Fall 12-21-2018

Beyond Enlightenment: The Evolution of Agency and the Modularity of the Mind in a Post-Darwinian World

Derek Elliott

BEYOND ENLIGHTENMENT:

THE EVOLUTION OF AGENCY AND THE MODULARITY OF THE MIND

IN A POST-DARWINIAN WORLD

A Dissertation

Submitted to McAnulty College and Graduate School of Liberal Arts

Duquesne University

In partial fulfillment of the requirements for

the degree of Doctor of Philosophy

By

Derek Anthony Elliott

December 2018

BEYOND ENLIGHTENMENT:

THE EVOLUTION OF AGENCY AND THE MODULARITY OF THE MIND

IN A POST-DARWINIAN WORLD

By

Derek Anthony Elliott

Approved September 13, 2018:

Patrick Lee Miller Associate Professor of Philosophy (Committee Chair)	James Swindal Dean, McAnulty College and Graduate School of Liberal Arts Professor of Philosophy (Committee Member)
Jeffrey McCurry Directory, Simon Silverman Phenomenology Center (Committee Member)	
James Swindal Dean, McAnulty College and Graduate School of Liberal Arts Professor of Philosophy	Ronald Polansky Chair, Philosophy Department Professor of Philosophy

ABSTRACT

BEYOND ENLIGHTENMENT:

THE EVOLUTION OF AGENCY AND THE MODULARITY OF THE MIND

IN A POST-DARWINIAN WORLD

By

Derek Anthony Elliott

December 2018

Dissertation supervised by Patrick Lee Miller

Working out of the social and philosophical revolutions from the Enlightenment, contemporary action theory has unwittingly inherited several Cartesian ideas regarding the human mind: that it is unified, rational, and transparent. As a result, we have for too long conceived of action as intimately bound up with reason such that to act at all is to act *for* a reason, leaving us with theoretical difficulties in accounting for the behavior of non-human animals as well as irrational behavior in human beings.

But rather than propose that such difficulties can be resolved by retreating to a pre-Enlightenment view of human nature, the solution is to make the philosophical turn and embrace the insights that have been secured by Charles Darwin. It is a post-Darwinian evolutionary worldview that can shed some new light on these traditional problems. Two such innovations from the theory of evolution have been *evolutionary*

explanations, which attempt to understand the functions of organisms as having developed in response to environmental pressures, and *modular theory*, which views organisms as composed of parts with highly specialized functions.

Taking these evolutionary ideas together along with the assumption of biological continuity—that there is a developmental history shared by living organisms—we can begin to conceive of more robust theories of action, mind, and human nature. Contrary to Enlightenment conceptions, reason emerges as just one mental process alongside many, the mind appears anything but Cartesian, and agency begins far earlier along the spectrum of life than we have been supposing.

DEDICATION

To Family

ACKNOWLEDGEMENT

I would like to thank first and foremost my wife who has had all of the patience in the world with me and has been exactly what I needed, when I needed it.

I want to thank my director, Patrick Lee Miller, without whom, none of this would have been possible. In Socratic fashion, you have helped me remember how to read, write, and think. Your philosophical guidance helped organize my thoughts in a way that I would have never imagined.

I would like to thank Duquesne University, Gumberg Library, and the Philosophy Department, especially Joan Thompson, Linda Rendulic, and Ronald Polansky for their unending help. I also but especially thank Dean Swindal and Jeffrey McCurry for your support during one of the most difficult times in my life.

Lastly, I cannot forget my friends and family, including my brother Chad, who kindly entertains my ideas from time to time and challenges them, Tim Vickers and Dom Barnabei for reading parts of my work, and Curt Simcox and Todd Diedrich for their moral support.

TABLE OF CONTENTS

LIST OF FIGURES

Introduction: In the Beginning, the Logos

> And we can only wonder why one must become ill in order to have access to such truth.
>
> — Sigmund Freud, "Mourning and Melancholia"

0.1 | The End of an Era

The Enlightenment has ended.

