytelling* is such a recognizable criticism of those approaches that the narrative approach stands in contrast to, once we see that these criticisms were rooted in the use of evidence and the interpretation of evolutionary theory, we can see that there is room for — and need for — an analysis of narrative as an explanatory structure in these new hypotheses.

The diversity of approaches to human cognitive evolution discussed above, and their growing recognition within and outside academia, is surely a good thing, in contrast to having a single dominant way of looking at the history of the mind. Hence this dissertation, in which I look at the overarching features of prominent hypotheses across the disciplinary lines referenced above, and that feature seems to me to be *narrative*. Narrative or historical explanation

[6] See also Olmos (2022).

is its own topic of persistent interest in philosophy of science, but also more broadly – in literature, rhetoric, (general) philosophy and more. The interaction between evolutionary explanation and narrative in these other disciplines has been happening throughout the lifetime of evolutionary theory and philosophy of biology.

This makes sense, because narrative is a pretty broad topic – there are many questions to explore about narrative beyond those about when or how narratives function as *explanations*. Areas where people are particularly concerned about narrative *explanation* are history and the historical sciences – for good reason, as these are the places where we're (generally) concerned to gain knowledge about the past natural world. Whether or not narrative explanation is a distinct kind of explanation has implications for philosophers trying to understand the diversity of explanation in science, for philosophers and historians concerned with whether their explanations are distinct and whether history can share explanatory strategies with science. Carl Hempel somewhat famously argued that historical explanations (which I'll use interchangeably with narrative explanations) can't be explanatory *at all* – they can only be sketches of explanations, which won't be explanatory until we expand them to reveal a general law that they follow (Hempel, 1942).

Since then, others have challenged this idea and presented narrative explanation as a category of its own that applied in situations that traditional models of explanation, e.g., the DN model, did not seem to apply. The topic persisted throughout the latter half of the 20$^{\text{th}}$ century, crossing disciplinary lines in philosophy of science and history, e.g., Richards (1992), Hull (1975), Gallie (1968), Mink (1966), Danto (1962). The last decade has seen a resurgence of attention to narrative and historical explanation, especially from philosophers of the life sciences, although the topic is still quite interdisciplinary – for example, Paul Roth's *The Philosophical Structure of Historical Explanation* (2020) takes an analytic history perspective and his work has had significant uptake within philosophy of science.

A productive debate among John Beatty, Eric Desjardins, Marc Ereshefsky, and Derek

Turner (Ereshefsky and Turner, 2020; Beatty, 2017, 2016; Desjardins, 2011), along with work from Kim Sterelny (Sterelny, 2016), has shown that something like contingency or path-dependence is important to narrative explanation – my proposal about instability in Chapter 1 responds to this discussion. Other contributions come from the mechanistic explanation and philosophy of historical sciences literatures (Swaim, 2021; Currie and Swaim, 2021; Swaim, 2019; Currie, 2014; Glennan, 2010). In focusing on causal explanation and historical explanation of evolved traits, I aim to respond and contribute to these discussions as well.

In the context of human cognitive evolution, I see narrative as an explanatory structure, rooted in the causal properties of the explanatory target itself. My approach to understanding narrative explanation is, rather than to look at a narrative and ask what is explanatory about it, to ask what a narrative looks like, and then ask what circumstances might produce an explanation that *looks like that.* There is a fruitful body of work in the philosophy of scientific explanation that can be brought to bear on questions about narrative explanation and about explanations of human cognitive evolution. Of course, this intersection itself is not new: evolutionary explanation is an important philosophical topic in and of itself. But what is novel is bringing to bear recent developments in manipulationist theories of *causal* explanation on these questions, especially analyses of the dimensions along which causal relationships can vary and thus be compared. In addition to addressing the question of what a causal relationship is and how such relationships can be discovered, philosophers of causal explanation have developed concepts to capture variation among causal relationships. These include properties of specificity (Woodward, 2010; Waters, 2007; Lewis, 2000), stability and proportionality (Woodward, 2010; Hitchcock and Woodward, 2003), and reversibility (Ross and Woodward, 2022). These properties can be measured and tested against the world and can allow scientists to compare causal relationships and evaluate them with respect to their explanatory value. E.g., in some contexts, even though identifying any causal relationships would constitute at least a weak causal explanation, identifying a specific causal relationship would be a *better* explanation compared to another less specific contributing cause. That is to say, some of the work in these disciplines is guided by the causal properties of the sequence they are interested in and thus produce narrative explanations. I will show in Chapter 2 that much work on narrative explanation is explicitly but imprecisely causal, and much of the rest has been implicitly causal. Thus I contribute a more precise application of principles of causal explanation to narratives of human cognitive evolution.

My approach in my dissertation is to first understand what kind of explanation is going on in these story-like hypotheses, and then to understand how that affects the strategies and reasoning patterns that theorists take, and also to evaluate the approach and its challenges. Ultimately, I show that once these hypotheses are characterized and evaluated *as narrative explanations*, we can better understand this recent, promising approach to human cognitive evolution. And so to answer the question of how narrative structure affects the strategies and challenges for theorists of human cognitive evolution, especially compared to alternative approaches, I argue in Chapters 2-4 respectively that (1) Narrative explanation is causal explanation of token phenomena with unstable beginnings (2) instability guides the construction of narratives in the evolution of human cognition, and (3) those strategies have weak points, and even good theorists have faced problems with those weak points. This dissertation is intended to be the beginning of a fuller understanding of the 'narrative approach' as a distinct, important paradigm of the evolution of human cognition, standing in contrast to other approaches to an evolutionary understanding of our psychology.

Chapter 2

Causal properties of narrative explanations tell us what they are and when we'll find them

2.1 Chapter Abstract

Some scientists produce explanations that seem to have a story-like, or *narrative* form. Such explanations are especially prominent in the historical sciences, including evolutionary history. To achieve an understanding of explanatory practice in historical science as well as a fuller understanding of the diversity of explanatory practices across the sciences, philosophers of science have proposed accounts of narrative explanation that seek to identify their unique explanatory features. However, such accounts neglect or misidentify the features that distinguish mere narratives from narrative explanations, and narrative explanations from other causal explanations. In this chapter, I propose a novel account of narrative explanation that situates it within an interventionist account of causal explanation. Using my account,

I propose three dimensions along which narrative explanations can be evaluated and draw several consequences for outstanding disagreements in the literature on narrative explanations. Ultimately, my account clarifies and justifies narrative explanation as a legitimate and distinct form of explanation.

2.2 Introduction

Some scientific explanations appear to take the form of a story – especially those in paradigmatically historical sciences like geology, paleontology, and evolutionary biology. Mass extinction events, extreme climate shifts, and the evolution of traits, e.g., the trichromatic eye, are typically explained by reference to the relevant events and phenomena that preceded them (Turner and AboHamad, 2023; Currie, 2019b). Philosophers of science have often been skeptical of the explanatory power of these stories, beginning with Carl Hempel (1942), who argued that such stories were only sketches of explanations waiting for a scientist to subsume them under a law-like regularity. "Telling stories" is even a pejorative for a scientist – for example, Stephen Jay Gould and Richard Lewontin's infamous accusation that human sociobiologists were only telling "just-so stories" to explain modern human behaviors (Gould and Lewontin, 1979; Gould, 1978). More recently, philosophers of science, especially philosophers of historical science, have been interested in whether or not such stories are explanatory, and if so, what makes them explanatory. That is, do the stories that historical scientists give count as a distinctive category of scientific explanation? Answering this question would help us to understand explanatory practice in historical science as well as the diversity of explanatory practices across the sciences.

Such accounts either locate the explanatory features of narrative in the configuration of a narrative, or in the relationship between an event and its history. I argue that existing configuration and historical accounts fail to identify the distinct features of narrative explanation and I propose a novel account of narrative explanation that situates it within an interventionist account of causal explanation. Using my account, I propose three dimensions along which narrative explanations can be evaluated: change in the likelihood of the outcome given the introduction of proposed events, empirical support for the occurrence of the proposed events, and empirical support for the causal relationships between proposed events. I then draw several consequences for outstanding disagreements in the literature on narrative explanations.

The chapter proceeds as follows: in 2.3, I describe what is required for an account of narrative explanation to show that it is indeed a distinct kind of explanation and argue that existing accounts of narrative and narrative explanation have not met those requirements. In 2.4, I present my account and illustrate it with an example of a narrative explanation of the evolution of human cognition. In 2.5, I propose some implications that follow from my account, including three dimensions along which we may evaluate narrative explanations. In 2.6, I use my account to resolve several outstanding questions in the literature on narrative explanation, and in 2.7 I discuss the distinctions between my account and mechanistic explanation. In 2.8 I conclude and offer some suggestions for further applications.

2.3 Narratives and narrative explanation

To get our bearings for an account of narrative *explanation*, we can set out a few characteristic features of narrative without yet considering how they might explain. "Narrative" may refer to a story simpliciter. But "narrative" may also refer to something more specific, like the *plot* of a story, or it may refer to something more abstract, like the organizing structure of the story, exclusive of other elements like plot, setting, character development, etc. Moreover, narratives have all kinds of functions, epistemic and otherwise: in addition to explanation, they may communicate, entertain, evoke feelings or memories, exemplify, or teach. Narrative is studied in history, literature, and rhetoric, and the extensive literature on narrative and narrative explanation in philosophy and philosophy of science is itself only a portion of the broader literature. Kim M. Hajek points out that many scholars even find it undesirable to circumscribe narrative as a concept; preferring a flexible concept that admits of degree when applied to activities or representations of one kind or another (Hajek, 2022). Due to the wide range in which we find use of narrative, accounts of narrative and narrative explanation are quite diverse as well. Because I will defend an account of narrative explanation, I will discuss only the broad characteristics of narrative representation before turning to existing accounts of narrative explanation.

One central feature of narrative as a form of representation is that a narrative presents a set of ordered and related pieces of information – a narrative is not a list (Roth, 2020; Morgan, 2022, 2017; Roth, 2017; Richards, 1992; Hull, 1975). Moreover, although authors disagree about whether a narrative must be some kind of *temporal* ordering, in any case it cannot be *simply* temporal – a narrative is not a chronicle of time-ordered but otherwise unrelated events. Another feature of a narrative is that its events are organized by some kind of relevance to a *central subject* (Hull, 1975). Hull argues that what makes a narrative seem like a cohesive story is that it is about the same entity throughout. The central subject is a historical individual in Hull's sense, meaning that its identity is maintained through time through some continuity in its constitutive parts. The central subject unifies the narrative – each part of the story is somehow relevant to that subject (Hull, 1975). Narratives also characteristically represent some kind of 'retrospective' information (Ereshefsky and Turner, 2020; Roth, 2020; Danto, 1962). Accounts vary in the way that they flesh out this feature. On some accounts (including my own) this only means that a narrative represents a token event and the relevant events which precede it. On others, the notion of retrospective information is developed semantically, as in Arthur Danto and later Paul Roth's notion of *narrative sentences*, which reference past entities such that the sentence could not have meaning until after the event has occurred (Roth, 2020; Danto, 1962). The foregoing introduction is about as specific as one can get in giving a general definition of ”narrative.” If we try to get more specific than that, we quickly see that it is a broad concept, with various meanings in different contexts.

My aim is to identify what it is about the narratives that scientists give that that makes them explanatory, and under what conditions a scientist is likely to discover or prefer a narrative explanation. I will argue that narrative explanation is best understood as a special case of causal explanation that emerges when the explanandum is a token event that stands in an unstable causal relationship with the first relevant cause in its causal history. I will need to elaborate on the notions of “stability” and “first relevant cause.” As I will explain further in 2.4, causal stability will be construed in James Woodward’s (2003) interventionist account of causation and causal explanation and relevance will be construed as *explanatory* relevance determined by the particular explanandum.

There are, broadly, two ways of approaching an account of narrative (and narrative explanation) in science: configuration accounts and causal accounts (Morgan, 2022; Carrier et al., 2021). Configuration accounts initially emerged in resistance to the mainstream consensus of the Deductive-Nomological (DN) account of scientific explanation (e.g., Richards (1992); Hull (1975); Mink (1966); Danto (1962)). DN explanations made the law-like regularities an essential part of explanation; narrative scholars maintained that there must be some other explanatory property that did not have to do with regularities. Configuration accounts locate the explanatory feature of a narrative in the holistic structure or organization of the events and relations among them. On Hull’s view, for example, there is distinctly explanatory value in seeing the entire narrative arranged into a singular whole (Hull, 1975). Mary Morgan, Derek Turner, Marc Ereshefsky, and Kim Sterelny all offer accounts that highlight the explanatory role of a narrative’s configuration (Sterelny, 2021; Ereshefsky and Turner, 2020; Morgan, 2017). Morgan argues that narratives explain through the ordering processes that produce them: *colligation* – collecting the information mutually relevant to a topic –

and *juxtaposition* – comparing and contrasting the colligated information and generating a question, which the scientist then uses the narrative to answer (Morgan, 2017). Thus, the narrative as a whole represents a scientific question, its answer, and the processes that the scientist followed. Turner and Ereshefsky argue, along similar lines as Hull, that a narrative's "ability to integrate reports of events into a single cohesive description" together with the presentation of those events in a correct temporal sequence give a narrative its explanatory force (Ereshefsky and Turner, 2020). Kim Sterelny, in a discussion of methodology for constructing a narrative of human evolution, emphasizes the "achievement" in producing a coherent narrative for a complex explanatory target – bringing together many different events over an extended period is so difficult that producing an internally consistent narrative *at all* should have some epistemic weight (Sterelny, 2021).

In general, on a configuration account, it is *coherence* and *consistency* that make narratives explanatory. We can think of coherence as the events of the narrative being mutually relevant to one another in some way or another, and consistency as the property that events in the narrative do not logically contradict one another. We can see this idea of coherence and consistency in Sterelny's description of the achievement of producing a historical narrative, and in Hull's account, and in Ereshefsky & Turner's emphasis on a "single cohesive description." Coherence is roughly the same idea as Hull's shared relevance to some central subject – relevance to the central subject is what makes a narrative coherent.

It seems intuitive that seeing the whole picture of some phenomenon and facts that are somehow relevant to it is in some sense illuminating. But this intuition is somewhat misleading. In the context of scientific explanation, a proposed narrative explanation is a hypothesis. Consistency and coherence are requirements for any scientific hypothesis. In other scientific contexts, these would be considered a fairly low bar. We would not even consider a hypothesis that is not internally consistent in its constitutive claims or the predictions it makes. Sterelny is probably right that it is much more *difficult* to generate a consistent and coherent hypothesis to explain a very complex token event, because the demands for evidence might be diverse and difficult to access (Turner, 2005). Nonetheless, internal consistency is a requirement to consider a hypothesis, not reason to believe it. The explanatory force of a narrative, then, does not come from its coherence or consistency. This suggests that the explanatory force of the narrative will (at least implicitly) rely on the explanatory nature of the relations among the elements of the narrative, at least in the context of scientific explanation. Causal accounts are therefore more promising, since they relate narrative explanation to the common scientific practice of causal explanation.

Causal accounts locate the explanatory power of a narrative in the relationship between the explanandum and the relevant events that precede it – they hold that the body of the narrative explains its final event. Informally, the beginning and middle of the story explain the end because those events cause the outcome. Historical narratives are characteristically ordered in time and consist in events and relations between them. Historical narratives are sequences of events causally related to one another (Ereshefsky and Turner, 2020; Beatty, 2017, 2016; Sterelny, 2016; Currie, 2014; Beatty and Carrera, 2011; Gould, 1990). To construct such explanations, scientists empirically investigate phenomena to discover what that sequence was (Currie and Swaim, 2021) – they aim to find "the best approximation of the true causal history" (Swaim, 2019).

The various causal accounts of narrative explanation agree that a narrative's explanatory power is due, at least in part, to the causal relationships that it represents. They also agree that a narrative is not just a standard causal explanation, however. If narrative explanation is a distinct category of explanation, there must of course be features that distinguish *narrative* explanation from other types of explanation – a requirement articulated nicely by Roth (2017, 2020). Roth argues that an explanation that could just as well be represented non-narratively is not an "essentially narrative explanation" (Roth 2020, 42).[1] For example, the operation

[1] This distinction appears in another form by Currie (2014), who argues that there are *simple* and *complex* narrative explanations in historical science. The latter correspond to Roth's essentially narrative explana-

of a car engine has a paradigmatically mechanistic explanation. I may verbally describe the story of gasoline entering the tank, being ignited in the cylinders, moving the pistons, and powering the wheels to move the car, but it is the mechanism doing the explanatory work, not the narrative description. Thus we are looking for the distinctive features of narrative explanation apart from narrative representation, because not every narrative explains in a way related essentially owing to its narrative form. As we saw earlier, narratives are typically invoked to explain *token* events that are not instances of stable generalizations (laws or otherwise). This observation has motivated several accounts to identify what it is about such events that leads to a narrative explanation, as opposed to a standard causal explanation. The historical accounts proposed thus far in the literature, however, have not successfully shown that narrative explanations are in some way distinctive among causal explanations.

One proposed feature is that narratives typically make reference to *alternative possible histories.* For example, John Beatty argues that narratives are branching historical paths connecting events. They are explanatory because they represent relationships of counterfactual dependence between events (Beatty, 2016, 2017). They are especially apt at representing the possible alternative paths that history might have taken (Beatty and Carrera, 2011). Indeed the idea that narratives represent some kind of counterfactual information appears so salient that Daniel Swaim (2019) identifies this as a *distinctive* feature of narrative explanations. He argues that narratives explain actual outcomes in contrast to alternative histories (Swaim, 2019). Swaim takes this to suggest that narratives characteristically deal with *underdetermined* outcomes. But we must be careful – in the historical sciences, alternative histories are only alternative hypotheses. While it may seem intuitive that narratives compare alternative "ways things might have gone," this is not in principle different from considering one hypothesis in comparison to another. If it is true that narratives are invoked in situations of underdetermination, then, it is not because of the form of a narrative explanation, as

tions.

underdetermination is itself not unique to historical hypotheses.[2] If there is something that distinguishes narrative explanation, it must be something else.

Philosophers have identified a number of other candidate properties, either of narrative explanations themselves or the phenomena that they explain. These concepts are *contingency* ((Beatty, 2017, 2016; Beatty and Carrera, 2011), *robustness* (Sterelny, 2016), and *peculiarity* (Currie, 2019a).[3] On Beatty's account, narrative explanation is not itself a unique category of explanation that we discover, but rather a useful format for representing highly *contingent* histories. He characterizes a narrative as a branching path of causally connected events, where each branch point represents an event that is "contingent *per se*" (Beatty 2016, 36). A contingent *per se* event was, in a strong metaphysical sense, *not yet determined* at the time of its preceding events – it is an event whose occurrence sets the historical trajectory off down a path when it *could just as well* have gone down a different one. And it is those contingent events that belong in the narrative, because they are the truly puzzling points that explain why some outcome occurred and not another. It is only histories made up of such choice points that are "worth narrating" (Beatty 2016, 33); that is, the right kind of history to represent in a narrative explanation. "The more forks in the road on the way to the actual outcome, the more points at which history matters," he writes with Carrera, noting that "[a] narrative is just right for emphasizing the critical intermediary points, and would be superfluous without them" (Beatty and Carrera 2011, 491–93).

The claim of indeterminacy, however, is a hard pill to swallow for many scientists. And, if nothing else, it is certainly the case that scientists who reject metaphysical indeterminacy construct and put forth narrative hypotheses, and presumably they are doing so on some other basis than locating the points of metaphysical indeterminacy. So even if one is not committed to determinism, it is useful for understanding the explanatory practices of scien-

[2]Underdetermination may instead be due to a lack of available relevant evidence, although whether less evidence is available in the historical sciences is contentious (Currie, 2018; Turner, 2005)

[3]Swaim (2019) also has an additional feature related to mechanistic explanation; for discussion see 2.7 on mechanism.

tists to have an account of narrative explanation that does not depend on such choice points. But notice that without this notion of contingency as indeterminacy, we lose the force of Beatty's account of narrative explanation as distinctive, because we are left only with causal dependence, and this does not help us distinguish which sequences of causal dependence are, or ought to be, represented as narratives. Eric Desjardins gives a formulation of Beatty's account that does not include metaphysical contingency but in which he argues that historicity involves multiple possible past states, multiple possible outcomes, and causal dependence (Desjardins (2011); see also Beatty and Desjardins (2009)). Desjardins emphasizes that the path a narrative takes is sensitive to initial conditions (Desjardins, 2011). But Turner and Ereshefsky point out that on Beatty's and Desjardins' account, covering-law explanations would count as narrative explanations, because the occurrence of an upstream cause, e.g., placing a piece of copper in sulfuric acid, may be contingent *per se* (we may not have done so), and the occurrence of an effect, e.g., the dissolution of the piece of copper, is counterfactually contingent upon the upstream event (Ereshefsky and Turner 2020, 50). Thus, it seems that this account has not identified what is distinctive about narrative explanations over and above their status as causal explanations.

Adrian Currie has developed an account of narrative explanation in the context of the historical sciences in which he seeks additional properties or features that make narratives distinctive among causal explanations. His account, however, has a similar issue with identifying the properties that distinguish narrative explanations from other causal explanations. Currie's early view on narrative explanation picks out what he calls "embeddedness" and "detail" (Currie, 2014). "Detail is a measure of the specificity, complexity and diffusion of the explanans required for explanatory adequacy" while embeddedness describes those explananda that are "accounted for as a token of a type; an instance of a regularity" (Currie 2014, 1169). These two properties lead Currie to distinguish between "simple" and "complex" narratives – complex narratives are high in detail and not embedded, while simple narratives are low in detail and embedded (Currie 2014, 1169). But note here that Currie's simple narratives then have the same trouble that Ereshefsky & Turner pointed out in Beatty's work: paradigmatic covering-law explanations count as simple narratives. Complex narratives are characterized only descriptively: they have high detail and their explananda are not tokens of type-level regularities, but it is not clear why this is the case. In more recent work, Currie introduces a concept that he argues underlies the properties we find in complex narratives: *peculiarity* (Currie, 2019a). Peculiarity is a feature of an explanandum (in Currie's account, called a "target"). Certain targets are more peculiar than others, and peculiar targets require narrative explanation. Peculiarity is, in Currie's words, "the extent to which [a target's] modal profile depends on the processes that generate and maintain it" (Currie 2019a, 42). Targets, in Currie's words, "arise from, and are maintained by, processes," and a target's modal profile is "the likelihood of a target's properties staying the same or changing" combined with "the ways a target may transform under various conditions" resulting in "a map of the various conditions under which a target might change" (Currie 2019a, 42). He then writes, "When peculiarity holds, the target's modal profile is sensitive to some features of these processes" (Currie 2019a, 42). But note that if "processes" are just those things that generate and maintain targets, then the properties of some target (along with all its potential properties, or potential for change) always depend on the processes that produce them. The degree of peculiarity is supposed to determine whether a target requires a narrative explanation, but there does not seem to be a dimension along which peculiarity could vary. Thus, Currie's account, like Beatty's, boils down to a causal dependence claim. But it is not distinct from other forms of causal explanation and so it cannot pick out distinctive features of narrative explanation.