Jean d'Alembert's "century of philosophy *par excellence*" became aware of the first symptoms of its terminal illness in 1781, when Immanuel Kant discovered buried in its intellectual development unresolvable tensions known as *antinomies*, contradictions that emerge from a set of beliefs that otherwise appear reasonable (Gillespie 2008, 258–9). The most notable of these, argues philosopher Michael Gillespie,[1] is the Third Antinomy (see figure 0.A), in which Kant exposes the absurdity that underlies our efforts to reconcile human freedom with causal determinism (259).

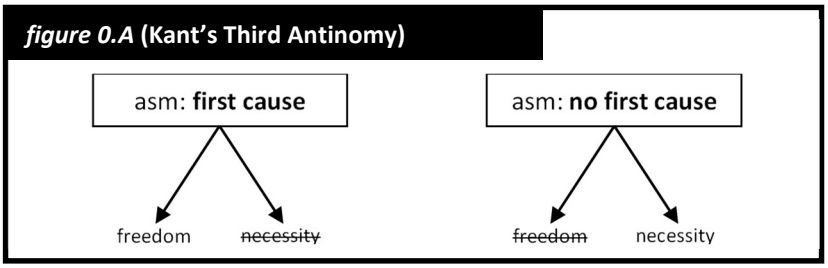
In his *Critique of Pure Reason*, Kant observes with his Third Antinomy that if we assume there is a first cause of the universe, it is, by definition, free from determinism. Proof that there is such a first cause rests in the fact there can be no causation at all without something first setting the chain of events into motion (*Critique of Pure Reason*, A446/B474). Such a consequence therefore yields the possibility that other such beings can similarly act freely (A450/B478).

Yet, on the other hand, the assumption of determinism requires that we suppose events occur *necessarily*, a metaphysical requirement that is incompatible with freedom,

---

[1] Though the symptoms presented themselves most clearly in Kant's writings, both David Hume and Jean-Jacques Rousseau can be said to have sounded the alarm bells earlier, imploring their intellectual peers to reconsider some of their cherished assumptions. Gillespie acknowledges in particular Rousseau's anxiety that modernity might be making human beings worse and Hume's metaphysical skepticism that there is no necessary connection between cause and effect. In spite of these efforts, explains Gillespie, "while such criticism did not go unheard, European intellectuals were still overwhelmingly committed to modern thought in the broadest sense in 1789" (Gillespie 2009, 256).

including freedom of a first cause (*Critique of Pure Reason*, A445/B473). Our faith in this determinist thesis rests in the fact that we have discovered laws of nature, enabling us to make predictions and reliably replicate our discoveries (A447/B475). Contrary to the freedom thesis, we need not imagine a first cause if we assume that this dynamical universe has always existed (A449/B477).



Unable to decisively prove either assumption is true or false and having good reason for both, Kant reveals to us how the use of reason sometimes issues in a manifest contradiction. How is this possible? Perhaps we can argue that this is an empty, nonsensical problem guising itself in the clothing of meaning, a clever presentation of something like a liar's paradox with no intelligible answer? Or maybe, to the abhorrence of Enlightenment thinkers, this points to a defect in human reason? Is reason itself somehow fatally flawed?

Kant rejects both of these possibilities.[2] As he sees it, the problem resides neither with semantics nor a defect in reason, nor even nature herself. Gillespie elaborates:

> The apparent contradiction of reason with itself is thus the consequence not of the contradictory character of existence or the inadequacy of reason but of the *misuse*

[2] Kant's solution, as presented in the *Critique of Pure Reason*, is more complicated than this, requiring a discussion of his philosophical division between noumena and phenomena to appreciate fully. Though he rejects both theses, he attempts to preserve the idea that the natural world is determined and that human beings are free by arguing that because we experience ourselves as free, it is *possible* that we are. Whether we *actually* are or are not is a question, he thinks, whose answer goes beyond what can be known by reason. For an introductory, detailed discussion of this, as well as the issues that Kant's solution raises, see Alfred Ewing's *A Short Commentary on Kant's Critique of Pure Reason* (Ewing 1967, 226–40), Allen Wood's *Kant* (Wood 2005, 89–100), and Sebastian Gardner's *Kant and the Critique of Pure Reason* (Gardner 1999, 257–64).

> of reason. The correct use of reason, Kant believes, will make possible the mastery of nature and attainment of human freedom, which will produce prosperity and morality and consequently political liberty and perpetual peace. (Gillespie 2008, 260; emphasis added)

Guided by his Enlightenment faith in the power of reason, Kant implores that we rein in these rational horses from the arid deserts of speculation so that they might run free in the prairies of empirical science and morality, their natural habitats. And yet, this exhortation was not satisfying to those intellectuals who followed in his wake, believing instead that Kant sidestepped the true difficulties raised by the antinomies (260).

Though the Enlightenment entered its final stage of life in 1781, the death knell officially sounded twelve years later in 1793. Inspired by the success of the American Revolution, a revolution widely viewed as epitomizing some of the core values of Enlightenment, such as the inevitability of progress and the rule of reason, the French would eventually set out to overthrow the *ancien régime* whose social institutions privileged the clergy and nobility, beginning with the execution of King Louis XVI (Gillespie, 255). Now that there was no turning back, the age of reason was supposed to have officially ushered in an era of peace, freedom, and prosperity. The results were anything but.

What followed in the aftermath of the French Revolution was the *Reign of Terror*, a brief but horrifying period during which the new government, in an effort to protect and promote its noble interests in Enlightenment values, sanctioned executions of anyone who dissented with the new ideals,[3] leaving more than thirty thousand dead by July of

[3] While the official number of deaths is recorded as 16,594, this number refers to those who faced *judicial* execution by way of guillotine. There were also *summary* executions during this time—immediate executions in the absence of a trial—and so the exact number is more challenging to estimate, although historian Donald Greer places this number at 25,000 (Greer 1935, 115). It is also important to keep in mind

1794 (255–6). It was the *Terror* that officially punctuated the end of Enlightenment, as philosopher William Bristow explains:

> The French revolutionaries meant to establish in place of the *ancien régime* a new reason-based order instituting the Enlightenment ideals of liberty and equality. Though the Enlightenment, as a diverse intellectual and social movement, has no definite end, the devolution of the French Revolution into the Terror in the 1790s, corresponding, as it roughly does, with the end of the eighteenth century and the rise of opposed movements, such as Romanticism, can serve as a convenient marker of the end of the Enlightenment, conceived as an historical period. (Bristow 2017, 2)

That faith in the dawn of an era of reason and peace should be expected to follow from the seeds of violent revolution, widespread paranoia, and intellectual defenses of the importance of terror is an irony that should not go unnoticed, betraying the presence of contradiction even in the *practice* of Enlightenment.

If there were any single person who embodied both the Enlightenment and its hidden contradictions, it was Maximilien Robespierre.

0.2 | The Antinomy of Impure Practical Reason

Born Maximilien-François-Marie-Isidore de Robespierre on May 6th, 1758, the oldest son of a lawyer would eventually become the symbol of the best and worst of the French Revolution. From early on, he had every reason to resent his lot in life. His mother,

---

that the *Terror* reached beyond executions as well. French historian François Furet puts into perspective its tragically long reach:

> The number of arrests from March 1793 to the end of July 1794 was far higher, probably close to half-million: this figure gives some idea of the shock caused by a repressive wave of these dimensions. It also indicates that there were not only acquittals but also, occasionally, penalties other than the death sentence, as well as "suspects" who languished in prison until 9 Thermidor [the 11th month in the French Republican Calendar] without being tried. The Terror's victims came from all levels of society, with each conflict producing its own characteristic shadings: more peasants in the Vendée, more bourgeois in Paris, Lyons, and Nîmes. In proportion to their relatively small numbers, the upper classes and clergy were comparatively hard-hit. (Furet 1989, 143)

Jacqueline, died at the young age of twenty-nine during childbirth, and shortly afterwards, his father, François, ran up his family's debts before running out altogether. Although his sister Charlotte recalled that Robespierre could never think of his mother without tears, and although he experienced poverty as a result of his father's decisions, he nonetheless developed an intense work ethic, one that earned him a scholarship to the prestigious Collége Louis-le-Grand before pursuing a career in law himself. In spite of his success, he always lived his life very frugally and humbly (Cavendish 2008).