The final account I discuss comes from Kim Sterelny. He seeks to identify what it is about certain events in history that leads historians to explain them with either a "robust process explanation" or an "actual-sequence explanation," a distinction he develops following Jackson and Pettit (1992) (Sterelny 2016, 523). Sterelny argues that some historical events can be explained by robust processes while others require a detailed causal sequence leading up to the occurrence of the event. Similar to Desjardins' account, on Sterelny's view, to be contingent is to be sensitive to small perturbations in some initial condition. For example, Sterelny argues that the first World War was *not* contingent, because there were many conditions in place such that the specific nature of the triggering event (the assassination of Franz Ferdinand) did not make much of a difference (Sterelny 2016, 527–29). WWI was therefore the result of a robust process – it was going to happen regardless of the triggering event that, in fact, began the war. In contrast, the outcome of some battle is contingent, e.g., "If a battle is lost because the admiral in command has a stroke at the crucial moment. . . [I]ts causes are not likely to be captured by an analysis of the relevant military factors before the battle" (Sterelny 2016, 522). Here, an actual-sequence explanation – a detailed causal story – is required because the occurrence of the event is not very robust – that is, whether or not it would happen in many possible worlds – is comparatively low. The decisions of particular individuals often require such explanations. This is the reason that some larger-scale historical events end up contingent, because human societies developed "command-and-control systems that put individuals in power such that their idiosyncratic decisions, rather than the more robust aggregate behaviors of groups, end up with more causal influence (Sterelny 2016, 534–37). In human history, events that are the result of individuals' decisions may require actual-sequence explanations because those individual decisions are much more fickle than the aggregate behaviors of groups.

But Sterelny's explanation types are tied to the *occurrence of the event* – they answer the question of whether, in a global sense, the event *would have happened anyway* under somewhat different circumstances. Accordingly, the examples of events that are explained by robust processes answer this question. But they do so by pointing out that the events in question were causally overdetermined. Sterelny's examples of robust process explanations – WWI, the Nuer conquest in Sudan, the retreat of the Bubonic plague in western and central Europe, the European demographic transition to smaller family size – are all examples of causal overdetermination. For each case, Sterelny cites a number of causes that overwhelm the possibilities of history turning out otherwise, such that the particular events in each case cease to make much of a difference taken individually. Because the need for an actual-sequence explanation depends only upon whether or not the event would occur in other possible histories, Sterelny's account leaves open the possibility of an event whose occurrence is highly robust but for which there is no robust process to explain it. The robust processes described in Sterelny's examples are cases of causal overdetermination which, in a sense, *amount* to a robust process: situations of various kinds of instability, conflict, and anticipation of conflict lead reliably to wars. The many causes that make WWI's occurrence robust are *related* to one another – there are similarities in these causes that make the situation overall an instance of a generalization, a token of a type. But it is consistent with Sterelny's account that an number of causal histories, unrelated to one another (except for the fact that they produce the same outcome), could simultaneously occur. Suppose that just as my cat begins to push a glass vase off the edge of my kitchen table, my partner bumps into the table as he walks by, and at the same time, an earthquake strikes, shaking the table and causing the vase to fall off. The event is counterfactually robust, but it could not be explained by a robust process. Sterelny's account would tell us we ought to find a single robust process, where instead what explains why the event happened or not is the occurrence of three convergent histories. Therefore, while actual-sequence explanation tracks with a narrative explanation, Sterelny has not identified under what conditions we need or should expect to find one. Sensitivity to initial conditions will exclude events that are counterfactually robust but not instances of robust generalizations.

Because narratives explain token events that are not simply instances of generalizations, something like contingency or sensitivity to initial conditions or path dependence likely *is* at the root of what calls for a narrative explanation. What each of these accounts fails to do is distinguish narrative explanations from other simply causal explanations. We might conclude that therefore narrative explanations are nothing more than causal explanations, without any special formal features – perhaps this means it is only their content that makes them narrative or historical explanations, or that their distinctive features are not in their explanatory form but in their use as rhetorical or communicative tools. I do not think that we ought to take this road, despite my criticisms of these authors, because I think we can say something slightly more nuanced, without giving up that narrative is a useful category of explanation (even though I will argue that it is a useful *sub*category of *causal explanation*). The persistent challenge is to find what it is about a narrative explanation that not every causal explanation has, or, in Sterelny's case, to do so in an appropriately specific way. I can now set out the task for my account much more clearly: what is it about an event of interest that leads us to explain it with a narrative? Is it a feature of the event itself, of its causal relationships with prior events, or a particular set of explanatory aims? The answer, I argue, is a combination of the three: explanatory aims determine the first relevant cause of a token explanandum event, and when that causal relationship is unstable, a narrative is required to fill in the space between. My argument for this claim will involve making good on the notion of a first relevant cause, as well as a concept of causal stability that I promised earlier. In the next section, I argue that a narrative explanation is the causal sequence that links a token event to its first relevant cause when that overall causal relationship is *unstable*. Instability is construed as in James Woodward's interventionist account of causation, as are the rest of the causal explanatory properties of narratives. Ultimately, my account says that it is not entirely accurate to speak of conditions that *require* narratives as if we have a choice of the type of explanation we might use; rather there exist causal sequences, that, as a result of their causal properties and our explanatory aims, *have* the properties that characterize narratives: central subjects, retrospective representation, coherence and consistency, etc. Thus, when we represent or relay that causal sequence in response to a why-question, we are giving a narrative explanation.

2.4 Narrative explanations and causal instability

Because not every narrative explains, it is a mistake to begin an account of narrative explanation by asking how the characteristic features of narrative representation might be explanatory. I approach the problem instead by asking what conditions could produce an explanation that resembles a narrative. I argue that under certain conditions, the *causal* explanation of a target of interest takes on the characteristic features of a narrative, and this is what a narrative explanation is. My account is summarized as follows:

> A narrative explanation is the causal explanation that results when the explanatory target is a token event that stands in an unstable relationship with the first relevant cause in its causal history.

In other words, a narrative explanation is a causal explanation with narrative properties. I will now illustrate this account with an example from human cognitive evolution. I use this example because explanations that resemble narratives in evolutionary history are particularly controversial – indeed, "storytelling" is a commonplace pejorative for explanations in this field (stemming from critiques by Stephen Jay Gould and Richard Lewontin (1979; 1978). By examining this example, we will see how narrative explanation works and that there is nothing inherently suspicious about an explanation that sounds like a story. The case I will examine is Sarah Blaffer Hrdy's narrative hypothesis for the evolution of uniquely human cognition, as presented in her book *Mothers and Others* (Hrdy, 2009). While this hypothesis is ultimately incomplete,[4] it has been accepted as identifying an important component of hominin[5] evolutionary history, and it is also an illustrative example of a genuinely narrative explanation.

[4]Theorists typically accept that the emergence of cooperative breeding was an important event in the evolution of human cognition but contend that other factors were equally or more important.

[5]"Hominin" refers to humans and all of the extinct species with whom we share more recent common ancestors than the LCA that we share with chimpanzees & bonobos.

Figure 2.1 shows a graphic summary of Hrdy's narrative. The end of Hrdy's narrative is the explanatory target: the emergence of the particular set of species-typical cognitive capacities of modern humans. These include capacities for exceptionally prosocial and cooperative behaviors, linguistic capacities, and complex theory of mind. Hrdy also agrees with the consensus view that the evolution of these capacities in hominin brains was due in part to an increase in brain size. The beginning of Hrdy's narrative is an earlier event: the emergence of a species with a particular set of cognitive capacities, ancestral to modern humans and our closest extant relatives, chimpanzees and bonobos. This last common ancestor (LCA) could, among other things, engage with and imitate others at least in infancy and had some theory of mind capacities. The beginning of a narrative is the hypothesized *first relevant cause*. The first relevant cause is determined by the explanatory aims we set out at the start of inquiry and the background theory that informs the explanatory contrast of interest. Here, the explanatory contrast of interest is why modern humans developed the set of cognitive traits that we have, rather than some other possible set. Background evolutionary theory tells us that humans share ancestors with modern species that have a wide range of different sets of cognitive traits, and that the relevant *uniquely human* traits are those possessed by us, and not by our closest evolutionary relatives or the most recent common ancestor that we share with them. Phylogenetic history tells us that changes that determined the unique set of cognitive capacities of modern humans occurred during the period of time between the emergence of the LCA species and the present (more specifically, sometime during the Late Miocene and Pleistocene). The challenge for theorists like Hrdy is to figure out why the hominin lineage ended up with our particular set of cognitive capacities and not another set.

The middle of the narrative consists of several major events that connect the first relevant cause to the explanatory target. Hrdy's primary proposal is the emergence of cooperative breeding in the hominin lineage, though she first postulates a less specific event – some ecological change that introduced the selective pressures that led to the emergence of coop-

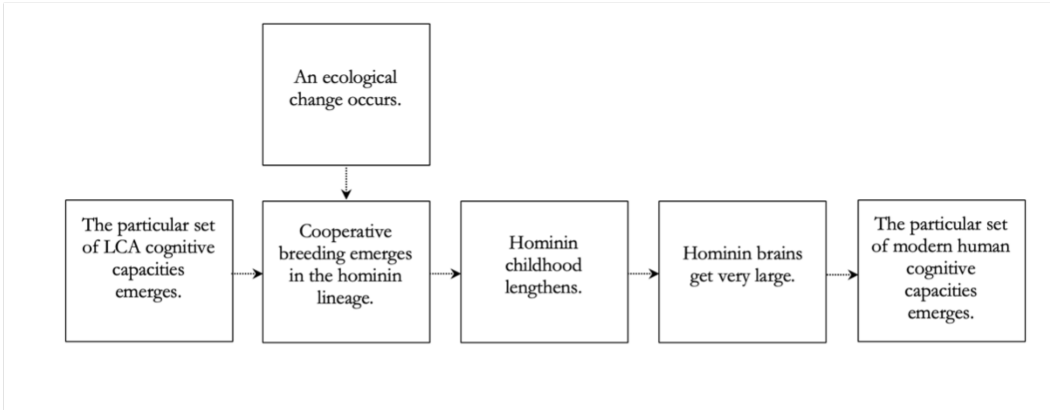


Figure 2.1: Summary diagram of Hrdy's hypothesis for the evolution of uniquely human cognition

erative breeding. The rest of the events are consequences of this important development, in combination with selective pressures that were largely already present but unable to affect the hominin lineage prior to the emergence of cooperative breeding. Cooperative breeding allows offspring to spend more time developing in childhood. With more time and energy to devote to development, brain size could then increase (given that selective pressures for larger brains were already present). Then, with larger brains to work with, selective pressures began to shape the cognitive capacities that we now observe in modern humans. Hrdy's explanation has those properties that characterize narrative representation. She describes a set of events, ordered in time, ending with a token event. Its central subject is the hominin lineage. But Hrdy's narrative is clearly presented as an *explanation* for the emergence of uniquely human cognitive capacities. Why does the resulting explanation have the properties of narrative representation?

Hrdy wants to explain why our particular set of cognitive capacities evolved, rather than some other set. She poses this question as "why us and not them?"(Hrdy 2009, 63) – that is, why did the hominin lineage acquire, e.g., complex theory of mind, while other lineages that share the LCA did not? This is the relevant *contrastive focus* for the explanatory target – it constrains the many possible ways of answering why human cognition is the way it is to the causal factors that make the difference between the hominin lineage and its neighboring

$$\begin{tikzcd}
 & C \arrow[d] & & & \\
S \arrow[r] & H \arrow[r] & L \arrow[r] & V \arrow[r] & E
\end{tikzcd}$$

Figure 2.2: Causal graph of Hrdy's narrative explanation. S = the set of LCA cognitive capacities, C = the occurrence of a particular ecological change, H = the emergence of cooperative breeding, L = the length of hominin childhood, V = hominin brain size, E = the particular set of modern human cognitive capacities.

lineages.[6] Hrdy does not give her story in the form of a precise causal graph, but we can do so easily (see Figure 2.2).

The explanatory target and the emergence of the LCA can be represented as causal variables: S for the beginning set of cognitive capacities, and E for the end, with, say, s the particular set of the LCA and e the particular set of modern humans. Note that before Hrdy even begins to fill in the narrative, there is a causal relationship here. Adopting a minimal criterion for causation from Woodward (2003), the relationship between S and E is causal:

> X causes Y if and only if there are background circumstances B such that if some (single) intervention that changes the value of X (and no other variable) were to occur in B, then Y or the probability distribution of Y would change (Woodward 2010, 63).

The causal relationship $S \rightarrow E$ is borderline trivial; the traits of an ancestral species are obviously causally related to the traits of a descendant species. And in fact other accounts of narrative and narrative explanation have identified counterfactual dependence as an important feature that drive their explanatory force (Beatty, 2016, 2017). But observing that construction of the narrative begins with a hypothesized causal relationship between the beginning and the end is crucial to clarifying its distinctive properties. The relationship

[6] My use of the term "contrastive focus" comes from Lauren Ross (2019), who argues that setting this focus is a necessary part of fully specifying the explanatory target of interest.

$S \rightarrow E$ on its own only holds in very few sets of background conditions. The set of cognitive capacities of an ancestral species fully specifies the set of cognitive capacities of a descendant species only in circumstances in which *no evolutionary change occurs* – that is, when $s = e$. Thus, even though a minimal causal-explanatory view suggests that S explains E, it is trivially explanatory at this point. Nevertheless, our choice of explanatory target and our background evolutionary theory dictates that our story starts here, with the cognitive capacities of the LCA, and ends with the cognitive capacities of modern humans. This property of a causal relationship holding only under very few sets of background conditions is called "instability" in the interventionist framework (Woodward, 2010).

Woodward's concept of instability applies quite naturally to understanding narrative explanation; in fact, the paradigmatically unstable causal relationships he uses to illustrate the concept sound like narratives. For this, he modifies David Lewis's (1986) illustration of what he calls "sensitive" causal relationships:

> Lewis writes a letter of recommendation L that causes X to get a job she would not otherwise have got. This in turn has various other effects: X meets and marries a colleague she would not have married if she had not taken the job, they have children and grandchildren that would not exist in the absence of Lewis' letter, these grandchildren do various things A and so on. Now consider the following claim:
>
> (2.1) Lewis' writing the letter L caused X's grandchildren to exist and to do A.
>
> Given the facts just specified, it follows from Lewis' own theory of causation, as well as the account specified in **M**, that (2.1) is true. Whether or not we accept this judgment, virtually everyone will agree that there is something non-standard, or misleading about (2.1). Lewis traces this to the fact that (2.1) is highly sensitive or unstable. The counterfactual dependence associated with (2.1) may hold in the actual background circumstances but if these had been different,

> in a variety of "small" ways, then if Lewis had written the letter, X's actual grandchildren would not have existed and would not have done A. This might have happened if, for example, X's future spouse Y had not also taken a job at the same school as X, if other contingencies had led X not to marry Y and so on (Woodward 2010, 292).

The token event that ends the causal sequence is that X's grandchildren exist and do A. But the relationship between L and A could easily be broken by changes to a wide variety of other factors. Note also that each claim in the 'story' references X whom we might call the central subject. One aspect of this passage, however, seems to run counter to the account I am developing here. Woodward says that virtually everyone agrees that there is something strange about starting with L. It does not seem very relevant to the implicit contrast class for A (though it does seem more relevant to the contrast class for X's grandchildren's existence). Woodward sees stability as an explanatory virtue, and thus sees this as a reason that we might prefer factors other than L when explaining that X's grandchildren do A. Specifically, we might prefer factors more proximate to A:

> To further explore this notion, consider the connection between stability of a relationship and how proximate or distal it is. Obviously, as a general rule, more distal causal relationships with many intermediate links will be less stable than the individual links themselves. Suppose that we have a chain of causal relationships $X_1 \rightarrow X_2$, $X_2 \rightarrow X_3 ... X_{n-1} \rightarrow X_n$ which holds in the actual circumstances B in the sense that each individual link satisfies **M** in circumstances B and that in addition there is an overall relation of counterfactual dependence in the sense of **M** between X_n and X_1. Suppose that $X_1 \rightarrow X_2$ would fail to hold in some set of circumstances B_1, $X_2 \rightarrow X_3$ would fail to hold in set of circumstances B_2, and so on. Then (assuming no additional complications such as backup mechanisms) the overall dependence from $X_1 \rightarrow X_n$ will be disrupted

> if any one of the circumstances in B_1 or B_2 or B_{n-1} holds. So unless $B_1 ... B_{n-1}$ are strongly overlapping (e.g., most members of B_2 are already in B_1 etc.) the overall $X_1 \to X_n$ relationship will be less stable than any of the individual links $X_i \to X_{i+1}$. Thus to the extent that we value finding stable causal relationships, we will often be able to accomplish this goal by looking for more proximate causal relationships that mediate distal relationships. It follows that a concern with stability can sometimes (but need not always) drive us in a 'reductive' direction, toward the identification of more fine-grained, 'micro' relationships (Woodward 2010, 294–95).

But here we should note a few important observations about the explanatory context for a narrative explanation: the beginning of the narrative – the first relevant cause – was fixed by the explanatory contrast of interest for the ending token event. In Hrdy's narrative, the story must begin with the LCA, because of the explanatory contrast set up for modern human cognition. This means that the overall unstable causal relationship must be part of an adequate explanation. Theorists will prefer causes more 'proximate' to the outcome of interest, but this will not license them to exclude the first relevant cause from the explanation. So theorists are driven to add causal variables that mediate the unstable relationship while retaining the first relevant cause. Rather than dropping away the unstable beginning, as Woodward seems to imply should happen with L in the toy example, the causal sequence will be built up, eventually becoming a satisfactory narrative.

If the causal relationship $S \to E$ were very stable, and we had evidence for the occurrence of b and e, then the causal explanation might be adequate. But it is not so stable. This observation allows us to make sense of what theorists do next, which is to *fill in the space* between the beginning and the end. Each time that Hrdy proposes an *event* in the narrative, she is at the same time introducing a causal variable into her explanation. The rest of Figure 1 is straightforwardly represented by the causal graph in Figure 2. The causal graph in Figure

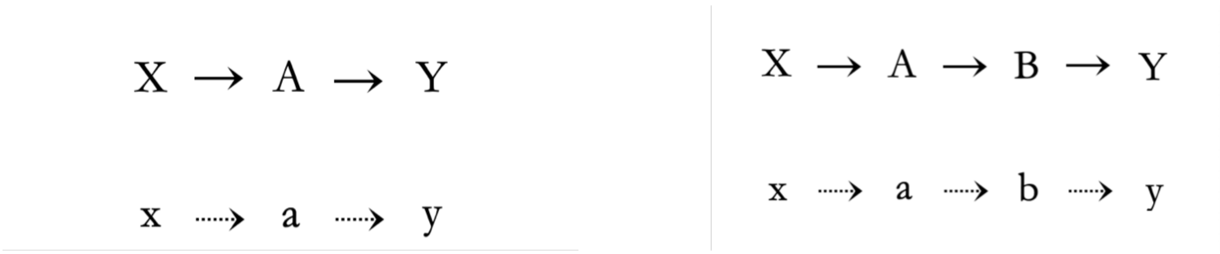
2 represents the narrative that Hrdy proposes and illustrates that it is a legitimate candidate causal explanation. But it is the instability of the causal relationship $S \rightarrow E$ that demanded additional causal steps be added. The stability of the causal relationship $S \rightarrow E$ is a fixed quantity; adding more steps does not stabilize the causal relationship. What theorists seek in terms of explanation is to uncover the factors that produced the outcome of interest *despite* the instability of its connection to its first relevant cause. To this end, adding steps to the causal sequence is akin to *importing* features of the background conditions to the explanation itself or to revealing those particular factors within B (and their values) such that $S \rightarrow E$ holds and e occurred.

Further, steps within the narrative may be more stable than the overall causal connection $S \rightarrow E$. Suppose, as Hrdy argues, that in hominins the length of childhood is stably related to the average brain size. That is to say, the causal relationship $V \rightarrow E$ is relatively stable. Thus, including V and its causal relationships within the explanation does not increase the stability of the causal relationship $S \rightarrow E$ but the explanation does contain a more stable relationship than giving $S \rightarrow E$ alone.[7]

As a matter of practice, however, when Hrdy builds the narrative, she does not proceed primarily by hypothesizing causal variables to introduce. Narrative theorists add *events* – values of the variables that we represent in the causal graph. Many factors guide a theorist in deciding which events ought to be included in the narrative. For example, certain events may be thought to be causally relevant due to independent evidence for their occurrence in combination with background theory that establishes their relevance.

To summarize: a narrative explanation is a causal explanation with narrative properties, which we find in situations in which the explanatory target and its beginning stand in an unstable causal relationship. Narrative construction proceeds by the introduction of events;

[7] I am especially grateful to Lauren Ross for discussion on the foregoing points regarding the introduction of causal steps.



(a) (b)

Figure 2.3: Causal graphs and narrative diagrams for two narrative explanations. As in Figs. 2.1 & 2.2, causal graphs are represented with solid arrow and capital letters for variables while narrative diagrams are represented with dotted arrows and lowercase letters for instances.

the relevant causal relationships added are at least as stable as the overall causal relationship, and ideally, more stable. A narrative explanation is not chosen from among other possible explanatory strategies,[8] rather, it is the form that a causal explanation takes under certain conditions.

2.5 Evaluating narrative explanations

This causal interpretation allows us to evaluate narrative explanations along three dimensions: the number of events included, empirical support for the occurrence of the proposed events, and empirical support for the causal relationships between proposed events.

The difference between Figure 2.3a and 2.3b captures a comparison of the first dimension. As discussed above, the inclusion of additional causal steps in the narrative reveals background conditions that are relevant to the occurrence of the explanatory target and introduces potentially more stable component causal relationships into the explanation. This helps make sense of something puzzling in several discussions of narrative explanation. Currie (2014) observes that narrative explanations have more detail than other explanations. He states that "detail" is a "measure of the specificity, complexity, and diffusion of the explanans" (Currie

[8]That narrative explanation is not always a pragmatic choice is also a feature of Currie's (2014) account.

2014, 1168). But Currie leaves unanswered just how these measures apply to an explanans, and more importantly, just why we should expect a narrative explanation to have more detail. Marc Ereshefsky and Derek Turner (2020) similarly argue that a "thicker" narrative (one which contains more causally relevant events) is better than a "thinner" one, all else equal. They argue that a thicker narrative brings the historical path to the outcome "into sharper focus" (Ereshefsky and Turner 2020, 53). But this metaphor only begs the question; to bring something into sharper focus is only to increase resolution. It does not tell us *why* higher resolution is better. Recall from 2.3 that the motivation for configuration accounts of narrative: that producing a narrative alone is producing an explanation of some kind, and especially the emphasis that Sterelny placed upon the achievement of a narrative and how this speaks in favor of its truth. Why should the internal consistency of a long, complicated narrative speak in its favor? Including more steps shows how and why an unstable causal connection in fact obtained, and may add more stability to the explanation overall in the sense of showing more stable component steps.

The second two dimensions for evaluating narrative explanation concern empirical support. In Figure 3a, I have represented both the causal graph and the chain of events for a short narrative. In order to establish the adequacy of this explanation, we need evidence that the *causal relationships* $X \rightarrow A$ and $A \rightarrow Y$ *and* evidence that events x, a, and y occurred. Surely some evidence may play both roles, but the two can come apart. Hrdy's evidence that cooperative breeding causes an increase in tolerance for food-sharing is drawn from studies on various extant monkey species (Hrdy 2009, 92–98). But separate evidence is needed to claim that the emergence of cooperative breeding and an increase in tolerance of food-sharing occurred in the hominin lineage. This separation captures an intuition that many philosophers hold about narratives: there is a way for a narrative explanation to be somehow epistemically *good* before we know whether or not it is *true*. If there is a substantial amount of empirical support for the causal relationships $X \rightarrow A \rightarrow Y$, the narrative has empirical support, but not yet support for the occurrence of particular values x, a, and y.

2.6 Implications of a causal interpretation

My account resolves several disagreements in the literature on narrative and historical explanation. First, notice that Hrdy's narrative contains events that happen to the hominin lineage, and 'external' events like ecological changes that are causally relevant to events that happen to the lineage. From this we can see that the idea of a central subject as an organizing feature of a narrative is underlain by causal-explanatory relevance: causal variables either represent events that reference the central subject or are causally relevant to events that reference the central subject. Second, the question of whether retrospective knowledge is essential to narratives and what impact this has on the way that they explain is clear in a causal framework: that a narrative explanation is retrospective is derivative of the fact that it causally explains a token event. Third, the broad scope of a causal framework like interventionism means that there is no restriction on the kind of cause that is included in the narrative explanation (whether it is an event or some regularity that becomes causally relevant). This resolves questions in the narrative explanation literature about the status of regularities in historical explanations (Kranke, 2022; Currie, 2018, 2014; Turner, 2009; Jeffares, 2008; Kosso, 2001). Nothing in the form of a narrative explanation dictates whether the causes involved are part of regularities.

The two dimensions of evidence for narratives that I propose make sense of why philosophers are sometimes tempted to treat proposed narrative explanations as "how-possibly explanations" (e.g., Kitcher (2011)). Others have articulated that a how-possibly explanation only shows that some outcome is *possible* under certain conditions in contrast to *impossible* (Reydon, 2012). Proposed narratives that explain why one event happened in contrast to another do not do this. They are proposed hypotheses for the *actual* course of history. The temptation to call these a kind of *explanation* makes sense once we see how a narrative explanation can have empirical support for the truth of the causal claims it makes *without* (yet) having empirical support for the occurrence of the events that instantiate those causal claims.

Producing an internally consistent narrative is an achievement, but not an explanatory one.

2.7 Narratives and mechanisms

An important question remains: Why is a narrative not a mechanism? Mechanistic explanation is a prominent type of explanation in science, especially in the life sciences (see, e.g., Machamer, Darden, and Craver (2000). There are several accounts that characterize narrative explanation not as a distinctive category within causal explanations more broadly, but as a certain kind of mechanistic explanation. There have been two primary strategies for doing so: the first is to characterize narratives and/or histories as a kind of one-off mechanism. This is the strategy taken by Stuart Glennan (2010) and Daniel Swaim (2019; 2021). The other is to characterize a historical explanation as a string of mechanistic explanations ordered in time, an approach developed by Brett Calcott (2009).

Glennan proposes that narratives and histories are "ephemeral mechanisms" — they have all the properties of mechanisms, except they only occur once. The relationships among the 'entities' are stable and change-relating, but the entities do not come together very often, so the mechanism has only had one chance to run. He writes,

> Specifically, I take an ephemeral mechanism to be a collection of interacting parts where:
>
> 1. the interactions between parts can be characterized by direct, invariant, change-relating generalizations
> 2. the configuration of parts may be the product of chance or exogenous factors
> 3. the configuration of parts is short-lived and non-stable, and is not an instance of a multiply-realized type (Glennan 2010, 260).

Similarly, Swaim introduces the idea of a "mechanistic setup" (Swaim 2019, 7). Both accounts seek to legitimize the causal explanatory nature of a narrative by showing that it is characterized by all the same features as a mechanism except for the overall regularity of the mechanism. But this does not apply to, for example, Hrdy's narrative. Some of the causal relationships in Hrdy are not invariant, and some are, but the *overall* relationship between beginning and end is not. The causal relationships within the history may be any degree of invariant, and the pieces of the narrative do not have straightforward interpretations as entities, because they are events. But without the regularity of a mechanism's overall operation and the stable change-relating relationship between entities, it is neither descriptively accurate to call a narrative a mechanism nor explanatorily helpful, because the regularity of the operation of a mechanism is integral to its status as an explanation.

Brett Calcott follows a different line of argument. He introduces the notion of a *lineage explanation*, with the aim not to dissolve the idea of distinctly narrative explanations, but rather to be descriptively accurate to the explanations generated in evolutionary developmental biology. Some authors of human cognitive evolution narratives even take their ideal or aim to be something like Calcott's lineage explanation (e.g., Sterelny and Planer (2021, 5)). Lineage explanations trace how a mechanism could change over time to show continuity between one form and another. In the evolutionary-developmental explanation of the evolution of the focused lens eye, he writes, "the same three layers of cells are present throughout the stages. Their shape, size and relation to one another are modified, but the credibility of continuity between the stages is justified because there is no radical change in parts, or their properties, during the transitions (Calcott 2009, 64)." This example illustrates the two-dimensional nature of a lineage explanation: a production dimension and a continuity dimension. The production dimension captures that each stage of the explanation is an independent mechanistic explanation of how a character is produced. Crucially, what it is that links the stages is not important. He also argues that natural selection does not play a significant role, if any, in a lineage explanation of a trait:

> Likewise, in a lineage explanation, the fact that natural selection caused the change is not part of the explanation. What matters is that we describe the mechanism in a way that we can see how a small modification to some component could have, by whatever process, enabled such a change in the mechanism (Calcott 2009, 67).

But narratives like Hrdy's about are not lineage explanations. A mechanistic explanation at each stage of hominin cognitive evolution is not even desirable, leaving aside the empirical challenge of ever accessing that kind of information. Further, in Calcott's lineage explanations, the contintuity of the sequence of stages only comes from their being *close enough* to one another – "the credibility of continuity between the stages is justified because there is *no radical change* in parts, or their properties, during the transitions" (Calcott 2009, 64, emphasis mine). But here Calcott is relying on an unspecified sense of "radical change." In the case of the evolutionary-developmental explanation of the eye, this sense of "radical change" is probably based on a theorist's trained sense of plausibility relating to mechanical complexity, evolvability, number or kind of mutation that could produce the posited change, etc. This is effectively an informal appeal to legitimate background theory. In contrast to Calcott's characterization, the judgment that one stage is close enough to the next is crucial to the explanation overall. In an evolutionary-developmental explanation of the mammalian eye, informal judgments that a change is small enough may be harmless, but only because they happen to be good proxies for judgments justified by background theory.

But in the context of human cognitive evolution, these evolutionary reconstructions will indeed be stories of natural selection – changes in selective regimes will often form the link between one stage and the next. In this context it is much easier to see that the transitions from one stage to the next must be empirically supported, not idiosyncratic judgements of whether a difference is small enough. Thus, Calcott's account does not accommodate the features of historical narratives, and so cannot account for their unique explanatory features.

Further, given that the judgments of "no radical change" in the evolutionary-development case are in fact proxies for judgments supported by legitimate background theory, it is possible that Calcott's account does not capture all such evolutionary-development explanations either. They may instead be narrative explanations in the sense I argue for in this paper. In any case, neither Glennan and Swaim nor Calcott's accounts of mechanism can account for narrative explanation.

2.8 So, just storytelling? Conclusions and applications

On my account, narrative explanations are causal explanations that result from a set of conditions involving both our explanatory aims and the causal relationships rendered relevant by those explanatory aims. The explanatory value does not arise from the narrative form; rather the form of the narrative emerges from the causal properties of the event we are interested in explaining. My account situates narrative explanation within a well-established framework of scientific explanation, clarifying the explanatory form and justifying its power. It also, as I have shown, resolves several ambiguities in the literature on narrative explanation.

I conclude with a note on applications of my account. One especially salient context for narrative explanation, as we have seen, is evolutionary history. Despite the past criticisms of storytelling, a number of recent, well-received hypotheses for the evolution of uniquely human cognition which are presented as narratives – not only Hrdy's narrative as discussed here, but in hypotheses from Michael Tomasello, Celia Heyes, Kyle Stanford, Kim Sterelny, and Ronald Planer.[9] Why should these theorists choose to present narratives? My account shows that there is nothing suspect about these explanations in virtue of their narrative form alone.[10] Producing a story is not just storytelling; it is the construction of a legitimate causal

[9]See Heyes (2018), Sterelny (2021; 2012), Sterelny & Planer (2021), Stanford (2017b), Tomasello (2014; 2016; 2019).

[10]See also Olmos (2022). Other authors have responded to just-so story criticisms by pointing out other

explanation demanded by the properties of the causal history of the explanatory target. My account may even distinguish these hypotheses as an altogether different approach to the evolution of human cognition compared to alternatives like sociobiology and Evolutionary Psychology. Further, because of its evaluative power, my account offers a framework in which to understand other features of this approach. A more precise understanding of narrative explanation will therefore be invaluable to making progress in paradigmatically narrative fields of inquiry like evolutionary history.

non-explanatory epistemic functions, e.g., Lennox (1991).

Chapter 3

Implications of instability for constructing narratives

3.1 Chapter Abstract

The instability that underlies the structure of narrative explanation also underlies one of its characteristic features – narratives seem to be most satisfying when they include a lot of detail. This connection explains two inferential strategies that theorists use to build narratives that help them to add detail under the constraints of limited evidence. These two strategies are (1) to show that one capacity emerged at a certain point in evolutionary history by connecting it functionally with another capacity whose emergence is evidenced in the material record, and (2) to show that past traits have high evolvability with respect to their enhancement toward traits in modern humans. Both such strategies inherently invoke what I call a "close-enough judgment," wherein a theorist judges that the functional difference between one capacity and another, or that the evolutionary steps between a capacity and an enhanced version of that capacity, are *small* enough or *close* enough in an evolutionarily

relevant sense. Close-enough judgments are not formally problematic but given the challenge of sparse material evidence in the hominin lineage, they constitute a potential weak point in narrative hypotheses.

3.2 Introduction

In Chapter 2, I showed that narrative explanation is a distinctive type of causal explanation that we find when there is instability between a token explanatory target and its first relevant cause. In this chapter, we will see some of the implications of this instability for constructing explanations of the evolution of human cognition. Instability leads theorists to maximize detail in their hypotheses, which explains some of their inferential strategies. This, in turn, highlights some of the potential pitfalls associated with those strategies in the context of human cognitive evolution.

One feature that appears in several accounts of narrative explanation discussed in the previous chapter is their relatively high amount of detail. This seems to be a salient feature of narratives, although of course it is not unique to narrative explanation. As I introduced in Chapter 2, what makes an unstable causal sequence acquire the properties of a narrative is the inclusion of intermediate causal steps. But in human cognitive evolution, material evidence is scarce. To maximize detail in the face of limited evidence, theorists employ particular inferential strategies to broaden the reach of material evidence. These strategies involve what I call a *close-enough* judgment, which, though not inherently problematic, constitutes a potential weak point in evolutionary narrative construction.

The first claim of this chapter is thus an elaboration on a point from the first chapter, where I suggested that narrative explanations can be evaluated on the basis of the number of causal steps they incorporate. In this chapter, I argue that the instability of the beginning-end causal relationship underlies the explanatory standard that evolutionary narratives ought to be relatively high in detail.

The second claim is that causal instability helps to make sense of two inferential strategies that theorists use in building evolutionary narratives. These are (1) inferring that a population had some capacity for which we do not have material or trace evidence by showing that it is functionally associated with another capacity for which we do have such evidence, and (2) inferring that some capacity or ability could be gradually enhanced so as to produce significant change in ability without significant change in underlying cognitive features. Understanding these two inferential strategies highlights an area of potential problems (and I will show examples of such problems in Chapter 4): both (1) and (2) involve a judgment that some capacity is close enough to another in the relevant sense. The demand for detail in narratives leads to these inferential strategies which involve close-enough judgments. In contexts of scarce evidence, and especially in the context of human cognitive evolution, the use and support of close enough judgments can be a point especially susceptible to error.

The chapter is structured as follows: In 3.3, I argue that instability underlies the explanatory standard of high detail in evolutionary narratives. In 3.4 I discuss the challenge presented to narrative theorists by the relative scarcity of material evidence: maximize detail under evidential scarcity. I describe two inferential strategies that theorists use to meet this challenge using Kim Sterelny and Ronald Planer's *From Signal to Symbol* (2021) as an example. In 3.5 I argue that both strategies involve what I call *close-enough* judgments and suggest that these are especially susceptible to error. I conclude by using the insights gained by examining the relationship between stability and detail in narrative explanation to compare narrative construction with explanations pursued with evolutionary models.

3.3 Stability demands detail

Because narratives in general (not just narrative explanations) are stories, it is natural to think that there is something important about the amount and nature of the details they include. There are also several senses in which causal explanations can be more or less detailed: they can include many or few contributing causes for an effect, the cause-and-effect variables may have many or few possible values, and a sequence of causes and effects can include many or few intermediary causal steps. A causal explanation can vary along independent among all these dimensions. As we will see, the lattermost sense is most relevant to narrative explanation.

The intuitive importance of detail in narratives has stood out to philosophers giving accounts of narrative explanation. For example, Adrian Currie argues that detail is in fact a defining feature of narrative explanation (Currie, 2014). In Chapter 2, I discussed Currie's distinction between simple and complex narratives and noted that it is his notion of complex narrative that is comparable to my account of narrative explanation. One of the features of complex narratives, in his view, is that they contain more detail than simple narratives, where detail "tracks the number of explanans, their specificity, and their diffusion: their spread across temporal and spatial grain" (Currie 2014, 1168). He also elaborates that low-detail explanations may have fewer explanans that are each more informative compared to high-detail explanations.

Currie's notion of detail covers a wide range of causal properties, including specificity and the temporal range represented by the explanation. It does not seem appropriate to hang detail directly on the temporal range, because an explanation with less detail could surely span a wide temporal or spatial range. Specificity of the relationship between cause and effect could be construed as a kind of detail, but it is more appropriately related to the property of proportionality (c.f. Woodward (2010)). But the number of explanans does seem to capture the dimension of detail that matters to a narrative. Relative to an explanation of a regularity (at a similar scale, level, etc), we should expect a narrative explanation to have more causes cited in the explanans. We can recast Currie's notion in Woodward's framework by specifying that the number of explanans is the number of causal variables included in the causal sequence.

But importantly, and in contrast to Currie's account, the amount of detail, even construed as the number of variables preceding the final variable represented in the causal system, is not distinctive of narrative explanation. Mechanistic explanations, for example, are often very detailed, depending on the explanandum. So why should be expect in general that narratives have more detail than other explanations? What drives the need for detail in narrative explanation is not detail in and of itself, but the instability of the beginning-end causal relationship. Instability explains why narrative explanations typically want for detail – the causal relationships linking the added detail are at least as stable as the overall relationship and ideally more stable. It also explains why detail seems to be related to spatiotemporal range. Instability will often track with the time or distance covered by the causal explanation, but not always Woodward (2010). Thus detail construed as number of explanatory causal variables will track with the features that Currie includes in his notion of detail often, but not always, because what it really tracks is instability.

Woodward discusses a connection between causal *stability* and *proximity* where "as a general rule, more distal causal relationships with many intermediate links will be less stable than the individual links themselves" (Woodward 2010, 294). Woodward notes that stability is not identical with proximity, but that adding causal steps is a way of responding to instability: "Thus to the extent that we value finding stable causal relationships, we will often be able to accomplish this goal by looking for more proximate causal relationships that mediate distal relationships. It follows that a concern with stability can sometimes (but need not always) drive us in a 'reductive' direction, toward the identification of more fine-grained, 'micro' relationships" (Woodward 2010, 294–95). This helps to show that it is stability that underlies the number of links – the amount of detail – in a satisfactory narrative explanation.

3.4 How to Cook a Wolf: Narrative-building under evidential scarcity

3.4.1 Sparse Evidence

Material traces, such as fossils or preserved artifacts, are often considered the prototypical source of evidence for historical scientists (Currie, 2018; Turner, 2005; Cleland, 2002). But as philosophers have often pointed out, the material record in general is a sparse representation of the traits and variation of past organisms, and the fossil record in particular for the hominin lineage is very sparse (Turner, 2005; Kaplan, 2002). One way of responding to the problem of sparse evidence in the hominin lineage is to adopt a conservative methodology, relying only on very direct inferences from material traces and accepting that the resulting hypotheses will be less informative or less certain than hypotheses in other disciplines (Kaplan, 2002; Lewontin, 1998; Hawkes, 1954).

Skeptical attitudes of this kind have arisen periodically throughout the history of archaeology more broadly, beyond the study of human cognitive evolution, e.g., Derek Turner (2016, 2009, 2005). As Alison Wylie and Robert Chapman show in their (2016) book *Evidential Reasoning in Archaeology*, the question of just how much material evidence can tell us about past cultures has led to a century of "crisis debates," in which archaeologists, anthropologists, historians, and philosophers oscillate between views that emphasize the limitations of sparse material evidence and views that resist this conservatism or embrace the interpretive nature of archaeological theorizing.

How to Cook a Wolf (1942) is a cookbook, memoir, and survival guide by M.F.K. Fisher (and the namesake of the eponymous restaurant in Seattle, Washington). The title plays on the metaphorical wolf of Charlotte Perkins Gilman's poem "The Wolf at the Door," a representation of urgent hunger and the threat of starvation. Fisher's cookbook instructs readers in creative uses of minimal, basic ingredients in the context of scarcity brought on by the Great Depression. Narrative theorists of human cognitive evolution look for ways to cook a wolf – how to make the very most of sparse evidence and produce a good hypothesis.

Wylie and Chapman reject the idea that the availability and use of material traces in archaeology leads to either of the two extreme positions taken by theorists of the earlier crisis debates. While archaeologists should not aspire to theorizing by anything like deduction based on objective readings of material traces, neither should they abandon all attempts to precisely tie archeological theorizing to scientific standards and practices. Along with this rejection of the crisis debate dilemma, Wylie and Chapman emphasize the need to examine the details of particular cases of archaeological theorizing and draw methodological principles and guidelines from the best practices of archaeology and the sciences.

In line with Wylie and Chapman, historical scientists and philosophers respond to the challenge of sparse evidence by developing and analyzing other ways that historical scientists use other evidence, in combination with material traces, to make less direct inferences that go beyond the direct observation of trace evidence but can still be tested and asserted with proportional confidence. Currie (2018, 2015) has referred to this broad approach as "methodological omnivory." This more optimistic approach is not fundamentally at odds with methodological conservatism, as it does not imply that anyone should be making claims without good evidence. Rather, historical scientists and philosophers thereof argue that there is more information to be extracted from material trace evidence once other, relevant sources of evidence are brought into play. In human cognitive evolution, for example, philosophers have defended the carefully constrained use of modern ethnographic data to aid in making inferences about early hominin ancestors in combination with other auxiliary evidence, including material traces (Sterelny, 2022; Currie, 2016; Wylie, 1985). The inferential strategies in *From Signal to Symbol* discussed below are two ways that narrative theorists cook a wolf. These strategies join others identified by philosophers of the historical sciences, including searching for 'smoking guns' (Cleland, 2002, 2001) and minimum-capacity inference (Pain, 2021; Killin and Pain, 2021; Currie and Killin, 2019).

However, note that a major motivating factor in the more conservative methodological views is a concern about unchecked speculation. These more nuanced analyses of evidence in the historical sciences, especially in human cognitive evolution, run the risk of inviting or invoking speculation. Of course, they ideally would not invoke any blatantly unsupported speculation, and some of the same philosophers have argued for the epistemic value of certain kinds of speculation (Currie and Sterelny, 2017). But rather than revitalize the crisis debates that Wylie and Chapman showed to be conceptually misguided, we are now attempting to thread the needle, balancing responsible use of evidence with methodological innovation. Exemplary in this respect is Anton Killin and Ross Pain's analysis of the use of data from WEIRD (White, Educated, Industrialized, Rich, and Democratic) societies in making inferences about past cultures, which shows that diverse samples in psychology and sociology are also of crucial importance to cognitive archaeology because these sciences are invoked as middle-range theory (Killin and Pain, 2023). While it is a basic expectation for scientific reasoning that theorists avoid bad practices like wishful thinking, baseless claims, wild speculation, etc, it is useful to identify the 'weak points' or the parts of a heterogenous, omnivorous methodology, where extra regulation or extra care may be needed. Theorists building narratives of human cognitive evolution are pulled in opposite directions: the instability in the causal connection of interest drives them to add detail, but the problem of sparse evidence poses a challenge to just that.

3.4.2 Two Inferential Strategies in From Signal to Symbol

There is a major pattern of inference that theorists use to establish the presence of certain cognitive capacities in ancestral hominins under scarce evidence, articulated and applied clearly by Kim Sterelny and collaborators in several of their evolutionary hypotheses.[1] They make this strategy explicit when they write, "Our standard pattern of inference is to use the material record to identify the availability of cognitive and social resources for particular communicative capacities, and to use that same record to identify communicative needs that select for those capacities. If capacities were both available and useful, they probably had them" (Sterelny and Planer 2021, 27).