While studying at Louis-le-Grand, Robespierre excelled in the classics and dreamed of the Roman Republic, reading inspirational passages from the volumes of Cato and Catullus. He had also fallen in love with the philosophical and political writings from Montesquieu and Jean-Jacques Rousseau, authors who inspired him to believe strongly in equality for all and the goodness of the will of the people. Though timid, nervous, and shy, he would somehow manage—eventually and not without enduring derision early in his political career—to deliver commanding speeches (Mantel 2000). Though possessing inferior charisma to his peers, it would be the content rather than the mode of delivery that resonated most with those who listened.

From the outset, Robespierre was swept away by dreams of democratic principles and progressivism, fruits from the tree of Enlightenment thought. He initially gained a reputation among the people by fighting for the working class while practicing law, and he was evidently good enough that it was starting to concern the wealthier citizens who were on the receiving end of his pugilistic prosecution (Cavendish 2008). This same fighting spirit was taken up into his political life, where he fiercely advocated for the ideas that the poor and the Jews should have equal rights, and he railed against slavery,

warfare, capital punishment, and government censorship of the press (Mantel 2000). It is hard to believe that such an otherwise quiet and shy, even perhaps pathetic,[4] person would become synonymous with the voice of the revolution.

His political life began right before the age of thirty-one in the year 1789, when King Louis XVI summoned the assembly of an Estates-General to discuss the nation's finances, the first assembled in over 175 years. Comprised of representatives of the three estates in France—the Clergy, Nobles, and Commons—elected officials were expected to attend conventions, air any grievances, and vote on matters accordingly. It just so happened that Robespierre would be one of the elected officials for the Commons, associating himself with a populist leftwing minority known as the Jacobin Club, among whom he became wildly influential (Mantel 2000).

Having declared war on Prussia and Austria, the French military appeared to be on the brink of defeat, and with an increasing lack of faith in an outdated political institution, socioeconomic tensions reached a fever pitch in 1792. Convinced that Louis XVI was conspiring against the revolutionary impulse, thanks in no small part to the efforts of Robespierre and his speeches in the Assembly, the Tuileries Palace was stormed on August 10th, and the monarchy overthrown. With Louis XVI arrested only three days later, his fate would be the subject of debate amongst the members of the new government, the National Convention, which had declared that France was now a republic. One of the most vocal leaders who called for his execution was none other than

[4] Robespierre was often described as having twitches, ulcers, skin infections, and nosebleeds, amongst other problems, and more recently, historical forensic scientists have retrospectively diagnosed him with the auto-immune disease, sarcoidosis, a condition that likely took its toll on him, ramping up especially in the last few years of his life. For more on this, see "Robespierre: the oldest case of sarcoidosis?" (Charlier and Froesch 2013).

Robespierre (Linton 2006). How could the man who only a handful of years earlier wished to outlaw capital punishment now be vociferously calling for its implementation in Louis XVI's case?

After the deposed king's execution on January 21st, 1793, it did not take long for the minority party of Robespierre, the Jacobins, to become frustrated with their political counterparts, the Girondins, during this period of political upheaval. They had become convinced that it was the Girondins who were principally responsible for slowing the movement down. Thus, only six months later in June, backed by the lower-class common folk known as the *sans-culotte*, the Jacobins violently overthrew and arrested the Girondins, quickly establishing a Committee of Public Safety as a war cabinet. When one of the twelve members of the committee had fallen ill, it was Robespierre who was called upon to replace him (Linton 2006).