I show here that there is another component kind of inference that that Sterelny and Planer argue for the availability part of the aforementioned strategy, which I call a *close-enough judgment*. Such judgments involve the determination that one cognitive capacity is 'close enough' to another in a relevant sense, and conclude that the hominins in question had the capacity in question. In this context, that sense of 'close enough' is specified either as functionally close enough in the same individuals, or historically in the sense that one is a small evolutionary step from the former. The former sense, which I call a *functional-association inference* is the inference that some population of ancestral hominins had a particular cognitive capacity based on its inferred association with other capacities for which we do have material or archeological evidence.

The second, which I call *historical-evolvability inference* involves the inference that a modern human cognitive capacity is an 'enhanced' version of an ancestral capacity such that gradual evolution (typically, via incremental coevolutionary feedback loops) could produce the modern 'version' by such fine increments that no significant novelty in cognitive architecture was required.

[1]This paper will focus on *From Signal to Symbol,* but see also e.g., *The Evolved Apprentice* (Sterelny, 2012) and *The Pleistocene Social Contract* (Sterelny, 2021).

I first illustrate these inferential strategies, and in 3.5 I show that each strategy

Kim Sterelny and Ronald Planer's *From Signal to Symbol* presents a narrative explanation that traces the gradual emergence of modern human language beginning with the abilities of the LCA. The central novel claim of the hypothesis is that by the time of the Erectines, all of the core capacities for modern human language were present in the use of Erectine protolanguage. The emergence of modern human language throughout subsequent evolution therefore was not a result of any major novel cognitive capacities. Selection on the core capacities present in Erectines enhanced those capacities through the operation of coevolutionary feedback loops.

Sterelny and Planer's narrative has two major parts: first, they show that the capacities of the LCA in combination with material evidence support the inference that Erectines had a protolanguage consisting of "a small set of signs with some fluency to coordinate core subsistence activities" (Sterelny and Planer 2021, 65). Second, they argue that modern human language emerged from the gradual enhancement of those capacities through coevolution among social, cultural, and cognitive features in the changing environment of the Pliocene and Pleistocene.

The LCA, they argue, probably had communicative capacities similar to those shared among the Great Apes (Sterelny and Planer 2021, 32–33). Great Apes have little top-down control of their vocalizations, but have a large number of gestures used for communication. While there is evidence of some contextual flexibility in vocalization, Great Apes show significantly more voluntary control and contextual sensitivity in their use of gestures, leading Sterelny and Planer to claim that Erectine protolanguage was initiated by the expansion of gestural communication (Sterelny and Planer 2021, 36–37). They also attribute to the LCA the following capacities, in forms "roughly similar to those of living great apes" (Sterelny and Planer 2021, 37). They include: short-term memory, control of online-processing capacities, semantic memory, mental models, theory of mind, along with social learning and prosociality

(Sterelny and Planer 2021, 37–47). These are the "main cognitive mechanisms upon which language relies" (Sterelny and Planer 2021, 37).

Certain changes evidenced by the fossil record support Sterelny and Planer's case that descendant hominins with LCA cognitive capacities responded to selective pressures for elaboration in communication. These include the transition to obligate bipedalism and the emergence of Oldowan technology (Sterelny and Planer 2021, 55–63). But the primary evidence for the existence of Erectine protolanguage comes from Sterelny and Planer's strategy of inferring the existence of capacities by supporting their availability and usefulness to the hominins in question. The *availability* of capacities for protolanguage is supported by their argument that rudimentary versions of those capacities were already present in the LCA and therefore the earliest proto-linguistic hominins as well.

Their argument for the *usefulness* of protolanguage comes from five features of Erectine lifeways: hunting, diet, life history patterns, technology, and geographic distribution. Archaeological evidence for each of these features of Erectine lifeways, especially the emergence of Acheulian technology, is coupled with adaptive reasoning about why enhancements in the capacities for protolanguage would have been favored in their own right, primarily through the pressures to elaborate in communication as each of these lifeway features emerged (Sterelny and Planer 2021, 65–77). On their view, then, protolanguage was both available and useful for Erectines, and therefore they had it.

For the second part, Sterelny and Planer seek to explain the expansion of the lexicon of Erectine protolanguage. Hominins eventually arrived at an ability to "expand their lexicon at will" (Sterelny and Planer 2021, 86). This ability resulted from the coevolution of theory of mind and causal reasoning with communication and cooperation. "As both theory of mind and causal reasoning became more sophisticated in our line, inventing signals would have gradually become more intelligent, more comprehending... less dependent on fortuitous coincidence and more dependent on intelligent prediction and interpretation" (Sterelny and

Planer 2021, 86). In other words, small enhancements to theory of mind and causal reasoning would enable small increases in the efficiency of adding new signals to the repertoire, which would in turn create the cooperative conditions supporting further enhancements of theory of mind and causal reasoning, and so on.

Importantly, they argue that the initiation and operation of this feedback loop requires nothing more than the capacities they already argued were present in Erectines: "The establishment of this basic engine for expanding a group's signaling repertoire would not have required a massive upgrade to the cognitive abilities of great apes; certainly, nothing like the theory-of-mind or causal-reasoning abilities of modern humans would have been required, though birds-eye view representation is important" (Sterelny and Planer 2021, 87–88). The initial LCA capacities, combined with the right selective pressures and only small upgrades can support the start of the feedback loop which can continue building upon itself to eventually reach a much more sophisticated lexicon.

3.5 The inferential strategies include close-enough judgments

As I discussed in 3.3, theorists respond to instability in the causal relationship by adding detail, and the resulting causal sequence acquires the properties of a narrative. And stability can be a principled reason to prefer one causal relationship over another as explanatory of a phenomenon of interest (Woodward, 2010; Hitchcock and Woodward, 2003). And this can (but need not) drive theorists to add more and more causal intermediates. But if this is generally true, we must ask when a theorist is justified in deciding they have added enough causal intermediates. There are probably many relevant considerations involving explanatory aims and evidence and the answer is probably different in different contexts. As

I am primarily concerned with evolutionary narrative explanations of human cognition, the more specific question is, when should an evolutionary theorist be satisfied with the level of detail they have achieved?

At this point it is useful to consider again Brett Calcott's (2009) notion of a lineage explanation. It is particularly appropriate here, because Sterelny and Planer explicitly consider *From Signal to Symbol* to be an attempt at a lineage explanation (Sterelny and Planer 2021, 5).[2] Lineage explanations trace how a mechanism changed over time, showing the operation of a mechanism at each stage as well as the continuity between one stage and another. Lineage explanations have two dimensions: a production dimension and a continuity dimension. The production dimension captures that each stage of the explanation is an independent mechanistic explanation of how a character is produced. Calcott's primary example is the evolution of the mammalian eye: three layers of cells in the structure of the eye are maintained over evolutionary time, but the sequence is divided into separate stages as their size, shape, and relation to one another change (Calcott 2009, 64). On Calcott's characterization, continuity of the lineage explanation – the relation *between* one stage and the next – is neither causal nor in need of precise identification. In the evolutionary-developmental explanation of the evolution of the focused lens eye, he writes, "the same three layers of cells are present throughout the stages. Their shape, size and relation to one another are modified, but the credibility of continuity between the stages is justified because there is no radical change in parts, or their properties, during the transitions (Calcott 2009, 64)." That is to say, stages need only be *close enough* to one another – there is "*no radical change* in parts, or their properties, during the transitions" (Calcott 2009, 64, emphasis mine). Calcott requires that the stages in a lineage explanation are not separated by a radical change. To make such a determination is to make a close-enough judgment.

But Calcott takes that judgment for granted – how should theorists judge that one stage

[2]Sterelny also considers *The Evolved Apprentice* to be an attempt at a lineage explanation (Sterelny 2012, 25).

involves no *radical* change from the previous? His characterization here invokes an implicit assumption about what "no radical change" means. Because he developed the concept of lineage explanation in the context of evolutionary-developmental biology, it is likely that in the case of the mammalian eye, the implicit sense of "no radical change" is in fact based on a theorist's trained sense of plausibility relating to mechanical complexity, evolvability, number or kind of mutation that could produce the posited change, etc. In other words, an evolutionary-developmental biologist's intuitions about the plausibility of particular transitions might actually be an informal appeal to legitimate background theory. That is to say, the close-enough judgments are supported by legitimate, albeit unspecified, background theory.

In an evolutionary-developmental explanation of the mammalian eye, informal appeals to background theory may be epistemically harmless, if it is indeed the case that evolutionary developmental biologists can more or less take for granted their implicit beliefs about biomechanical or genetic plausibility; thus, it makes sense that Calcott would conceive of the transitions between stages as he does. In my view, this actually represents a significant issue with Calcott's account of lineage explanation, because he suggests that the continuity dimension is *characteristically* one about changes small enough to be left implicit. But it only happens that in evo-devo, this is, by good fortune of available auxiliary theory, not a problem. The background theory about biomechanics is likely to involve fairly stable causal relationships, and this is perhaps one reason that it seems harmless not to specify them, but they are present all the same. And so if we want to apply the idea of a lineage explanation to human cognitive and behavioral evolution,[3] we will still need to find the analogous close-enough judgments and evaluate their bases. The bases on which these judgments are made will be absolutely crucial.

So, instability drives us toward more detail, the decision that we have enough detail is a

[3]Calcott does not explicitly make this application of his lineage explanations, but Sterelny does so when he presents his hypotheses as attempts at lineage explanations.

close-enough judgment that must, either implicitly or explicitly, appeal to some background theory, and on top of all this, sparse evidence stands in the way of theorists trying to add detail to their narratives. What do we need for the close-enough judgment in the two inferential strategies? For functional-association inferences, we need evidence that ties those functions together cognitively in a way that does not only observe that the two are somehow related in modern humans – their association is only relevant if it is present at the relevant point in the evolutionary past.

For historical-evolvability inferences, we need information about evolvability. Evolvability is a dispositional property of a population – as Rachel Brown argues, it is "a measure of their capacity to change over time with respect to some future state given some starting state of affairs" (Brown 2014, 15). It is therefore, she argues, "in part a measure of how easy it is for a population to move over time in particular directions through design space" (Brown 2014, 16). Recognizing that historical close-enough claims are claims about evolvability helps to identify the kinds of information that can serve as evidence, e.g., "population structure, mutation rate, genetic constraint and developmental compartmentalisation" (Brown 2014, 21).

3.5.1 Examples in From Signal to Symbol

The first examples show the functional-association inference and the close-enough judgment it involves. In Example 3.1, a small-scale case is presented for illustrative purposes. Example 3.2 represents a larger scale case in the sense that the functional association underlies a significant portion of the overall hypothesis.

Examples 3.1 and 3.2 both come from Sterelny and Planer's proposed explanation for the emergence of grammatical structure – in their words, "...an incremental account of some of the basic cognitive capacities on which the structural features of language depend" (Sterelny and Planer 2021, 120). Sterelny and Planer argue on the basis of behavioral evidence from extant great apes that "our close primate relatives have much if the cognitive machinery needed to understand simple forms of syntax" (Sterelny and Planer 2021, 125). These examples also show nicely how the functional-association strategy is used in combination with other strategies, because this claim about great apes comes from a minimum capacity inference:

> [O]ur focus here will not be on the proto-sentences themselves, but on the evolution of the cognitive capacities needed to produce and interpret them. In our view, the empirical ground here is rather firmer, for we think those cognitive capacities were borrowed from ones needed for stone toolmaking. We have a reasonable indication of the techniques needed to make the different kinds of tools found in the record, and of the cognitive demands imposed by those techniques" (Sterelny and Planer 2021, 119).

But for the selective pressures that operate on that baseline, the capacities underlying linguistic structure are connected to other capacities, for which material evidence is available. Thus, here is one example of a functional-association claim:

> **Example 3.1** Importantly, merely getting others to attend to the right agents would rarely work well in messaging about the social domain... Moreover, given the foundational role of such communication in shaping cooperation and collaboration, the costs of miscommunication in this domain can be steep... we can connect the evolution of grammatical devices like linear syntax to archaeological traces of enhanced forms of cooperation and collaboration. It is of course possible that grammar emerged even earlier and/or in response to other communicative needs of our ancestors. But we do think that communication about social events,

> actual and possible, must have been a key driver (Sterelny and Planer 2021, 127–28).

The functional association is between capacities for cooperation and collaboration on the one hand and linguistic/communicative capacities on the other, which are *close enough* because they are employed in many of the same contexts and in service of one another. This association shows what is probably a good close-enough judgment: Sterelny and Planer argue throughout the book and elsewhere (e.g., Sterelny (2012)) for the close relationship between communicative capacities and cooperative capacities.

The following functional-association claim is used to explain the emergence of the capacity to use hierarchical structure in communication. This claim connects on the one hand hierarchical organization for planning and recognizing action, and on the other, hierarchical organization for producing complex linguistic constructions.

> **Example 3.2** The thought, then, is this. Technological evolution in our line drove the evolution of a system for the hierarchical control of complex, extended, and precise action plans, and in social learning for the parsing of complex actions of other agents. This system was an elaboration of capacities already possessed by the ancestors of great apes and hominins: it did not come from nowhere. Once this was in place, our ancestors had all they needed, cognitively speaking, to make use of hierarchical structure in language. They would have been able to represent sentences as having nested part-whole structure. This would have made possible the eventual evolution of conventions that gave a role to such structure in encoding meaning. They also had all they needed to put to use hierarchical structure below the sentence level, yielding morphological principles, and perhaps above this level as well (Sterelny and Planer 2021, 150).

In this case, the capacity to use hierarchical structure in language is associated with the capacity for hierarchical structure in toolmaking and tool use. In fact they are, on Sterelny and Planer's view, really the same capacity. The close-enough judgment is to say that the contexts of tool use and linguistic use are close enough that the same capacity could accomplish both. The support for the association comes from neuroscientific studies on Broca's area, which is involved in hierarchical cognition and some scientists describe as "supramodal" because it is, in modern humans, "involved in processing everything from toolmaking to language to music to arithmetic" (Sterelny and Planer 2021, 141). This provides the basis for arguing that the capacity for grammatical structure is close enough to the capacity for hierarchically structure tool making and use. In this case, they argue that they are in fact the *same* capacity:

> It is true that we are accustomed to thinking of action and linguistic communication as fundamentally distinct. But they are not really. Language is a special case of complex intentional action… the neuroscience is consistent with this assumption. For as was noted earlier, Broca's area and other nearby regions are paradigmatically multimodal. It seems that the areas that carry out the relevant computations do not "care" whether it is toolmaking or language or any other complex intentional activity that they control or identify (Sterelny and Planer 2021, 150).

The second inferential strategy is the historical-evolvability type. Again, Example 3.3 will be smaller-scale to illustrate the inferential strategy, and the Example 3.4 will show where the strategy is used in a larger-scale claim of the hypothesis. In Example 3.3, Sterelny and Planer are tracing the emergence of intentionally shaped Oldowan stone tools:

> **Example 3.3** It may well be, then, that our last common ancestor with chimps and bonobos used stones as hammers, and in so doing produced usable flakes

> as a by-product. A habitat shift requiring more intense extractive foraging — more resistant plant matter, such as roots; the marrow concealed in bones and skulls of large animals…; raw meat needing tenderizing – could have easily led to the intentional use of these primitive flakes, and then, somewhat later, their intentional production. No significant cognitive or cultural advance would have been required (Sterelny and Planer 2021, 142).

They claim that the capacity for modified wooden or unmodified stone tools, under the specified selection pressures, requires "no significant cognitive or cultural advance" to progress so as to give rise to intentionally shaped stone tools. The capacity or capacities for intentionally shaping stone tools are close enough to, or perhaps the same as, those needed for intentional wooden tool production or unmodified stone tool use.

The historical-evolvability inferential strategy is also used at a larger scale in arguing for the operation of a "cognitive-cultural engine" that expands the lexicon of Erectine hominins:

> **Example 3.4** First, the signaling repertoire of the group can become somewhat standardized. This simplifies both signal acquisition and real-time production and interpretation…Second, and more importantly still, it makes signal creation strongly cumulative… a successful innovation will spread to peers in the same generation, and to the next generation. The establishment of this basic engine for expanding a group's signaling repertoire would not have required a massive upgrade to the cognitive abilities of great apes; certainly, nothing like the theory-of-mind or causal-reasoning abilities of modern humans would have been required, though bird's-eye view representation is important. A capacity for basic metarepresentation, together with an understanding of simple intentional states (*sees, hears, attends, desires*), was probably enough to kick-start the process. As these abilities were gradually transformed, this engine would

> have become more powerful still, with a higher success rate, higher bandwidth, higher fidelity" (Sterelny and Planer 2021, 87–88).

The ability to expand the lexicon indefinitely, then, is within the reach of a chain of close-enough changes that the operation of existing selection pressures and cumulative processes can enhance the basic abilities to reach that point. My aim in this section has not been to critique the claims made in the examples above, nor to suggest that the close-enough judgments are illegitimate or poorly evidenced; rather, I mean to show that this kind of judgment is involved in both strategies.

3.6 Conclusions & implications

The inferential strategies used in *From Signal to Symbol* exemplify the narrative approach's response to maximizing detail under evidential scarcity. They bring with them a necessary close-enough judgment that, to be well-supported, must appeal either to the right sort of cognitive evidence or to evolvability. This analysis highlights one feature of the narrative approach as a way of studying human cognitive evolution and also highlights a point of potential problems: close-enough judgments are hard to make with confidence and, as we will see in Chapter 4, present an inroad for problematic forms of evidence.

I conclude by considering some of the insights we can gain from examining the way that instability structures inferences in narrative construction. In Chapter 2, I distinguished narrative explanation from other general forms of explanation; e.g., mechanism. Here, I begin to compare the narrative approach to human cognitive evolution to other approaches by comparing their explanatory features. Understanding the role of (in)stability in directing the narrative approach goes some way in distinguishing it from other approaches to studying human cognitive and behavioral evolution. One such approach is the formal modeling of evolutionary processes explored in evolutionary game theory.

I should note that evolutionary game theoretic models (like all kinds of scientific models, and like narratives!) have many epistemic roles in scientific practice and are not always used for explanatory purposes.[4] But formal models of evolution are, at times, taken to be explanatory, and one suggestion for how they do so is by serving as how-possibly explanations (Verreault-Julien, 2019; Rice, 2016; Grüne-Yanoff, 2013; Craver, 2006). As we will see, it is natural to ask whether the distinction between how-possibly and how-actually explanation makes the difference between the explanations generated by narrative and formal approaches to the evolution of human cognition and behavior. I will argue below that it does not. Instead, these explanations are distinguished by their use with respect to causal stability.

But first I must briefly argue that the narrative explanations of interest here are themselves how-actually, not how-possibly, explanations. This requires some argumentation, because a significant vein of early work on how-possibly explanation in philosophy of biology was concerned with exactly this question of vindicating adaptive hypotheses as not only 'just-so stories' but as a special kind of explanation, namely, how-possibly explanation. Ultimately, my view is that they are instead proposed how-actually explanations (and this is just what we usually mean when we call something a "hypothesis," so it suffices to simply call them "hypotheses").

William Dray is typically cited as giving the original formulation of how-possibly explanation as an alternative to hypothetico-deductive explanation that is frequently found in history (Dray, 1957). In Dray's original illustrative example, he considers a toy example (though it does describe a real event) as a paradigmatic how-possibly explanation. In this example, a radio announcer commentating on a baseball game tells listeners that the center fielder caught a long fly ball headed toward the top of the 20-foot fence – shocking listeners until it was revealed that there happened to be a high platform for the scorekeeper near the

[4] E.g., O'Connor (2017), Parker (2020), among others; see O'Connor (2020) for a review.

fence, accessible via a ladder, which the center fielder climbed up to catch the ball. Of this explanation, Dray writes,

> Explanation is called for because what happened seemed *impossible* under the circumstances. What *were* the circumstances? It may seem perhaps that these have been disingenuously misrepresented as I have presented the problem. For the question seems at first to be: 'How could the fielder have caught the ball at the twenty-foot mark, with absolutely nothing to stand on?', whereas in fact, there was a perfectly solid platform available, with a ladder attached. We assumed that we were dealing with a case of a fielder catching the ball 20 feet in the air, whereas it was really a case of his catching it from a 20-foot platform (Dray 1957, 160).

Thus, for Dray, how-possibly explanation is distinguished by a correction of an audience's puzzlement about the facts that made an apparently impossible event actually happen. Robert Brandon and David Resnik both gave contrasting accounts of how-possibly explanation that instead give a definition in terms of their degree of confirmation (Resnik, 1991; Brandon, 1990). Later, Patrick Forber argued that these previous accounts were conceptually misguided and instead there are three relevant types: how-actually, global how-possibly, and local how-possibly explanations (Forber, 2010). Thomas Reydon criticizes Forber's account by showing that local how-possibly explanations are not explanatory at all and only conjectures (which Forber (2012) happily grants), and that global how-possibly explanations are indeed explanations but should not be construed as *how-possibly*. Reydon ultimately maintains a form of Dray's original conception of how-possibly explanation (Reydon, 2023). Reydon and Forber thus ultimately agree that some hypotheses that have been called "how-possibly explanations" are in fact something else – hypotheses, or as Reydon calls them, "conjectures."

Alisa Bokulich has argued that abstract models of high-level processes like self-organization can effectively be both how-actually and how-possibly explanations relative to precisely what it is they explain and at what level of abstraction that explanandum is found (Bokulich, 2014). That is, the behavior of a model of self-organization in, e.g, the tiger bush vegetation pattern, is a how-actually explanation of the abstract organization process that it represents, but it is left open whether or not particular systems in the world are instantiations of that abstract process. Thus, with respect to those candidate systems in the world, it is a how-possibly explanation. Bokulich thus responds to the debate between Forber and Reydon by clarifying what sort of explanandum can be explained by a formal model of an abstract, constructed phenomenon. The model's behavior (actually) explains features of the abstract system that it defines.

What we can conclude from this debate is that the central defining feature of a genuine how-possibly explanation has to do with whether its explanandum is itself a contrast about possibility. Note that in some cases, Bokulich's how-possibly explanatory role for abstract models does effectively treat them as hypotheses for the operation of a particular system. But in other cases, i.e., for other explanatory contrasts, the model might show that, e.g., while relevant theory might have led you to believe that a tiger bush pattern could not be produced by a self-organizing system, in fact it can. The model is a genuine how-possibly explanation for this explanatory target whose contrast is about possibility. This is the contrast that Dray had in mind, after all, although his example is misleading, because it is also a regular old (how-actually) causal explanation. A well-supported causal explanation always does both jobs of showing possibility and actuality; a genuine how-possibly explanation for Dray's toy example would only need to show that a baseball player could catch a ball near the top of a fence by standing on something tall. Because the contrastive focus in these explanations is possibility versus impossibility, we cannot, at least in the context of scientific explanation, hang how-possibly on the mere fact that the audience did not have the relevant facts to see the possibility of the event; there have to be good positive reasons *to think it impossible* in the scientific context. For example, many 20$^{\text{th}}$ century evolutionary theorists believed that altruistic behaviors *could not possibly* evolve by natural selection. Perhaps altruistic behaviors in nature are only apparent, or else they came about by some mechanism other than evolution by natural selection. But formal models of altruistic behavior show that it is in fact *possible* for altruistic behaviors to evolve within the typical constraints of natural selection.[5] The relevant explanatory contrast is, again, whether some causal history is *possible* versus *impossible*.

Authors of narrative hypotheses for human cognitive evolution, or supporters thereof, sometimes defend these hypotheses as provisional, merely how-possibly explanations. Some do so explicitly, e.g., Kitcher 2011. But to call these hypotheses "how-possibly explanations" is only to hedge in asserting them as hypotheses (or "conjectures," in Reydon's terminology) – perhaps in the interest of protecting speculation. To call them so is only to express lower confidence in the hypothesis, relative to hypotheses one might make in other contexts. This is not the relevant explanatory contrast for the narrative hypotheses under consideration here. They propose how human evolutionary history *did* happen in contrast to other ways it might have happened. They begin with the premise that (at least some features of) human cognition evolved; there is no question about the *possibility* of the evolution of human cognition at all. Moreover, it is not the case that all of these theorists even present themselves as proposing a low-confidence suggestion of how things went. As we saw earlier, Sterelny identifies his work as proposed lineage explanations. More strongly, Tomasello concludes *A Natural History of Human Thinking* with the assertion that "something like the Shared Intentionality Hypothesis just must be true" (Tomasello 2014, 153). This is not a statement of conceptual possibility. It is an assertion of the actual course of evolutionary history.

Once we see that narrative explanations are not themselves how-possibly explanations it is natural to ask whether this is a point of divergence that characterizes different approaches

[5]This example is discussed widely, but especially clearly in Sober (2024), though he uses the term "proof of concept" rather than "how-possibly explanation."

to human cognitive evolution: that narrative explanations are how-actually explanations while modeling explanations are (genuine) how-possibly. There is an immediate sense in which this distinction does not apply, given that modelers pursue questions that are clearly about the evolution of human cognition and behavior that are intended to be and ultimately serve as how-actually explanations for abstract, constructed phenomena. Evolutionary game theorists can provide either how-actually explanations of abstract phenomena or genuine how-possibly explanations for things that our understanding of evolution would lead us to believe were impossible.

So what does make the difference? One important difference is related to (in)stability. One way that formal models can be shown to be explanatory is by robustness testing (O'Connor, 2020; Weisberg, 2013; Levins, 1966). Robustness testing, Weisberg writes, proceeds in three steps. First, theorists study a "diverse set" of "similar but distinct models" to identify robust behavior across the different models. Second, they find the "core structure which gives rise to the robust property" shared by the models. In the final step, theorists combine the results from the first two steps to "formulate robust theorems," which have the form of a conditional statement connecting the core structure common to each of the models to the robust behavior that they display, qualified by a *ceteris paribus* statement (Weisberg 2013, 158–59). Ideally, at the 'end' of robustness testing, a modeler is confident that their model's core structure explains the behavior of the system it represents (whether that is an abstract, constructed, or real-world system).

Contrast this with the way that narrative explanation construction proceeds: narrative explanation owes its form to the instability of the connection between beginning and end; there is no robust core to be found (even if certain causal links in the narrative are themselves robust). When there is a core to be found, robustness analysis should find it; narrative is what we get when there is no core to be found; it is another way of responding to instability. To put it briefly, modelers strip away detail to find the robust core of the model while narrative theorists, with no robust core to find, build up detail to add more stable causal relationships to the sequence.

To close, I will briefly address another side of the idea that these hypotheses are provisional, which is related to the idea of judging close-enough. No theorist discussed in this dissertation would claim that they have given the definitive and complete narrative explanation of human cognitive evolution, and likely none of them would turn down additional relevant detail, and often the hypotheses are to a high degree complementary to one another. And, of course, sharing these hypotheses has other epistemic benefit aside from stating what the theorist believes to be ultimately true (furthering debate, sharing evidence that has been neglected, reframing certain questions and assumptions). But evolutionary theorists, perhaps in contrast to some historians and historiographers, do seem to hold at least an ideal of discovering *the* true history of human cognitive evolution. And to publish a hypothesis is to assert it with a certain, relatively high, degree of confidence, even if authors are self-effacing about which parts of the hypothesis are more speculative than other parts. With this goal in mind, and with the necessary role of close-enough judgments, it is an important part of evaluating hypotheses that take the narrative approach to examine how these judgments are made, and what it is that justifies a theorist in claiming that they have 'enough' detail for a satisfying narrative explanation.

Chapter 4

What would imaginary ancestors do?

4.1 Chapter Abstract

The reconstruction of the evolutionary history of human cognition is a complex, interdisciplinary science. Human evolutionary theorists generally use a wide variety of evidential sources to construct hypotheses for the emergence of novel cognitive capacities in the hominin lineage. But sometimes, they take certain associations among cognitive capacities to be intuitively plausible – so much so that they accept and include them in their hypotheses. I argue that when they make these plausibility judgments, they are actually attempting (explicitly or implicitly) to reason from the perspective of an ancestral hominin. They reason about whether some inference would be easy, natural, or obvious for a hominin with only a hypothesized set of ancestral capacities. When they do this, they invoke their intuitive sense of ease or obviousness as evidence for the association among cognitive capacities. But such intuitions are systematically misleading, because they are influenced by the very cognitive capacity whose evolution the theorist seeks to explain. Intuitions that stem from the evolved capacity in question thus make some associations among cognitive capacities appear

artificially plausible. I analyze examples of this kind of reasoning from three recent hypotheses for human cognitive evolution. I argue that this is a particularly pervasive challenge for evolutionary theorists. Even though they are typically highly attuned to similar errors, theorists should expect to keep contending with this challenge, even in our best approaches to human cognitive evolution.

4.2 Introduction

It is notoriously difficult to reconstruct the evolutionary history of the unique features of human cognition and behavior, but the fact remains that human cognition *did* evolve. Our complex, plastic minds have an evolutionary history that, at least in principle, can be discovered. And we ought to try. As Kim Sterelny has remarked, "[o]ur evolutionary history *must* be an important part of the explanation of our being the kind of creature that we are" (Sterelny 1995, 365, emphasis mine). The persistent question for evolutionary theorists is how to reconstruct that history in light of the associated challenges, most notably, that evidence for past cognition and behavior is relatively scarce and difficult to interpret. In this paper, I examine a related challenge whose detail and significance has not yet been recognized.

When theorists reconstruct the evolution of human cognition, they aim to build a detailed history: they want to know what the important events were, what difference these events made to the selective pressures affecting ancestral hominins, and how this impacted their behavior, survival, and reproduction. Ultimately, they aim to uncover when and how our uniquely human cognitive capacities evolved, for example, capacities for mindreading (also called "theory of mind"), moral psychology, hyper-cooperation, etc.[1] In the hypotheses I will

[1]By "uniquely human cognitive capacities," what I mean is that humans have some distinctively human version of those capacities, and not that no other species has a capacity or behavior that could appropriately be called, e.g., "mindreading." The existence and nature of such autapomorphic traits is itself an active

be concerned with in this paper, the resulting explanation is a broad-scope narrative that synthesizes the causal history of uniquely human cognitive traits. Such narratives are one type of explanation pursued in the emerging tradition of cognitive paleoanthropology (Currie et al., 2024). In this paper, I focus not on these narrative hypotheses overall, but on two specific kinds of claim within them that theorists sometimes take to be intuitively plausible: claims that the capacity of interest is associated with another capacity, and claims that the capacity of interest could easily emerge in populations with certain ancestral capacities. I will argue that in judging these claims to be plausible, evolutionary theorists sometimes invoke a systematically unreliable kind of evidence: intuitions about the associations among cognitive capacities.

It happens like this: in order to establish that one cognitive capacity emerged at a particular time or in response to a particular selective pressure, the theorist judges that one capacity is *close enough* to another, such that either the presence of one indicates the presence of the other, or the emergence of the other following the emergence of the first is a small enough evolutionary step. To do this, theorists explicitly or implicitly try to take the perspective of imagined hominins to reason about whether some inference would be easy, natural, or obvious for a hominin given only a hypothesized set of ancestral capacities. To put it abstractly, theorists make a judgment about a novel capacity Z in one of the following ways: (1) reasoning that, given that an ancestral hominin had X and Y, it is plausible that they also had Z, or (2) reasoning that in a population of hominins with X and Y, it is plausible that Z could readily emerge.

Formally, these judgments are unproblematic. But they are sometimes supported not by independent reasons for the association of Z with X and Y but by evolutionarily biased intuition. In this context, we are interested in capacities – theory of mind, moral psychology, etc. – that are adapted to facilitate specific kinds of inferences with ease. Modern humans

empirical question, and narrative explanations of these traits must involve or appeal to evidence regarding the trait's uniqueness.

do not find those inferences easy because they are easy in a principled evolutionary sense; we find them easy because to make them, we employ a cognitive capacity that is specialized for just that kind of inference. But that is the very cognitive capacity whose evolution we are trying to reconstruct, so we cannot take that sense of ease for granted. Our intuitions cannot help us determine whether X and Y are 'close' enough to Z.

Consider this toy example. Humans have a species-typical capacity for mindreading; the intentions of others often strike us as immediately evident. From early in development, humans attribute intentions to stimuli as minimal as shapes moving on a computer screen (Ratajska et al., 2020; Heider and Simmel, 1944). This capacity for mindreading is typically taken to be an evolved feature of uniquely human cognition.[2] When we experience the sense that the intentions of others are immediately evident, we are experiencing our evolved sensitivity to those intentions.[3] Suppose mindreading is the capacity of interest (Z, above). Suppose a theorist says, "Imagine an ancestral hominin, Hominin 1, whose attention is attuned to the actions of conspecifics (capacity X) and who can represent their own performance of a task as goal-directed (capacity Y). Picture Hominin 1 watching another hominin, Hominin 2, peel the bark off a stick. Hominin 1 sometimes peels the bark off sticks to make flexible foraging tools. With those existing capacities in that scenario, it seems that Hominin 1 would be able to represent Hominin 2's actions as goal-directed toward making a flexible foraging tool." While claiming that it is the scenario and the existing capacities that make it plausible, the theorist is leaving out an important question: what makes it seem like X and Y are enough for Hominin 1 to also represent Hominin 2's own intentions? When a theorist judges this association to be plausible, they invoke their intuitive sense of ease or closeness as evidence. But this intuitive sense of ease is actually the product of evolution, not its explanation. Of course, this toy example is exaggerated; the examples I will show are

[2] Though the extent to which hominin mindreading capacities are evolutionarily novel and to what extent they are shared with our closest evolutionary cousins, chimpanzees and bonobos, and therefore our shared LCA, is actively debated; see 4.6.

[3] This is true regardless of the particular evolutionary processes by which that capacity has evolved.

not nearly so coarse-grained.

Though the use of intuition to make plausibility judgments about the evolutionary relationships among cognitive capacities is related to anthropomorphism (as I discuss further in 4.6), this is not a case of the more common anthropomorphic error of simply projecting our abilities onto non-humans. Rather, theorists assess the ease of some inferences, given the presence of other capacities, by the light of their own experience. This has led theorists to explicitly appeal to intuition as evidence, mistake evidence from the operation of modern human cognition for evidence about its emergence, and in some cases to avoid attempting to answer certain questions about evolutionary history entirely.

I focus on three recent and prominent hypotheses: Michael Tomasello's Shared Intentionality Hypothesis, Kyle Stanford's correlated interaction account of moral externalization, and Kim Sterelny's apprentice model. Each of these hypotheses aims to explain why modern humans have one or more of the unique cognitive and behavioral features that we do. Sarah Blaffer Hrdy, in her own hypothesis of this kind, poses the central explanatory question as "Why us and not them?" (Hrdy 2009, 33–37). Why did humans end up with these capacities, while our closest evolutionary relatives did not? The project is broadly adaptationist in a methodological sense (Godfrey-Smith, 2001) in that theorists expect that the distinctive features of modern human cognition emerged as adaptations to the changing environment throughout the Late Miocene, Pliocene, and Pleistocene. However, they make no strong assumptions about optimality, and the resulting hypotheses are not strictly adaptationist – causes of any sort throughout the history of the hominin lineage may be relevant to the explanatory target. Unlike in the paradigmatic method of mainstream Evolutionary Psychology (see Barkow et al. (1992)), these theorists do not assume the modularity of cognitive systems with specific associated adaptations. Rather, they aim to find out what these cognitive capacities are and what drives their evolution – the details of the complex interaction of ecological environment, social and cultural interaction, and genetic and developmental change. Accordingly, we expect theorists to adhere to standards of good reasoning about adaptation, for example: adaptive hypotheses should be based on our best knowledge of the phylogenetic relationships among extant and extinct species beginning from the relevant Last Common Ancestor (LCA), we may not assume that a trait's current function is identical with the function for which it was selected throughout its emergence, and it must be developmentally, biomechanically, or ecologically plausible that the novel trait could have emerged as a modification of an ancestral one (see, e.g.,Griffiths (1996); Pigliucci and Kaplan (2000).

Meeting the standards of good adaptationist reasoning is already especially challenging in the context of human cognitive evolution, most notably because evidence for the cognition and behavior of ancestral hominin species is generally less available compared to other explanatory contexts. This is a challenge for the historical sciences more generally and has motivated skeptical accounts (Turner, 2005; Lewontin, 1998). Historical scientists continually rise to this challenge and have developed methods for discovering and analyzing the available evidence (see, e.g., Wylie (2011)), Currie (2018, 2015), Currie and Killin (2019)). Still, the challenge for human cognitive evolution remains. Even compared to other targets in historical reconstruction, the hominin lineage is relatively sparse with respect to number of closely related species and number of fossil specimens (Kaplan, 2002). Sparse evidence, together with the adaptationist requirement that the timing and causes of some trait's emergence are plausible relative to the lineage's ancestral traits, puts cognitive evolutionary theorists in a particularly difficult epistemic position where they are liable to reach for intuition that, by its nature, is apt to masquerade as evidence.

The paper proceeds as follows: in 4.3-4.5, I discuss examples from Tomasello's, Stanford's, and Sterelny's hypotheses. In each, I show the operation of intuitions about the association between one set of capacities and another or the evolutionary sequence from one set of capacities to another. In 4.6, I connect my discussion to the related issues of the ethnographic analogy in paleoanthropology and anthropomorphism in comparative psychology. I discuss the insights we can gain from those issues as well as the distinct features of the issue I raise here. In 4.7, I summarize and conclude with a discussion of the outlook for responding to this issue in ongoing work on human cognitive evolution.

4.3 Thought experiments in Tomasello's Shared Intentionality Hypothesis

Michael Tomasello, in his Shared Intentionality Hypothesis (SIH), proposes three major stages in the evolution of human cognition: individual intentionality, shared intentionality, and collective intentionality (Tomasello, 2019, 2016, 2014). Because I will discuss an extended example from Tomasello's presentation of the SIH, a summary of these stages and their transitions will be useful:

1. *Individual intentionality* is a type of thinking in which an organism can represent experiences, simulate the effects of actions it might take related to those experiences, and take such simulations into account in determining its own behavior. The thinker can represent the behavior of another, but only as a relevant feature of the environment. An organism with individual intentionality may, e.g., exploit cues from a competitor's behavior as a relevant factor in determining whether to fight or flee, but it does not conceive of other agents as having intentions like its own (Tomasello 2014, 9).

2. *Joint intentionality,* the second stage and Tomasello's primary novel proposal, is a kind of thinking in which at most two partners can conceive of themselves as jointly holding the same intention, while also representing their own respective individual roles and perspectives (Tomasello 2014, 33).

3. *Collective intentionality* is the abstract, group-oriented thinking of modern humans.

> Modern humans can coordinate with any member of the group, as opposed to the transient pairs that possessed joint intentionality. Modern humans form 'objective' representations (from 'anyone's' perspective), make inferences based on reasons that 'anyone' should find compelling, and adhere to norms that would apply to any member of the group (Tomasello 2014, 81).

On Tomasello's view, modern humans alone have capacities for joint and collective intentionality but share capacities for individual intentionality with great apes. Therefore, individual intentionality is attributed to the LCA and forms the starting point for the SIH. Tomasello proposes that joint intentionality emerged about 400,000 years ago as an intermediate stage between individual and collective intentionality (Tomasello, 2014).[4]

Tomasello frequently reasons about the conditions in which ancestral hominins would have found themselves and the way that their cognitive traits would respond to these conditions to make the kinds of claims introduced earlier: either that some capacities arose together, such that the presence of one indicates the presence of another, or that one capacity would soon emerge as a consequence of the other. Below, I will introduce and analyze a thought experiment that Tomasello presents when he explains the emergence of the capacity to create novel combinations of conventional gestures. The thought experiment does not represent the entirety of the SIH or the only kind of evidence that Tomasello considers overall. The overarching hypothesis – that joint intentionality emerged from ancestral individual intentionality and gave rise to modern collective intentionality – is made up of many smaller-scale claims, only one of which is presented in the thought experiment.

In the thought experiment quoted below, Tomasello proposes a way that early humans might have acquired the capacity to form abstract linguistic constructions by creating novel combinations of communicative conventions – a uniquely human capacity on the way to modern human linguistic capacities and mindreading. Tomasello claims that abstract linguistic con-

[4]For a more comprehensive review see Peters (2016).

structions emerged under pressure from situations in which a recipient does not understand the intended meaning of a communicator's gesture (Tomasello 2014, 99). "Linguistic constructions begin with such simple item-based schemas but then are elaborated and made more abstract through discourse interactions," he writes, "The key aspect of the process for current purposes is the communicative pressure – the demand for sufficient information – coming from the recipient. This forces the communicator to make things explicit that he otherwise might have left implicit" (Tomasello 2014, 99). We need to be convinced first that when a communicator with a certain set of capacities is faced with a situation in which it would help to make intentions explicit, they would recognize this fact and forge novel connections among conventions with ease, and second, that once such a novel connection is formed, it is similarly easy enough to generalize and apply this pattern to other conventional gestures. Because the wording and presentation of the thought experiment are important to my argument, I quote it at length. It runs as follows:

> If we imagine early modern humans with a small inventory of single-unit (holophrastic) communicative conventions, along with general cognitive abilities for creating novel mental combinations (possessed by all apes), we can easily imagine them creating multiunit linguistic combinations. And so, for example, perhaps there was a communicative convention for requesting eating by moving the hand to the open mouth. And perhaps there was an unrelated communicative convention for requesting going foraging for berries (miming a picking motion). It would not take a genius, in a situation in which someone offered something unpalatable to eat, to gesture eating followed by berries. Then, given an existing ability for schematization (possessed by apes and early humans, but now applied to conventions), one could imagine this individual generalizing the conventional eating gesture to other conventional food gestures, much like human toddlers do as they first say things like "more juice," followed soon by "more milk," "more berries,"

and so forth in a "more X" pattern (Tomasello 2014, 99).

The inferences described in this thought experiment – first the creation of the new conventional phrase "eat berries," and second the generalization of that combination's form to other conventional gestures – must be plausible enough to conclude that inferences of this type were easy enough for early humans who have only holophrastic communicative conventions and a general ability to create novel mental combinations and schematizations in certain domains.

We are supposed to find it plausible that an early human could have strung together conventional representations of the concepts "eat" and "go foraging for berries" in response to being offered something unpalatable and mean, "I do not want to eat that, and I want to eat berries instead, so let's go forage for some berries to eat." Further, we should find it plausible that the same individual could generalize that innovation and apply it to other conventional gestures. But to convince us, Tomasello asks us to imagine an early human with capacities for holophrastic conventions and novel mental combinations and then remarks that "we can easily imagine" them beginning to use multiunit conventions in novel mental combinations. It "would not take a genius" to do so. That is, if an early human has the first two capacities, the third is *close enough* that they probably had that third capacity, too. To convince us that the imagined early human could easily enough apply schematization to conventions, he remarks that "one can imagine" that they would generalize the way that human children do.

How easy is it to generate multiunit linguistic conventions, given only holophrastic conventions and the capacity for novel mental combinations in other domains? And how easy is it to apply the ability to schematize in other domains to these new, multiunit conventions? When we make these judgments, we are assessing how easy it would have been for an early human to make the kind of connections that we routinely make *using* our capacities for abstract linguistic construction. As a modern human with the ability to innovate with multiunit conventions, it *is* obvious to us that appending the gesture for "eat" to "forage for berries" should modify the meaning just so, and it *is* obvious that one could do this across a variety of communicative contexts. But we are not in a position to evaluate how accessible such inferences would be given only those ancestral capacities and *without* employing the capacities that we already have.

The events described in the thought experiment may indeed seem plausible. And they are entirely possible. But it is the alleged ease with which the imaginary hominin would be able to make the inferences described that renders plausible the claim that abstract linguistic constructions would be achievable for hominins with these other capacities. In this thought experiment, Tomasello has not simply overlooked the need for empirical evidence beyond that which supports ascribing to the LCA capacities for single-unit constructions and novel mental combinations; he explicitly relies on the thought experiment. Surely we can imagine a hominin with the capacities Tomasello described, but we cannot then intuit what *else* they would be able to do. We do not have access to a cognitive experience without our own evolved capacities; our intuition is a poor judge of how much of a stretch some previously unmade inference would be to our ancestors. Such judgments of plausibility cannot be the basis for accepting the proposed causal sequence. It is, in principle, possible that the intuition could be replaced by appropriate empirical evidence. In that case, it would be crucial that the intuition be *replaced*. Because it is sourced from an unreliable intuition, it cannot serve as evidence, even in conjunction with other, appropriate evidence.