Frustrated with the efforts of counter-revolutionaries, the *sans-culotte* demanded that this new government do something, and on September 17th, the Jacobins imbued their committees with the wide-reaching power and authority to arrest anyone suspected of trying to thwart the revolution's aims (Linton 2006). Nobody, they believed, should be lawfully permitted to stand in the way of the engine of progress. And thus, the *Reign of Terror* had begun.

With a brutally cold, calculating intellect, Robespierre justified this new policy in several public speeches over a series of months, a policy that was in direct contradiction with views he had espoused only years earlier. Given the content of his remarks, it was clear that though he had reached a different set of conclusions, he was convinced that this new course of action rationally followed from any good commitment to Enlightenment

ideals. The problem, as far as he was concerned, lies on the side of those who refuse to join their efforts, as well as politically corrupted and morally bankrupt revolutionaries. Indeed, Enlightenment demands that they purge their opponents for the sake of the greater good, for the sake of progress. Virtue, Robespierre believed, *requires* terror. Indeed, the two are inseparably logically intertwined.

The historian Marisa Linton calls our attention to this absurdity in one of Robespierre's speeches delivered in February 1794, spoken shortly after he had waxed poetically on the importance of peace, equality, and liberty:

> If the basis of popular government in peacetime is virtue, the basis of popular government during a revolution is both *virtue and terror*; virtue, without which terror is baneful; terror, without which virtue is powerless. Terror is nothing more than speedy, severe and inflexible justice; it is thus an emanation of virtue; it is less a principle in itself, than a consequence of the general principle of democracy, applied to the most pressing needs of the *patrie* [country]. (Linton 2006)

As fate would have it, Robespierre's enthusiasm for these new policies would eventually give way to paranoid fantasies that would lead to the execution of even those close to him, and in the summer of 1794, other influential members within the National Convention had grown weary with him, fearing that they would be next on the literal chopping block. As a result, they had reached their conclusion that he, too, had now become an enemy of the people and needed to be stopped (Bell 2012).

In July, after spending several weeks confining himself to his quarters out of increasing anxiety that there were few left he could trust, he finally worked up the courage to attend his first convention in over a month. There, he delivered his final speech that set out to, once again, justify his recent activities and decisions. During this speech, he had also enthusiastically promised that he had acquired evidence of political

misdeeds, but he refused to name those involved just yet (Linton 2006). This would be the last straw.

It was July 27th, only a day after his speech. Author Hilary Mantel describes the scene:

> He had written just two letters of his name, before a pistol shot shattered his jaw; whether he fired the shot himself, no one really knows. Lying in his own blood in an anteroom of the Committee of Public Safety, he gestured that he wished to write, but no one would give him a pen. "I would have given him a pen," Barras said later, uneasy at the cruelty and the lack of a possible disclosure. He was half-dead when he was taken to the scaffold, and his decapitated remains were buried near the Parc Monceau. (Mantel 2000)

On July 28th, this figurehead for the *sans-culotte*, this mouthpiece for the *vox populi*, had officially been swallowed up by the same tyrannical majority he had made his life's mission to defend, the same people in whom he saw the hope of a world of perpetual peace and prosperity for all promised by the Enlightenment.[5]

At every turn, Robespierre was certain that he was doing the right thing, and he took care to apply a rigorous logic in working out the details. The events during this time were striking enough that they attracted the attention of the German philosopher Georg Hegel, who sought to make sense of the madness. In one of his attempts, found in his *Lectures on the Philosophy of History*, he reckoned that the project was doomed for failure as soon as the people were expected to conform to ideals—in this case, the Enlightenment ideal, or, following Robespierre, what Hegel calls *virtue*. For, he argues, the only way to assess conformity *to* an ideal is *with* the ideal, and this naturally breeds suspicion, as everyone clamors to prove who belongs, that is, who is "one of us" and who is not. Nobody embodied the Enlightenment ideal more than Robespierre, but with

[5] For more on the life and complexity of Robespierre, see the collection of essays *Robespierre* (Haydon and Doyle 1999) as well as the biography *Robespierre: A Revolutionary Life* (McPhee 2012).

suspicion running amok, Hegel concludes, "virtue, as soon as it becomes liable to suspicion, is already condemned" (Hegel 1837, 450–1).