I have focused on this example in detail to illustrate the issue, but this is not an isolated instance. Tomasello gives similar thought experiments in at least two other instances in *A Natural History of Human Thinking* in which his argument is presented in the form of a request for the reader to imagine situations in evolutionary history (Tomasello 2014, 108–9). Still, one might think that this is a very specific inference in only one causal step in the SIH. I will discuss another example, this time from Tomasello's 2016 book, *A Natural History of*

Human Morality, to show that this issue arises in the context of another major explanatory target in the evolution of human cognition where we find strong, likely evolved, intuitions: human moral psychology. Tomasello (2016) argues that the evolution of moral psychology was part of the two-step cognitive evolution described in the SIH. A precursory morality first emerged alongside joint intentionality between pairs of partners, which was then scaled up to the group level. Throughout the emergence of joint intentionality, social norms emerged with the development of the cognitive capacities for jointly intentional collaborative activities – both driven by a selective pressure that pushed ancestral hominins toward obligate foraging (Tomasello, 2016). He claims that hominins with capacities for joint intentionality who were engaged in such collaborative activity would develop a conception *of agent-neutral roles* in that activity, which entails what he calls a "second-personal morality." Second-personal morality is a sense of moral self-other equivalence, i.e., because I could just as well serve in my partner's role in the activity, I see us as having a kind of equal standing in its execution. This recognition of self-other equivalence, he argues, comes along with a sense of mutual deservingness and respect of each other's equal status (Tomasello 2016, 60). As joint intentionality was gradually scaled up to collective intentionality, so the story goes, this sense of mutual dessert was amplified to a sense of responsibility to a group, imposed externally on all members equally.

But the association of self-other equivalence with the felt sense of mutual deservingness and respect turns out to be another case of Tomasello inferring the evolutionary association of traits based on their intuitive association in our own modern human experience. In a review of *A Natural History of Human Morality,* P. Kyle Stanford points out that on Tomasello's picture, the distinctly moral part of human moral psychology, supposedly explained by the SIH, is actually presupposed (Stanford, 2017a). Tomasello means to explain why modern humans feel a sense of obligation to adhere to moral norms. Moral norms differ from what Stanford calls "instrumental" norms. Partners could adhere to a norm simply because they are somehow motivated to behave toward each other in a particular way, e.g., by a mere preference for doing so (Stanford 2017a, 372). Stanford argues that the problem of jointly coordinating activity requires only instrumental norms. If this normative motivation relies on a kind of self-other equivalence, it need only be an instrumental self-other equivalence, not one of moral obligation. It is possible to recognize another agent as essential to some goal, and to recognize that to achieve that goal one should behave toward that agent in a way that modern humans would view as fair, all without feeling a sense of external obligation, deservingness, or respect. And without this particular sense, there would be no second-personal *morality* to scale up to the group level.

Stanford is rightly objecting to Tomasello's idea that moral obligation comes along with jointly intentional cooperation between partners. But this is not an oversight of Tomasello's. He sees self-other equivalence is a kind of "spandrel" – a structural support *for* the sense of moral obligation between dyads (Tomasello 2016, 80).[5] Tomasello recognizes that some collaborative activities between partners require only what he calls "strategic" expectations of one's partners (what Stanford might call "instrumental") but argues that trusting these expectations is risky, and that a mechanism for mitigating that risk must have been part of the evolution of jointly intentional cooperation. "What we need to take the risk, then," he writes, "is for each of us to trust one another more deeply, in a more committed way. What we need is for each of us to feel that we truly *ought* to follow through on our collaboration, that we truly *owe* it to one another" (Tomasello 2016, 64). A sense of mutual respect or deservingness *requires* an associated sense of obligation in order to reliably produce jointly intentional cooperative partners.

But why should it be a felt sense of obligation that mitigates the risk of strategic interaction, when non-moral motivations could produce the same behaviors? This particular connection, for Tomasello, is not so much an empirical question, but a conceptual matter

[5] A spandrel, in the sense articulated by Gould and Lewontin, is a structural feature of an organism that appears adaptive in its own right but is in fact only a consequence of its relationship to another adaptive trait, though it may later acquire its own adaptive function (Gould and Lewontin, 1979).

of what it means to feel obligation, which he supports with an appeal to Rousseau, writing that the "only possible source for a personal sense of 'ought' of this genuine kind is my identification with and deference to a larger (even idealized) body of which I myself am a part" (Tomasello 2016, 64). But it is this connection between the development of *some* sense of self-other equivalence, along with joint commitments, and a sense of *externally imposed obligation to one's* partner, that explains the emergence of distinctively moral psychology. As Tomasello writes, "The claim is not just that this was a modest first step on the way to a fully modern human morality, but rather, that this was the decisive moral step that bequeathed to modern human morality all of its most essential and distinctive elements" (Tomasello 2016, 78). Without this, Tomasello has not explained the emergence of humans' uniquely moral psychology. The fact that certain norms are experienced as externally imposed in virtue of one's membership in a group *just is* the morality of collectively intentional agents that needs empirically supported explanation.

Though in this case it does not come in the form of a thought experiment with an invitation for the reader to imagine,[6] Tomasello explicitly relies on the notion that the recognition of self-other equivalence concerning roles in a joint activity, in combination with contexts of equal bargaining power and the free-rider problem, "led to" or would cause early humans to "c[o]me to feel a sense of" a *moral* kind of self-other equivalence in which partners are mutually deserving of what *we* would call fair treatment (Tomasello 2016, 80). Without this implication between the two, there is no reason to expect that the external moral obligation of collectively intentional agents would emerge from the partnerships of jointly intentional agents.

But why has Tomasello committed to this implication at all? What is it about the recognition of (not yet moral) self-other equivalence and the contexts described that makes the emergence of a genuinely moral psychology plausible? This is, I argue, the same issue that we saw

[6]Though Tomasello does often use phrases like "we may picture. . . " or "The early humans we are picturing here. . . " when characterizing the early humans in question (Tomasello 2016, 65–68, 89).

in Tomasello's earlier thought experiment. Modern humans are already equipped with the moral psychology that prompts us to experience some norms as obligatory. For us, violations of norms in a joint activity are not isolated from modern human moral psychology more broadly, and we do not know what it is like to be a hominin with only joint intentionality. We use our fully modern, collectively-intentional moral psychology across the board – jointly cooperative dyads included. Suppose a stranger and I happen to be shopping in the same grocery store aisle, and we notice that there are two containers of chocolate-covered almonds remaining on the top shelf. We would each like to purchase some chocolate-covered almonds. A rolling ladder is nearby, and we agree that I will hold the ladder while my newfound partner climbs up and retrieves the almonds. We have formed a joint goal, unachievable by either of us alone, for which we are instrumentally equivalent (either of us could have held the ladder while the other climbed). But once my partner is safely on the ground, they take both containers and rush away without a word. We were engaged in a prototypical joint-intentional collaboration, which was purely instrumental. We had no social relationship nor investment in each other beyond the achievement of our joint goal. And yet it would be a sense of *moral* violation that I feel at their taking all the almonds. I, a modern human, would of course *feel wronged* by my partner because they took all the spoils of our joint effort. *For us*, instrumental self-other equivalence in a joint activity *does* result in a sense of moral obligation to one another. But this does not mean that early humans who were engaged in collaboration prior to the evolution of that sense of obligation would develop such a sense simply by recognizing instrumental self-other equivalence. It only seems so plausible that they would because for modern humans who *already have* our evolved moral psychology, even instrumental self-other equivalence triggers a moral response. Tomasello has – implicitly, this time – taken the perspective of early humans interacting *prior to the evolution of the capacity in question* and used his experience as a modern human to reach conclusions about the course of evolutionary history.

4.3.1 Minimum-capacity inference without artifacts?

Tomasello's inferences in these two examples share some features with what philosophers have called a *minimum-capacity inference* (Currie et al., 2024; Killin and Pain, 2021; Currie and Killin, 2019). Cognitive archaeologists and paleoanthropologists often make minimum-capacity inferences about material traces, especially artifacts like tools, and the abilities of the creatures that produced them. When making a minimum-capacity inference, an archaeologist first asks, what are the minimum necessary competences for an agent to produce that artifact? For example, what cognitive abilities are *required* for an agent to produce Acheulean stone tools? The archaeologist uses middle-range theory (e.g., about human cognition) and methods (e.g., studying modern human artisans) to determine what sorts of cognitive capacities were minimally required to produce the tools, e.g. forward planning and mental time-travel. Then, on the basis that these abilities are strictly required to produce the object, the archaeologist infers that the maker indeed had that cognitive capacity (Currie and Killin 2019, 268). Tomasello's SIH is a historical reconstruction, so understanding the relationship between these inferences and a core epistemic strategy of the historical sciences could offer insights or improvements.

The key insight that licenses minimum-capacity inference is that evidence for the production of stone tools (i.e., the preserved tool itself) is also evidence that its maker had the capacity for, e.g., mental time travel, provided that separate evidence supports the connection between the artifact and the capacities required to produce it. Compare this with Tomasello's inferences about the capacity for early humans to produce multiunit communicative conventions: if we have evidence that a hominin could produce holophrastic conventions and novel mental combinations, we also have evidence that they could produce multiunit conventions. Recognizing this similarity Tomasello's inferences makes clear what sort of evidence would work for Tomasello: good, independent evidence that the capacities in question are associated in such a way that they would evolve together. This is a strategy for supporting minimum-capacity inferences, proposed by Killin and Pain (2023) – developing better theories of the biological development and operation of modern human cognition can improve our understanding of the end point of the narrative.

However, Tomasello's inferences about novel multiunit conventions differ from minimum-capacity inferences in two interrelated ways: what sorts of things are connected to facilitate the inference, and whether that connection is one of minimal necessity. To make a minimal-capacity inference, archaeologists connect artifacts with cognitive capacities; in Tomasello's thought experiments, he connects cognitive capacities with other cognitive capacities. What difference does this make to the evidence required? Archaeologists seek justification that certain cognitive capacities are minimally necessary for producing particular artifacts; Tomasello needs evidence that two cognitive capacities are not only associated in modern humans, but that they emerged together in evolutionary history. Showing that the capacities are in some sense interdependent in modern humans would not give the right kind of necessity relationship for a minimum-capacity inference, as traits that depend in some way upon others in modern humans may or may not have originally emerged in evolutionary history with such dependence in place.[7] It is possible that cognitive traits emerge with such dependence, but if they did not (or without evidence to suggest they did) minimum-capacity inference is not possible here.

There is a further point of contrast with standard minimum-capacity inferences in Tomasello's argument about moral psychology. In that case, it is unclear what sort of change he is proposing for early humans to be "led to" or to "c[o]me to feel a sense of" genuinely moral obligation – whether these are the very same individuals coming to feel a new sense of moral obligation in their lifetime, the sense of moral obligation emerges over close generations, or the sense of moral obligation is itself another change in cognitive architecture. If either of the former

[7]This has been an important problem in showing traits to be adaptations, have particular functions, or most recently, to be unified traits in the first place – see e.g., Meneganzin and Currie (2025) Meneganzin et al. (2024), DiFrisco and Ramsey (2023), and Allen (2002).

options is what Tomasello intends, the issue is of the very same type as described in the earlier example from *A Natural History of Human Thinking.* If it is the latter, then the issue is slightly different: rather than an assumption of what would have been an easy or reasonable new inference for an individual, the claim is about what would have been *evolutionarily* 'easy' for the lineage to achieve – explaining the new cognitive capacity by its evolutionary 'closeness,' where that closeness is supported by intuition. This is quite different from a minimum-capacity inference, because we are no longer reasoning about what capacities an early human had *at one point in time*, but the plausible evolutionary connection, in sequence, between one and the other. This would be a claim about evolvability and thus the issue may be more general than making assumptions about what an imaginary individual could achieve.

4.4 Stanford's hypothesis for the emergence of moral psychology

Tomasello's thought experiments are illustrative because in them, he makes explicit appeals to intuition in his assessment of whether a certain kind of inference was within the reasonable range of an ancestral hominin. But theorists also make less explicit use of this kind of evidence. Kyle Stanford (2017b) defends an adaptive hypothesis for the evolution of human moral psychology. Stanford first argues that the appropriate target for this evolutionary explanation is our sense that moral norms are *external.* That is to say, human moral psychology is characterized by the sense that we are motivated to follow certain norms not because we simply prefer to, but by something stronger – a sense of obligation that exists whether or not we prefer to follow that norm. Humans experience prototypically *moral* norms as external, such as those forbidding murder, but the capacity for externalization is flexible. Humans also come to externalize (and thereby moralize) other norms, resulting in cross-cultural variation regarding which particular norms are moral rather than conventional. Some norms may be more or less universally regarded as conventional, such as driving on the right-hand side of the road, and others may be so regarded as moral, such as murder in cold blood. But other norms can fall in-between extremes and be moralized to a greater or lesser extent depending on one's culture and context, e.g., burning one's national flag.

Stanford proposes that moral externalization is a cognitive capacity that emerged as an adaptation for ensuring correlated interaction among cooperative partners in ancestral hominin populations. By externalizing certain cooperative norms, early humans could be motivated to adhere to a norm themselves as well as observe the behavior of others and use *their* adherence to the norm to evaluate them as potential cooperative partners. This would protect increasingly cooperative hominins from exploitation, opening the possibility for developing the hyper-cooperative behaviors that we now observe in human populations. Hominins that were "prepared to adopt" increasingly cooperative norms would preferentially interact with one another, even as new norms were developed or extended (Stanford 2017b, 10). Critics of Stanford supply several alternative ways that the hominin lineage might have protected correlated interaction among cooperators *without* norm externalization (Stanford 2017b, 12–26). They argue that moral externalization is not the necessary solution to the problem of correlating interaction. Stanford replies that such suggestions miss the point – he does not mean to argue that moral externalization evolved because it *had* to. Instead, he argues that on an entirely empirical basis, we ought to accept that moral externalization *was in fact* the solution to the problem of correlated interaction that occurred, even though others were possible.

As Lee (2024) emphasizes, Stanford's empirical argument relies heavily on the connection between adhering to a norm and evaluating others on the basis of their adherence as evidenced by e.g., Skitka et al. (2005). Skitka et al. present evidence of a correlation between a person's own commitment to an externalized norm and the extent to which one excludes others for violating it. While suggestive, Stanford's use of this evidence from Skitka et al. in support of the hypothesis shows a similar problem to the one show in Tomasello's work. How? And what about the rest of Stanford's impressive compilation of evidence?

Stanford's empirical evidence is drawn from human psychology,[8] comparative psychology,[9] and ethnography.[10] This evidence establishes his three foundational claims: (1) human norms are situated on a scale of externalization, ranging from conventional to near-uniformly externalized, along which any particular norm could be found, (2) humans are spontaneously and flexibly hyper-cooperative, and such behaviors appear reliably at certain points along the developmental timeline, and (3) specifically supported by Skitka, Bauman, and Sargis (2005) and Goodwin and Darley (2012), that the degree to which people are comfortable socially engaging with others is directly related to both the degree to which one takes a shared norm to be objective and the degree to which one sees oneself as committed to and motivated by it (Stanford 2017b, 8). Note, however, that all three claims above apply to *modern* humans only. The empirical evidence provided certainly makes (1) and (2) seem likely, and (3) at least highly suggestive. But together with their associated evidence, (1) tells us what sorts of decisions, behaviors, or judgments are motivated by the degree of peers' adherence to norms along the scale of externalization *in modern humans,* (2) tells us that *modern humans* are strikingly different from chimpanzees in our cooperative tendencies, and (3) suggests that *modern human social preference* is systematically related to degree of externalization as well as one's own degree of commitment. Surely both (1) and (2) suggest that externalization and hyper-cooperation are important to the evolution of human cognition and warrant investigation into their origins, and (3) suggests that the two features of human cognition and behavior are related in modern humans. However, Stanford means for the empirical

[8]E.g., Sarkissian (2016), Skitka et al. (2005), Wright et al. (2014), Goodwin and Darley (2012, 2010, 2008), among a host of others.

[9]Including, but not limited to, studies from Tomasello, Warneken, Melis, and colleagues (Melis et al., 2013, 2011; Warneken et al., 2011; Melis et al., 2010; Melis and Semmann, 2010; Tomasello et al., 2007; Warneken and Tomasello, 2007).

[10]E.g., Baumard et al. (2013), Guala (2012).

evidence to show more than this. It is also supposed to show that hyper-cooperation and moralized personal motivation evolved *together* via the emergence of a capacity for moral externalization, in response to selective pressures for correlating interaction with cooperators and excluding defectors. Modern humans derive their motivation to adhere to a norm and their social preferences with respect to others' adherence from our capacity to externalize. It is the sense that the norm is external that, for us, *does both.* They follow so closely from the very idea of an externalized norm that is difficult for us to even imagine an externalized norm that does not impact both of these things. But this difficulty is an effect of their *current* systematic relationship, which need not have emerged together with the capacity to externalize. If, say, this systematic relationship developed later, then the selective advantage that Stanford proposes for the origin of moral externalization – that it regulates hominin cooperative behaviors *and* choice of cooperative partners – no longer holds.

Stanford's claim that we have "compelling empirical evidence that externalization is in fact how humans came to achieve the extraordinary degree and precision of correlated interaction that we do" (Stanford 2017b, 34) represents a similar issue to the one we saw in Tomasello's SIH. We must believe that those inferences and motivations that externalization grants us would are related such that they would emerge together in hominin populations gradually developing a capacity for externalization:

> The key to resolving this further puzzle is to recognize that although we might more simply and easily use subjective preferences or desires to motivate either our own conformity to a given set of norms or standards of behavior, or our social exclusion of those who do not similarly conform, experiencing moral motivation as externally imposed on both ourselves and others simultaneously is nonetheless an extremely effective and efficient way to ensure that these otherwise distinct motivations arise together and remain systematically connected to one another even as particular norms themselves come to be modified, applied, and extended

in new ways and into new circumstances, and even as entirely new norms come to be adopted (or come to be moralized) in the first place (Stanford 2017b, 8).

The implicit premise at work here is that judgments about the suitability of potential cooperators in relation to their adherence to norms *along with* one's own motivation to adhere to those norms would have been produced together by an emerging sense of external norms in ancestral hominins who previously possessed nothing of the sort. But it only appears to be a plausible evolutionary step because the relationships among our own adherence to norms, the adherence of others, and the sense of externality are obvious to us – a sense conferred only by their association in our modern cognition.

4.5 Feedback loops in Sterelny's apprentice model

Tomasello's thought experiments explicitly appeal to intuition in place of evidence, while Stanford's reliance on modern human data that misses an important connection appears to be a kind of oversight. In this section, I will introduce another example that contrasts with the previous examples in that it shows an intentional appeal to intuition that modifies the explanatory target. In another prominent hypothesis for the evolution of uniquely human cognition, Kim Sterelny presents an account centered around the operation of two feedback loops. One is between expertise, social learning, and life history, while the other is between expertise, individual adaptations for social learning, and organized learning environments (Sterelny 2012, 29). These two feedback loops describe the causal relationships that allowed the hominin lineage to accumulate social knowledge and respond adaptively to that accumulation, eventually transitioning to cultural knowledge and modern forms of human cognition.

In brief, Sterelny's hypothesis is that as hominins developed technical competence (e.g., toolmaking, or food and resource manipulation) they inadvertently enriched their environments so as to gradually increase the structured learning of other hominins, including their offspring. This initially accidental enrichment contributed to accumulation of competences across generations, providing selective pressures favoring longer childhood and improved social learning. These factors are interdependent, giving rise to the feedback loop that, on Sterelny's view, gradually snowballs to produce modern human cultural learning and uniquely human cognition.

Of particular importance in Sterelny's account is *apprentice learning,* a kind of social learning that Sterelny argues could have developed early in hominin history. In the general format of apprentice learning, an apprentice learns from an expert by observing the expert's actions, taking on small, relevant, informative tasks delegated by the expert, as well as learns from the structured environment that the expert maintains. He argues that in principle, apprentice learning can be constructed incrementally, in just the way that would make it attainable for gradualist evolution: intergenerational social knowledge can accumulate accidentally, without preexisting capacities for intentional teaching or learning. In addition to apprentice learning, the features of technology, tool use, and other material culture of early humans are of crucial importance. The interdependence among these factors forms the feedback loops that are characteristic of Sterelny's account (Sterelny 2012, 23–43).

Sterelny lays out these relationships early in *The Evolved Apprentice*, and throughout the rest of the book invokes archeological and anthropological evidence to support the causal relationships among these factors, as well as evidence of specific events supporting the idea that these relationships are responsible for the course of human cognitive evolution. The bulk of Sterelny's account concerns the operation of these feedback loops once they are set in motion. For my purposes here, we can grant that much of the evidence for the operation of the feedback loops as described is convincing. I am here concerned with an argument Sterelny presents early on in his hypothesis in order to make plausible the initial emergence of those feedback loops. This plausibility argument, then, plays a different role than the inferences discussed in the previous examples, but its role is still evidential.

The initial change that Sterelny proposes is that hominin offspring became curious and parents became tolerant – novelties that Sterelny calls "simple motivational changes" from the ancestral state (Sterelny 2012, 33). What kind of change triggered hominin offspring to find the technical activities of their parents salient enough to hang around and pay attention to them? And what triggered the required initial increase in parent tolerance of this offspring involvement? Sterelny gives an argument that relies on a combination of the ease of apprentice learning as well as the likelihood of accidental innovations occurring over time.

Any tool use or resource modification (in any species) modifies the environment and thus has the potential to, as a by-product, facilitate social learning in an organism cognitively equipped to exploit the modified environment. Sterelny argues that such opportunistic social learning was probably important in the development of material culture itself for hominins. This is, he says, "not idle speculation" (Sterelny 2012, 32). Rather, "the first traditions of using stone technologies may well have been established by such an epistemic side effect of innovation" (Sterelny 2012, 32). His proposal is that innovations themselves would be chancy and sporadic, and the conditions in which they occur may make them more likely to catch on – amounting to a kind of 'frozen accident' theory for the evolution of material culture. And if humans spend a lot of time in the same spaces, the probability of such accidents increases. The evidence for this idea consists in claims about the complexity of tools made by modern humans, and the by-product environment such activity generates (Sterelny 2012, 32–33).

This argument does not only serve the role of *illustrating* how technical competence created enriched environments; such frozen accidents are crucial to Sterelny's hypothesis. Because tool use or resource manipulation has this feature of modifying the nearby environment in any species that stays relatively local, but no other species has developed the same extensive, accumulated material culture (nor the cognitive capacities to produce such cultures), we ideally would want to know why hominins would have come to take advantage of such information-rich environments. For the incremental construction of apprentice learning, it is required that these motivational changes occur first, and Sterelny asks us to take them on assumption: "on the assumption that young hominins spent a good deal of time with their elders, and on the assumption that those elders were tolerant of children's presence, curiosity, and experimentation, the basic hybrid learning model applies to early hominins" (Sterelny 2012, 37–38). Given this tolerance and curiosity, it was then only a matter of time before the hominin lineage began a feedback loop of increasingly sophisticated social learning. The plausibility that such a feedback loop would emerge in such an environment rests on the assumption that motivations for both parental tolerance and offspring curiosity and experimentation had already emerged. Sterelny is relying on intuition to motivate the idea that the relevant changes in the salience of adult behaviors and parental tolerance for the involvement of children are "*simple* motivational changes" – that is, that it would be *easy* or simple enough that we can take it on assumption that these two things emerged together in ancestral hominin populations. Intuition motivates the idea that such a feedback loop could get started easily enough. Then, Sterelny can invoke evidence for the operation of these feedback loops without necessarily needing to provide evidence of just how they began.

I take this to be neither error nor oversight on Sterelny's part. Instead, it constitutes a shift in the explanatory target that effectively concedes one part of the overarching question, "why us and not them?" If Sterelny is right, then we know that at some point such feedback loops *did* get started, but we would still not truly know *how.* In fact, it is possible that we may never know how, depending on whether or not specific evidence emerges to replace intuition in Sterelny's argument. This is significant; the way that these unusual processes got off the ground seems to be a substantial part of the explanatory target. We might contrast Sterelny's change in explanatory target to the case of Behavioral Modernity in paleoanthropology. Meneganzin and Currie (2022) show that a canonical explanatory target for paleoanthropologists, Behavioral Modernity (an alleged revolution or period of rapid change during which all the characteristic behaviors of modern humans appeared in the lineage and transformed the resulting species into fully modern descendent humans) turned out to be an ill-conceived target. The changes at issue were incremental; there was no revolution to look for. In one sense, Sterelny's hypothesis embraces a similar dissolution – there is no moment (or single characteristic capacity) that sets off distinctively modern human cognition from the previous forms. But it is importantly different: although Sterelny proposes that it was a series of feedback loops rather than the emergence of a particular capacity, it remains that those feedback loops themselves have a causal history. They have a starting point – a set of conditions that we would want evidence for as those which produced it. Such evidence is probably diffuse and exceedingly difficult to identify – perhaps we ought not to expect a theorist to provide any more than what seems like a plausible story. That may be entirely reasonable, but it is important to recognize that this would be an evidential standard based on the expected availability of evidence, not based on an empirically motivated realization that the original target does not properly exist after all, as in the case of Behavioral Modernity.

Perhaps theorists ought not care that Sterelny cannot account for the very beginning of the narrative – maybe if the evidence for the operation of such a feedback loop is good, then that is good enough for us. Maybe we ought not to worry about the specifics of the startup, since Sterelny's primary claim is that no one feature of the initial conditions was a 'magic bullet' in setting off the process.[11] But this is no small change to the explanatory target – recall that in Sarah Hrdy's narrative in *Mothers and Others*, the fundamental question was laid out as "why us and not them?" (Hrdy 2009, 33–37). Why hominins and no other lineages? More specifically, why humans and not our closest evolutionary relatives? Sterelny cannot answer what seems to be a rather important part of the question – what makes the difference between the hominin lineage and the chimpanzee and bonobo lineage is that at some point hominins did experience the required motivational changes that enabled apprentice learning. But we

[11]See Downes (2013) for further discussion on magic bullets.

do not know *why* that happened for us and not for chimpanzees. If taking such motivational changes on assumption is ultimately the best we can get, we may have to be epistemically satisfied with an explanation in the form of a historical reconstruction of lower certainty than we would like. Perhaps this is not a terrible state to end up in, especially as Sterelny is probably right that many small changes in ancestral hominin life produced the conditions for coevolutionary feedback loops. But this shift in explanatory target is non-trivial, and its impact on the confidence we could have in the best, practically possible hypothesis for human cognitive evolution is worth recognizing. From an explanatory perspective, recognizing that the initial conditions for feedback loops are many and diffuse and that there was no magic bullet does not make the particulars of those conditions less important, but it does suggest they may be harder to specify. It represents a recognition of a limitation of evidence, and a practical adjustment of the type of evidence that theorists might realistically expect of one another. This is a change to what we might care about knowing, but not what facts about the hominin lineage could, in principle, be known.

4.6 Analogies and anthropomorphism

The issue that I have identified here is closely related to two other issues that have arisen in similar contexts – one is the use of ethnographic evidence from modern foraging communities to make inferences about ancestral populations, and the other is the familiar issue of anthropomorphism in the sciences of behavior and cognition across human and other animals.

The use of observational data from contemporary foraging societies to make inferences about ancestral hominin populations, sometimes called the "ethnographic analogy," has been controversial in paleoanthropology (Killin and Pain, 2023; Sterelny, 2022; Wylie, 1985). However, several philosophers have shown that making such analogical inferences is not inherently problematic. Wylie (1985) shows that in archaeology and elsewhere, analogy is not an inherently faulty reasoning pattern, and that analogies can be constrained and supported by relevant theories as well as rigorously evaluated. Currie (2016) shows that one cannot evaluate the general form of the ethnographic analogy; rather, the use of data from contemporary foragers (or modern humans in general) is sometimes appropriate and sometimes not. It depends on what other evidence is available and whether or not the behavior or practice in question is literally conserved over evolutionary history. Sterelny argues, similarly, that when ethnographic data can be legitimately generalized, it can guide inferences about past populations (Sterelny, 2022). Each of these accounts maintains that analogies must be appropriately constrained in order to be legitimate. Killin and Pain discuss one such constraint when they show that limitations on our current understanding of human psychological variation pose problems both for the use of ethnographic data and for minimum-capacity inferences. Our understanding of modern human cognition serves as middle-range theory, mediating projections onto ancestral hominins. As such the permissibility of those projections depends on qualities of that middle-range theory, e.g., sample diversity (Killin and Pain, 2023).

In the inferences that I am concerned with in this paper, whether or not our intuitions about cognitive ease are actually clues to their evolutionary relationship is not even on the table. Evolutionarily biased intuitions about relationships among cognitive capacities cannot be constrained; they can only be replaced with alternative evidence. While the use of some kinds of modern human evidence may stand or fall based on the particulars of the context and the explanatory target, intuitions about the ease of inferences or connectedness of cognitive capacities can never serve as evidence. The fact that it feels convincing is part of its evolved nature, not because of a systematic relationship with the truth.

The issue I have identified here is also similar to a familiar kind of anthropomorphism – we are apt to see human-like features (or abilities) in animals or objects that do not or even cannot have such features. Dacey (2017) considers this tendency to be a kind of cognitive bias that manifests primarily in our assumptions about the operation of non-human systems.[12] This kind of anthropomorphism has been discussed in the literature on comparative psychology, and theorists often take a cautious view. One example comes from primate behavioral scientists, who are often particularly interested in whether, or to what extent, chimpanzees have a theory of mind. The history of investigations into this question forms a cycle of making and correcting assumptions that turned out to be inappropriately anthropomorphic. I will use this example to help illustrate the similarities between the issue I raise here and this kind of anthropomorphism, as well as the points of departure that help illustrate why the context of early hominin cognition is especially difficult compared to other areas where scientists contend with anthropomorphic cognitive bias.

Daniel Povinelli and his research team performed an extensive series of experiments that appeared to demonstrate that chimpanzees do not have cognitive capacities for a theory of mind (Povinelli, 2000). Their most convincing empirical support came from object-choice experiments designed to use chimpanzees' gaze-following behavior to test whether they take the perspective of other agents into account when determining their own behavior. In one version of the object-choice experiment, chimpanzees and three-year-old children were taught to search for a hidden treat under one of two cups. In each trial, subjects were kept ignorant of the treat's location, but an experimenter, facing the subject with the cups between them, either (1) turned their head and looked *at* the correct cup or (2) turned their head and looked *above* the correct cup. Children selected the correct cup when the experimenter looked *at* it, but chose randomly when the experimenter looked *above* it, suggesting that they use the experimenter's gaze as informative of the treat's location. Chimpanzees, in contrast, show no difference between the two conditions (although they do eventually associate the direction of the experimenter's head with the treat's location). Povinelli took these results, along with similar results in other versions of the experiment, to indicate that chimpanzees have no theory of mind. He argues that these experiments reveal the anthropomorphic assumptions

[12] I thank an anonymous reviewer for pointing me to this reference.

underlying a popular competing framework – that of David Premack and Guy Woodruff (1978). Premack and Woodruff had argued that chimpanzees *do* have a theory of mind much like that of humans, but their argument was based primarily on an analogy they drew between chimpanzee and human behavior.

In a later set of experiments, Brian Hare, in collaboration with Tomasello, point out that Povinelli's object-choice experiments place chimpanzees in cooperative contexts, whereas competitive contexts are more representative of the cognitive challenges typically faced by chimpanzees in nature (Hare and Tomasello, 2004; Call et al., 2004). They note that the results from Povinelli appear to be in tension with observational studies, like those of Goodall (2000) and de Waal (1982), in which chimpanzees had been observed to refrain from retrieving available food until groupmates had left the area, and to use their hands to obscure signals like facial expressions from groupmates. Hare and Tomasello argue that Povinelli had implicitly assumed that a chimpanzee would understand an experimenter's intent to help it obtain food – an action completely alien to a chimpanzee who is much more inclined to *compete* for that resource. Chimpanzees do not (unless by accident) indicate the location of food to others when they could very well obtain it themselves, so why should we expect a chimpanzee to understand that a human means to do so? Hare and Tomasello's team instead test chimpanzees' ability to represent and utilize the mental states of known competitors – dominant in-group chimpanzees. They conclude that chimpanzees demonstrate a limited ability to change their behavior depending on whether or not their competitor can see a piece of food (Hare and Tomasello, 2004; Call et al., 2004; Hare et al., 2000). Given the choice between a piece of food that was visible to both the subject and their dominant competitor versus a piece that was visible only to the subject (hidden outside the competitor's visual range), subjects preferred to retrieve the piece unseen by their competitor.

Povinelli and his team find Premack and Woodruff's theory to be inappropriately anthropomorphic – behavioral similarities need not indicate similarities in underlying cognition.

Hare and Tomasello see Povinelli's experimental design as similarly flawed, though Povinelli and Vonk had already discussed this issue with the experimental paradigm they previously employed (Povinelli and Vonk, 2004, 2003). Povinelli and Vonk explicitly attribute this problem to the fact that it is so easy to rely on our own theory of mind to intuit what sorts of behaviors can only be produced by reasoning about mental states.

In my examples above, theorists (including Tomasello himself) make a similar error, this time with imaginary ancestral hominins, rather than living chimpanzees. The example of chimpanzee theory of mind helps to draw one similarity and two points of contrast between the chimpanzee case and the imaginary ancestral hominin case. First, this kind of projection is evidently difficult to avoid for chimpanzees and ancestral hominins alike, even for an expert highly attuned to the problem. Second, the chimpanzee theory of mind example illustrates by comparison the increased difficulty of identifying and correcting this kind of error in reasoning about imaginary ancestors. In the chimpanzee case, live autonomous research subjects were a crucial factor in probing the potential of chimpanzee cognition, but theorists constructing thought experiments on imaginary ancestral hominins just cannot have such subjects. When we reason about the perspectives, thoughts, or potential inferences of ancestral hominins, the experimental subjects are under our control. The lack of autonomous research subjects, which is so helpful in the chimpanzee case, only increases the challenge.

There is also a fundamental asymmetry between the kinds of assumptions underlying anthropomorphic projection onto species other than modern humans compared to projection onto extinct ancestors whose capacities modern humans inherited: other species might well produce the same functional outcome as humans, but in some non-modern-human way. In these cases, anthropomorphic biases can lead to 'false negatives' or 'false positives.' Meneganzin and Killin have shown that a bias toward modern human aesthetic capacities has led archaeologists to a kind of false negative; they have missed important evidence supporting a uniquely *Neanderthal* aesthetic sense (Meneganzin and Killin 2024).[13] We might make a false positive anthropomorphic error when we look at an octopus and suspect it is phenomenally conscious in precisely the way that humans are. But it could be the case that Neaderthals had an aesthetic sense and that octopuses are phenomenally conscious,[14] in their own ways and by their own independently evolved mechanisms. In both examples, we reached an inappropriately anthropomorphic conclusion.[15] But in these cases, we are interested in whether a species has some trait construed as a general *trait*, not the uniquely human cognitive capacity. Prima facie, we were never supposed to assume that that Neanderthals did not have some kind of aesthetic sense, or that octopuses are not in some way phenomenally conscious. And so even when thinking about the cognition of non-human animals, we ideally refrain from assuming they *do not* have the same kind of capacity as humans, because we allow that they *may* have such a capacity (or an extremely similar one) undergirded by some non-human mechanism. In contrast, when we imagine ancestral hominins at the time of the emergence of a novel cognitive capacity, we introduce a strong constraint: since we are imagining hominin species prior to the evolution of the capacity, we stipulate that those hominins *do not* have some way of producing the functional outcome at hand.[16] Because if the ancestral hominin *could* already achieve that functional outcome in some other way, we would no longer have a question about its emergence.

[13]Meneganzin and Killin's argument also serves the important purpose of showing that anthropomorphic bias can lead to *underestimating* other species' capacities, while most accounts are concerned with *overestimating* them.

[14]As many scientists and philosophers believe that they are.

[15]Though some scholars prefer the term "Sapiens-centric" rather than "anthropomorphic" to describe the misplaced assumptions about Neanderthals.

[16]Meneganzin and Killin are interested in the project of explaining the origins of interesting capacities *as generalized capacities*, e.g., aesthetic sense *of some kind*. That is, the target trait may be an ancestral aesthetic sense, shared with LCA of Sapiens and Neanderthals, which then underwent different changes throughout the divergent lineages. They therefore might argue that, in the broader aims of archaeology, the interest in explaining the unique capacities of modern humans (which I have said is the explanatory target in each of Tomasello, Stanford, and Sterelny's hypotheses) is itself a source of error for this ultimately more fruitful project of explaining the evolutionary history of interesting capacities, rather than uniquely modern human capacities.

4.7 Conclusion & outlook

I have argued that theorists have appealed to evolutionarily biased intuitions in their reconstructions of human cognitive evolution, either explicitly or implicitly. These appeals have come in the form of straightforward use of intuition as evidence, as in Tomasello's thought experiments, as an implicit connection between modern human psychology and evolutionary association, as in Stanford's hypothesis, and as assumptions about the evolutionary availability of certain 'startup' capacities, as in Sterelny's hypothesis. Each example features, in its own way, the question of what an ancestral hominin could think, infer, or do in a given context – that is, what would imaginary ancestors do? I have shown that claims about the association or evolutionary sequence of cognitive capacities should not be evidenced by such intuitions, since the reason they feel convincing to us is influenced by their evolutionary history. I will conclude by considering the outlook for reconstructing human cognitive evolution in light of the challenge I have described. Are theorists likely to encounter more situations in which this kind of intuition might seem convincing? And should we expect it to be simple enough to avoid the use of this kind of intuition as evidence?

To consider whether we should expect to keep encountering situations in which it is tempting to use this kind of intuition, I will return to the example from Sterelny. There is a broader aspect of his hypothesis that, in my view, suggests judgments about plausible causal sequences will continue to be important to any hypothesis of this kind, as well as difficult to make. This is the fact that existing *plasticity* in the behavior and cognition of the LCA and descendent hominins was probably crucial in the evolution of novel capacities. Sterelny makes a compelling case that cognitive adaptations for novel kinds of thinking probably did not emerge as a result of selection among small, random genetic mutations. Rather, plastic responses to novel environmental features likely played a major role in creating the selective environments in which such adaptations could emerge (Sterelny 2012, 23–43). He uses the concept of a Zone of Latent Solutions (ZLS) from comparative psychology to argue that the behaviors and cognitive capacities that organisms exhibit in novel environments, rather than simply representing a limit on the organism's current capacities, could instead provide the conditions for genetic assimilation. He even argues that theorists like Tennie, Call, and Tomasello underestimate the power of the cognitive responses near the limits of the ZLS in driving evolution (Sterelny 2012, 31). The ZLS represents counterfactual limits of cognition for some species – biologists interested in chimpanzee cognition might ask, just how far can we push a chimpanzee before its cognition cannot be 'stretched' any further? What novel environments can we create that will reveal the potential phenotypes of chimpanzee cognition?

As we saw in the earlier examples, the limits of plasticity in ancestral hominin cognition in novel environments are exactly the kind of thing we would like to know and to invoke in a good explanation. It makes good sense that these should figure in a hypothesis about the causal steps between stages or associations among cognitive capacities. And yet, these are exactly the kinds of things whose plausibility we are in an especially bad position to evaluate. The ZLS is hard enough to delimit for extant chimpanzees; inferring a ZLS for an extinct hominin species is quite another matter.

Thus, on the one hand, phenotypic plasticity of cognitive traits almost certainly has something to do with the way that selection modified existing capacities. But on the other, this is just the place where our intuitions about plausibility cannot pull their epistemic weight. The importance of plasticity in combination with scarce evidence suggests that the issue I have raised in this paper will persist. But again, inferences about the abilities of ancestral hominins in novel situations are not in principle impossible to make — they *could* be made on the basis of good empirical evidence. As I mentioned earlier, insights from work in cognitive paleoanthropology can help on this front, by pointing theorists toward better kinds of evidence.

As for eliminating the use of evolutionarily biased intuitions in these contexts, we should expect it to be difficult but not impossible. If our tendency to use our own cognitive capacities to decide what cognitive capacities were evolutionarily available to ancestral species is as persistent as other cognitive biases, e.g., confirmation bias, we cannot expect to correct the problem by simply asking theorists to adhere more closely to the standards of good adaptationist thinking. Dacey, in his work on anthropomorphism as a cognitive bias, argues that certain mental exercises could counteract anthropomorphism in comparative psychology research (Dacey 2017, 13–15). However, social psychology techniques of this kind are typically forms of psychological priming, whose effects are questionable enough that one should hesitate to wholly accept them as corrective measures for experimental design (Earp and Trafimow, 2015; Stroebe et al., 2012; John et al., 2012; Fanelli, 2010; Ioannidis, 2005). Furthermore, other studies suggest that correcting deeply rooted cognitive tendencies simply by being aware of them is difficult or even ineffective (Lilienfeld et al., 2009; Pronin, 2008).

Moreover, unlike in experimental psychology, the phases of hypothesis generation and evaluation for reconstructing hominin evolution are entirely intermixed. Even if Dacey's suggestions can be implemented in experimental comparative psychology, a theorist reconstructing an evolutionary history from a broad body of potential evidence cannot straightforwardly perform a priming exercise prior to hypothesizing and then independently examine the evidence. Because there is a reliable obstacle to our reasoning in this context, heightened skepticism is epistemically justified and ought not be considered pessimistic. Therefore, theorists constructing and evaluating hypotheses for the evolution of human cognitive features ought to (1) not use intuitions about the association or evolutionary sequence of cognitive capacities and (2) expect to do so inadvertently. This amounts to adopting a more skeptical attitude toward speculation than they might for other evolutionary hypotheses.

My aim in this paper has been to identify and analyze a particularly difficult problem for reasoning about the minds of our hominin ancestors. Human cognition is a special explanatory target for evolutionary reconstruction because our evolved cognition affects the way that we think about evolutionary possibilities. However, I will conclude by emphasizing that I do not mean to suggest that the project of reconstructing human cognitive evolution is in some way *doomed*. Recognizing the issues with evolutionarily biased intuition is itself important, and continuing to examine methods and reasoning patterns in human cognitive evolution can help theorists to understand and pursue the right kinds of evidence. The ultimate goal is, of course, to produce the best possible reconstruction of human cognitive evolution. Theorists ought to pursue any evidence that they can to replace speculation. But they also ought to understand the challenges we face in doing so.

Bibliography

Allen, C. (2002). Real Traits, Real Functions? In Ariew, A., Cummins, R., and Perlman, M., editors, *Functions*, pages 373–389. Oxford University PressOxford.

Barker, G. (2015). *Beyond Biofatalism: Human Nature for an Evolving World*. Columbia University Press, New York, NY.

Barkow, J. H., Tooby, J., and Cosmides, L. (1992). *The Adapted Mind*. Oxford University Press.

Baumard, N., André, J.-B., and Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, 36(1):59–78.

Beatty, J. (2016). What are narratives good for? *Studies in history and philosophy of biological and biomedical sciences*, 58:33–40.

Beatty, J. (2017). Narrative possibility and narrative explanation. *Studies in History and Philosophy of Science Part A*, 62:31–41. Publisher: Elsevier Ltd.

Beatty, J. and Carrera, I. (2011). When What Had to Happen Was Not Bound to Happen: History, Chance, Narrative, Evolution*. *Journal of the Philosophy of History*, 5(3):471–495. Publisher: Brill.

Beatty, J. and Desjardins, E. C. (2009). Natural selection and history. *Biology and Philosophy*, 24(2):231–246. Publisher: Springer.

Bokulich, A. (2014). How the Tiger Bush Got Its Stripes : 'How Possibly' vs. 'How Actually' Model Explanations. *The Monist*, 97(3):321–338.

Boyd, R. and Richerson, P. J. (1988). *Culture and the Evolutionary Process*. University of Chicago Press, Chicago, IL.

Brandon, R. N. (1990). *Adaptation and Environment*. Princeton University Press, Princeton.

Brown, R. L. (2014). What Evolvability Really Is. *The British Journal for the Philosophy of Science*, 65(3):549–572. Publisher: The University of Chicago Press.

Buller, D. J. (2005). *Adapting Minds*. MIT Press.

Buller, D. J. and Hardcastle, V. G. (2000). Evolutionary Psychology, Meet Developmental Neurobiology: Against Promiscuous Modularity. *Brain and Mind*, 1(3):307–325.

Calcott, B. (2009). Lineage explanations: Explaining how biological mechanisms change. *British Journal for the Philosophy of Science*, 60(1):51–78.

Call, J., Hare, B., Carpenter, M., and Tomasello, M. (2004). 'Unwilling' versus 'unable': Chimpanzees' understanding of human intentional action. *Developmental Science*, 7(4):488–498. Publisher: John Wiley & Sons, Ltd.

Carrier, M., Mertens, R., and Reinhardt, C., editors (2021). *Narratives and Comparisons: Adversaries or Allies in Understanding Science?* transcript Verlag.

Cleland, C. E. (2001). Historical science, experimental science, and the scientific method. *Geology*, 29(11):987–990. Publisher: Geological Society of America.

Cleland, C. E. (2002). Methodological and Epistemic Differences between Historical Science and Experimental Science. *Philosophy of Science*, 69(3):447–451.

Craver, C. F. (2006). When mechanistic models explain. *Synthese*, 153(3):355–376. Publisher: Springer.

Crawford, C., Smith, M., and Krebs, D., editors (1987). *Sociobiology and psychology: ideas, issues and applications*. Lawrence Erlbaum Associates, New York.

Currie, A. (2014). Narratives, mechanisms and progress in historical science. *Synthese*, 191(6):1163–1183.

Currie, A. (2015). Marsupial lions and methodological omnivory: function, success and reconstruction in paleobiology. *Biology and Philosophy*, 30(2):187–209. Publisher: Kluwer Academic Publishers.

Currie, A. (2016). Ethnographic analogy, the comparative method, and archaeological special pleading. *Studies in History and Philosophy of Science Part A*, 55:84–94.

Currie, A. (2018). *Rock, Bone, and Ruin*. MIT Press.

Currie, A. (2019a). Scientific Knowledge of the Deep Past: History Matters. *Elements in the Philosophy of Science*. ISBN: 9789896540821.

Currie, A. (2019b). Simplicity, one-shot hypotheses and paleobiological explanation. *History and Philosophy of the Life Sciences*, 41(1):10.

Currie, A. and Killin, A. (2019). From things to thinking: Cognitive archaeology. *Mind & Language*, 34(2):263–279. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mila.12230.

Currie, A., Killin, A., Lequin, M., Meneganzin, A., and Pain, R. (2024). Past materials, past minds: The philosophy of cognitive paleoanthropology. *Philosophy Compass*, 19(6):e13001. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/phc3.13001.

Currie, A. and Meneganzin, A. (2024). Hawkes' Ladder, Underdetermination, and the Mind's Capacities. In Wynn, T., Overmann, K. A., and Coolidge, F. L., editors, *Oxford Handbook of Cognitive Archaeology*, page 0. Oxford University Press.

Currie, A. and Sterelny, K. (2017). In defence of story-telling. *Studies in History and Philosophy of Science Part A*, 62:14–21.

Currie, A. and Swaim, D. (2021). Past Facts and the Nature of History. *Journal of the Philosophy of History*, 16(2):179–206. Publisher: Brill.

Dacey, M. (2017). Anthropomorphism as Cognitive Bias. *Philosophy of Science*, 84(5):1152–1164.

Danto, A. C. (1962). Narrative Sentences. *History and Theory*, 2(2):146–179.

Darwin, C. (1879). *The Descent of Man*. John Murray, London. Publication Title: Penguin Classics 2004.

de Waal, F. (1982). *Chimpanzee Politics*. Jonathan Cape, London.

Desjardins, E. (2011). Historicity and experimental evolution. *Biology and Philosophy*, 26(3):339–364. Publisher: Springer.

DiFrisco, J. and Ramsey, G. (2023). Adaptationism and Trait Individuation. *Philosophy of Science*, 90(5):1234–1243.

Downes, S. M. (2001). Some Recent Developments in Evolutionary Approaches to the Study of Human Cognition and Behavior. *Biology and Philosophy*, 16(5):575–594.

Downes, S. M. (2013). No Magic Bullet Explains the Evolution of Unique Human Traits. *Biological Theory*, 8(1):15–19.

Downes, S. M. (2016). Evolutionary Psychology. In *The Routledge Companion to Philosophy of Social Science*, pages 330–339. Routledge.

Dray, W. H. (1957). *Laws and Explanation in History*. Oxford University Press. Google-Books-ID: M7RmAAAAMAAJ.

Earp, B. D. and Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, 6:621.

Ereshefsky, M. and Turner, D. (2020). Historicity and explanation. *Studies in History and Philosophy of Science Part A*, 80:47–55.

Fanelli, D. (2010). "Positive" Results Increase Down the Hierarchy of the Sciences. *PLOS ONE*, 5(4):e10068. Publisher: Public Library of Science.

Fisher, M. (1942). *How to Cook a Wolf*. Duell, Sloan and Pearce, New York.

Forber, P. (2010). Confirmation and explaining how possible. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*, 41(1):32–40. Publisher: Pergamon.

Forber, P. (2012). Conjecture and explanation: A reply to Reydon. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(1):298–301.

Gallie, W. (1968). *Philosophy and the Historical Understanding*. Schocken Books, New York, 2 edition.

Glennan, S. (2010). Ephemeral mechanisms and historical explanation. *Erkenntnis*, 72(2):251–266.

Godfrey-Smith, P. (2001). Three Kinds of Adaptationism. In Orzack, S. H. and Sober, E., editors, *Adaptationism and Optimality*, pages 335–357. ISSN: 0012-7183.

Goodall, J. (2000). *In the Shadow of Man*. Houghton Mifflin Harcourt.

Goodwin, G. P. and Darley, J. M. (2008). The psychology of meta-ethics: Exploring objectivism. *Cognition*, 106(3):1339–1366.

Goodwin, G. P. and Darley, J. M. (2010). The Perceived Objectivity of Ethical Beliefs: Psychological Findings and Implications for Public Policy. *Review of Philosophy and Psychology*, 1(2):161–188.

Goodwin, G. P. and Darley, J. M. (2012). Why are some moral beliefs perceived to be more objective than others? *Journal of Experimental Social Psychology*, 48(1):250–256.

Gould, S. J. (1978). Sociobiology: the art of storytelling. *New Scientist*, 80(1129):530–533.

Gould, S. J. (1990). *Wonderful Life: The Burgess Shale and the Nature of History*. W. W. Norton & Company.

Gould, S. J. and Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society of London - Biological Sciences*, 205(1161):581–598.

Griffiths, P. E. (1996). The Historical Turn in the Study of Adaptation. *British Journal for the Philosophy of Science*, 47(4):511–532.

Griffiths, P. E. (2006). Evolutionary Psychology. In *The Philosophy of Science: An Encyclopedia*, pages 263–268. Routledge, New York.

Griffiths, P. E. (2011). Ethology, Sociobiology, and Evolutionary Psychology. In Sarkar, S. and Plutynski, A., editors, *A Companion to Philosophy of Biology*, Blackwell Companions to Philosophy, pages 393–414.

Grüne-Yanoff, T. (2013). Appraising models nonrepresentationally. *Philosophy of Science*, 80(5):850–861. Publisher: University of Chicago PressChicago, IL.

Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*, 35(1):1–15.

Hajek, K. M. (2022). What Is Narrative in Narrative Science? The Narrative Science Approach. In Morgan, M. S., Hajek, K. M., and Berry, D. J., editors, *Narrative Science*, pages 31–58. Cambridge University Press, 1 edition.

Hare, B., Call, J., Agnetta, B., and Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*, 59(4):771–785. Publisher: Academic Press.

Hare, B. and Tomasello, M. (2004). Chimpanzees are more skilful in competitive than in cooperative cognitive tasks. *Animal Behaviour*, 68(3):571–581.

Hawkes, C. (1954). Archeological Theory and Method: Some Suggestions from the Old World. *American Anthropologist*, 56(2):155–168. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1525/aa.1954.56.2.02a00660.

Heider, F. and Simmel, M. (1944). An Experimental Study of Apparent Behavior. *The American Journal of Psychology*, 57(2):243.

Hempel, C. G. (1942). The Function of General Laws in History. *The Journal of Philosophy*, 39(2):35–48.

Heyes, C. (2018). *Cognitive Gadgets: The Cultural Evolution of Thinking*. Belknap Press, Cambridge, MA.

Hitchcock, C. and Woodward, J. (2003). Explanatory Generalizations, Part II: Plumbing Explanatory Depth. *NOÛS*, 37(2):181–199.

Hrdy, S. B. (2009). *Mothers and Others: The Evolutionary Origins of Mutual Understanding*. The Belknap Press.

Hull, D. L. (1975). Central Subjects and Historical Narratives. *History and Theory*, 14(3):253–274.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, 2(8):e124. Publisher: Public Library of Science.

Jackson, F. and Pettit, P. (1992). In Defense of Explanatory Ecumenism. *Economics & Philosophy*, 8(1):1–21.

Jeffares, B. (2008). Testing times: regularities in the historical sciences. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*, 39(4):469–475. Publisher: Stud Hist Philos Biol Biomed Sci.

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, 23(5):524–532. Publisher: SAGE Publications Inc.

Joyce, R. (2006). *The Evolution of Morality.* MIT Press.

Kaplan, J. (2002). Historical Evidence and Human Adaptations. *Philosophy of Science*, 69(S3):S294–S304.

Killin, A. and Pain, R. (2021). Cognitive Archaeology and the Minimum Necessary Competence Problem. *Biological Theory*.

Killin, A. and Pain, R. (2023). How WEIRD is Cognitive Archaeology? Engaging with the Challenge of Cultural Variation and Sample Diversity. *Review of Philosophy and Psychology*, 14(2):539–563.

Kitcher, P. (1989). *Vaulting Ambition: Sociobiology and the Quest for Human Nature*. MIT Press, Cambridge, Massachusetts.

Kitcher, P. (2011). *The Ethical Project*. Harvard University Press.

Kosso, P. (2001). *Knowing the Past: Philosophical Issues of History and Archaeology*. Humanity Books, Amherst.

Kranke, N. (2022). Two kinds of historical explanation in Evolutionary Biology. *Biology and Philosophy*, 37(3):1–21. Publisher: Springer Science and Business Media B.V.

Lala, K. N. and Brown, G. R. (2011). *Sense and nonsense: evolutionary perspectives on human behaviour*. Oxford University Press, 2 edition.

Lee, J. (2024). Ice cream, Nazis, and the adaptive advantage of externalization.

Lennox, J. G. (1991). Darwinian Thought Experiments: A Function for Just-so Stories. In Horowitz, T. and Massey, G., editors, *Thought Experiments in Science and Philosophy*, pages 223–246. Rowman and Littlefield, Savage, MD.

Levins, R. (1966). The strategy of model building in population biology. *American Naturalist*, 54(4):421–431. ISBN: 0003-0996.

Lewis, D. (1986). Postscripts to 'Causation'. In Lewis, D., editor, *Philosophical Papers, Volume II*. Oxford University Press.

Lewis, D. (2000). Causation as Influence. *The Journal of Philosophy*, 97(4):182. Publisher: Philosophy Documentation Center.

Lewontin, R. (1998). The evolution of cognition: Questions we will never answer. In Scarborough, D. and Sternberg, S., editors, *Methods, models, and conceptual issues: An invitation to cognitive science*, volume 4, pages 106–132. MIT Press.

Lewontin, R. C. (1979). Sociobiology as an Adaptationist Program. *Behavioral Science*, 24(1).

Lilienfeld, S. O., Ammirati, R., and Landfield, K. (2009). Giving Debiasing Away: Can Psychological Research on Correcting Cognitive Errors Promote Human Welfare? *Perspectives on Psychological Science*, 4(4):390–398.

Lloyd, E. A. (1999). Evolutionary Psychology: The Burdens of Proof. Technical report. Publication Title: Biology and Philosophy Volume: 14.

Lloyd, E. A. (2003). Violence against Science: Rape and Evolution. In Travis, C. B., editor, *Evolution, Gender, and Rape*. MIT Press.

Lloyd, E. A. (2010). *Science, Politics, and Evolution*. Cambridge University Press.

Lloyd, E. A. and Feldman, M. W. (2002). Evolutionary Psychology: A View From Evolutionary Biology. *Psychological Inquiry*, 13(2):150–156.

Machamer, P., Darden, L., and Craver, C. F. (2000). Thinking About Mechanisms. *Philosophy of Science*, 67:1–25.

Melis, A. P., Altrichter, K., and Tomasello, M. (2013). Allocation of resources to collaborators and free-riders in 3-year-olds. *Journal of Experimental Child Psychology*, 114(2):364–370.

Melis, A. P., Schneider, A.-C., and Tomasello, M. (2011). Chimpanzees, Pan troglodytes, share food in the same way after collaborative and individual food acquisition. *Animal Behaviour*, 82(3):485–493.

Melis, A. P. and Semmann, D. (2010). How is human cooperation different? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1553):2663–2674.

Melis, A. P., Warneken, F., Jensen, K., Schneider, A.-C., Call, J., and Tomasello, M. (2010). Chimpanzees help conspecifics obtain food and non-food items. *Proceedings of the Royal Society B: Biological Sciences*, 278(1710):1405–1413. Publisher: Royal Society.

Meneganzin, A. and Currie, A. (2022). Behavioural modernity, investigative disintegration & Rubicon expectation. *Synthese*, 200(1):47.

Meneganzin, A. and Currie, A. (2025). Not wasted on the young: Childhood, trait complexes & human behavioral ecology. *Studies in History and Philosophy of Science*, 109:12–20.

Meneganzin, A. and Killin, A. (2024). Beyond reasonable doubt: reconsidering Neanderthal aesthetic capacity. *Phenomenology and the Cognitive Sciences*.

Meneganzin, A., Ramsey, G., and DiFrisco, J. (2024). What is a trait? Lessons from the human chin. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 342(2):65–75.

Mink, L. O. (1966). The Autonomy of Historical Understanding. *History and Theory*, 5(1):24–47. Publisher: [Wesleyan University, Wiley].

Morgan, M. S. (2017). Narrative ordering and explanation. *Studies in History and Philosophy of Science Part A*, 62:86–97. Publisher: Pergamon.

Morgan, M. S. (2022). Narrative: A General-Purpose Technology for Science. In Berry, D. J., Hajek, K. M., and Morgan, M. S., editors, *Narrative Science: Reasoning, Representing and Knowing since 1800*, pages 3–30. Cambridge University Press, Cambridge.

O'Connor, C. (2017). Modeling Minimal Conditions for Inequity.

O'Connor, C. (2020). *Games in the Philosophy of Biology*. Cambridge Elements. Cambridge University Press.

Olmos, P. (2022). Just-so What? In *Narrative Science: Reasoning, Representing, and Knowing since 1800*, pages 424–444. Cambridge University Press.

Pain, R. (2021). What Can the Lithic Record Tell Us About the Evolution of Hominin Cognition?

Pain, R. (2023). Gadgets Meet Artefacts: Aligning Heyes's Cultural Evolutionary Account With the Archaeological Record. *Perspectives on Psychological Science*, page 17456916231187405.

Parker, W. S. (2020). Model Evaluation: An Adequacy-for-Purpose View. *Philosophy of Science*, 87(3):457–477. Publisher: Cambridge University Press.

Peters, U. (2016). Human thinking, shared intentionality, and egocentric biases. *Biology & Philosophy*, 31(2):299–312.

Pigliucci, M. and Kaplan, J. (2000). The fall and rise of Dr Pangloss: adaptationism and the Spandrels paper 20 years later. *Trends in Ecology & Evolution*, 15(2):66–70.

Povinelli, D. J. (2000). *Folk Physics for Apes: The Chimpanzee's Theory of How the World Works*. Oxford University Press, Oxford.

Povinelli, D. J. and Vonk, J. (2003). Chimpanzee minds: suspiciously human? *Trends in Cognitive Sciences*, 7(4):157–160.

Povinelli, D. J. and Vonk, J. (2004). We Don't Need a Microscope to Explore the Chimpanzee's Mind. *Mind and Language*, 19(1):1–28. Publisher: John Wiley & Sons, Ltd.

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526. Publisher: Cambridge University Press.

Pronin, E. (2008). How We See Ourselves and How We See Others. *Science*, 320(5880):1177–1180.

Ratajska, A., Brown, M. I., and Chabris, C. F. (2020). Attributing social meaning to animated shapes: A new experimental study of apparent behavior. *The American Journal of Psychology*, 133(3):295–312.

Resnik, D. B. (1991). How-possibly explanations in biology. *Acta Biotheoretica*, 39(2):141–149.

Reydon, T. A. (2012). How-possibly explanations as genuine explanations and helpful heuristics: A comment on Forber. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(1):302–310.

Reydon, T. A. C. (2023). The proper role of history in evolutionary explanations. *Noûs*, 57(1):162–187. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/nous.12402.

Rice, C. C. (2016). Factive scientific understanding without accurate representation. *Biology and Philosophy*, 31(1):81–102. Publisher: Springer Netherlands.

Richards, R. J. (1992). The Structure of Narrative Explanation in History and Biology. In Nitecki, M. H. and Nitecki, D. V., editors, *History and Evolution*, pages 19–55. SUNY Press.

Ross, L. N. (2019). Causal Control: A Rationale for Causal Selection. In *Philosophical Perspectives on Causal Reasoning in Biology*, volume XXI of *Minnesota Studies in Philosophy of Science*.

Ross, L. N. and Woodward, J. F. (2022). Irreversible (One-hit) and Reversible (Sustaining) Causation. *Philosophy of Science*, 89(5):889–898.

Roth, P. A. (2017). Essentially narrative explanations. *Studies in History and Philosophy of Science Part A*, 62:42–50. Publisher: Pergamon.

Roth, P. A. (2020). *The Philosophical Structure of Historical Explanation*. Northwestern University Press.

Sarkissian, H. (2016). Aspects of Folk Morality: Objectivism and Relativism. In Sytsma, J. and Buckwalter, W., editors, *A Companion to Experimental Philosophy*, pages 212–224. Wiley, 1 edition.

Scher, S. J. and Rauscher, F., editors (2003). *Evolutionary Psychology: Alternative Approaches*. Kluwer Academic Publishers, Dordrecht.

Skitka, L. J., Bauman, C. W., and Sargis, E. G. (2005). Moral Conviction: Another Contributor to Attitude Strength or Something More? *Journal of Personality and Social Psychology*, 88(6):895–917.

Smith, J. M. and Price, G. R. (1973). The Logic of Animal Conflict. *Nature*, 246(5427):15–18. Publisher: Nature Publishing Group.

Sober, E. (2024). *The Philosophy of Evolutionary Theory*. Cambridge University Press.

Stanford, P. K. (2017a). Bending toward Justice. *Philosophy of Science*, 84(2):369–376.

Stanford, P. K. (2017b). The Difference Between Ice Cream and Nazis: Moral Externalization and the Evolution of Human Cooperation. *Behavioral and Brain Sciences*, (May):1–57.

Sterelny, K. (1992). Evolutionary explanations of human behaviour. *Australasian Journal of Philosophy*, 70(2):156–173.

Sterelny, K. (1995). Review: The Adapted Mind. *Biology and Philosophy*, 10:365–380. ISBN: 0195101073.

Sterelny, K. (2012). *The Evolved Apprentice: How Evolution Made Humans Unique*. MIT Press, Cambridge, MA. Publication Title: Jean Nicod Lectures.

Sterelny, K. (2016). Contingency and History. *Philosophy of Science*, 83(4):521–539. Publisher: University of Chicago PressChicago, IL.

Sterelny, K. (2021). *The Pleistocene Social Contract: Culture and Cooperation in Human Evolution*. Oxford University Press.

Sterelny, K. (2022). Ethnography, Archaeology, and the Late Pleistocene. *Philosophy of Science*, 89(3):415–433.

Sterelny, K. and Planer, R. J. (2021). *From Signal to Symbol: The Evolution of Language*. MIT Press.

Stroebe, W., Postmes, T., and Spears, R. (2012). Scientific Misconduct and the Myth of Self-Correction in Science. *Perspectives on Psychological Science*, 7(6):670–688. Publisher: SAGE Publications Inc.

Swaim, D. G. (2019). The Roles of Possibility and Mechanism in Narrative Explanation. *Philosophy of Science*, 86(5):858–868.

Swaim, D. G. (2021). What is narrative possibility? *Studies in History and Philosophy of Science Part A*, 89:257–266.

Tomasello, M. (2014). *A Natural History of Human Thinking*. Harvard University Press, Cambridge, Massachusetts.

Tomasello, M. (2016). *A Natural History of Human Morality*. Harvard University Press.

Tomasello, M. (2019). *Becoming Human: A Theory of Ontogeny*. Belknap Press, Cambridge, MA.

Tomasello, M., Hare, B., Lehmann, H., and Call, J. (2007). Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. *Journal of Human Evolution*, 52(3):314–320.

Trivers, R. L. (1971). The Evolution of Reciprocal Altruism. *The Quarterly Review of Biology*, 46(1):35–57. Publisher: The University of Chicago Press.

Turner, D. (2005). Local Underdetermination in Historical Science. *Philosophy of Science*, 72(1):209–230. Publisher: Cambridge University Press.

Turner, D. (2009). Beyond Detective Work: Empirical Testing in Paleontology. In Sepkoski, D. and Ruse, M., editors, *The Paleobiological Revolution: Essays on the Growth of Modern Paleontology*, page 0. University of Chicago Press.

Turner, D. (2016). A second look at the colors of the dinosaurs. *Studies in History and Philosophy of Science Part A*, 55:60–68.

Turner, D. and AboHamad, A. (2023). Narrative Explanation and Non-Epistemic Value. *Journal of the Philosophy of History*, 17(1):53–76. Publisher: Brill.

Verreault-Julien, P. (2019). How could models possibly provide how-possibly explanations? *Studies in History and Philosophy of Science Part A*, 73:22–33.

Warneken, F., Lohse, K., Melis, A. P., and Tomasello, M. (2011). Young Children Share the Spoils After Collaboration. *Psychological Science*, 22(2):267–273. Publisher: SAGE Publications Inc.

Warneken, F. and Tomasello, M. (2007). Helping and Cooperation at 14 Months of Age. *Infancy*, 11(3):271–294. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1532-7078.2007.tb00227.x.

Waters, C. K. (2007). Causes That Make a Difference. *Journal of Philosophy*, 104(11):551–579.

Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press.

Williams, G. C. and Williams, D. C. (1957). Natural Selection of Individually Harmful Social Adaptations Among Sibs With Special Reference to Social Insects. *Evolution*, 11(1):32–39. Publisher: [Society for the Study of Evolution, Wiley].

Wilson, E. (1975). *Sociobiology: The New Synthesis*. Harvard University Press, Cambridge, Massachusetts.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, New York, New York.

Woodward, J. (2010). Causation in biology: Stability, specificity, and the choice of levels of explanation. *Biology and Philosophy*, 25(3):287–318.

Wright, J. C., McWhite, C. B., and Grandjean, P. T. (2014). The Cognitive Mechanisms of Intolerance. In Knobe, J., Lombrozo, T., and Nichols, S., editors, *Oxford Studies in Experimental Philosophy*, pages 28–61. Oxford University Press.

Wylie, A. (1985). The Reaction against Analogy. *Advances in Archaeological Method and Theory*, 8:63–111. Publisher: Springer.

Wylie, A. (2011). Critical Distance: Stabilising Evidential Claims in Archaeology. In Dawid, P., Twining, W., and Vasilaki, M., editors, *Evidence, Inference and Enquiry*. British Academy.

Wylie, A. and Chapman, R. (2016). *Evidential Reasoning in Archaeology*. Bloomsbury Academic.