That these events had a profound impact on Hegel's thought is clear. Author John Rees summarizes some of these consequences:

> He began to think more critically about the legacy of the Enlightenment. The Jacobins generally, and Robespierre in particular, were the self-proclaimed followers of the Enlightenment thinkers. After all, was it not the Enlightenment belief that by altering men's environment we could improve their natures which stood behind Saint-Just's [one of Robespierre's closest friends] epigram, "It is for the legislator to make men into what he wants them to be"? Wasn't Robespierre Rousseau's ardent pupil? Hegel now began to question whether the stark project of the Enlightenment—to confront a recalcitrant world with the rational schemes of man—doesn't lead to the guillotine. (Rees 1998, 29)

And thus with Robespierre we find a warning that the pursuit of these Enlightenment ideals can not only prove self-defeating, but in his case, deadly.

0.3 | The Spirit of the Age

Unlike its death, there seems to be almost no consensus as to when the Enlightenment began, and for that matter, whether it can justly be characterized as one movement or a series of movements. Traditionally, historians tend to regard Francis Bacon, René Descartes, and Thomas Hobbes as belonging to some period before the Enlightenment, for which there are a number of proposed titles: "early Modern," "pre-Enlightenment," or even "post-Reformation. If that were not confusing enough, the British empiricist John Locke is sometimes considered "pre-Enlightenment" and sometimes as marking the beginning of the Enlightenment.[6] Likewise, we might ask to what degree the *philosophes*

[6] Philosopher Ernst Cassirer, for example, often refers to the age of Descartes as, rather appropriately, the "17th century," marking it off from the Enlightenment (Cassirer 1932 / 1979, 3; 6; 23). Historian John Spurr critically examines "post-Reformation," a term increasingly popular amongst historians of religion (Spurr 2002, 101). Historian Stephen Gaukroger, as well as Gillespie, tend to simply refer to it as "Modernity," although Gillespie believes that convention marks Locke as the beginning of Enlightenment (Gaukroger

of the French Enlightenment intellectually resemble the *commonsense philosophers* of the Scottish Enlightenment or their German counterparts, the *Wolffians* of the *Aufklärung*?

The confusion, however, is not a by-product of contemporary historians trying to piece together an ideological identity that unites the diverse thinkers of the 17th and 18th centuries, for it was a problem even at the time. Though usage of *Enlightenment* as a term had started to increase in popularity amongst those in the 18th century,[7] it was still vague enough even in its late stages to prompt Kant to explicitly address the idea in his 1784 essay, "Answer to the question: What Is Enlightenment?" There, he pronounces his faith in human reason, insisting that Enlightenment is the way forward from our own self-incurred immaturity and imprisonment, a way lit by the freedom from authority to use one's reason for all matters in life. Simultaneously imploring us while setting the motto for Enlightenment, he urges us: "*Think boldly!*" (Kant 1784, 135–6; 140).

Part of the issue with dating the Enlightenment comes down to whether it ought to be regarded as an historical period or as a movement, a worldview unified by a common set of ideas. For Kant, as well as many Enlightenment thinkers, it was a movement, and arguably even something greater than just that (Bristow 2017, 2–3). If it is regarded as the latter, then Enlightenment as a worldview clearly begins to take shape as the Scientific Revolution gets underway, a revolution that, to them, seemed to demonstrate the godlike power and promise of human reason. Given the absurdity witnessed during

2006; Gillespie 2008, 258). Historian Phyllis Leffler, however, has referred to the time period from 1660–1720 as "pre-Enlightenment" (Leffler 1976, 219).

[7] Gillespie claims that the term appeared in English for the first time with John Milton's *Paradise Lost* in 1667, but the *Oxford English Dictionary* now indicates that there were even earlier uses of it among religious texts, such as Robert Aylett's translation of the *Song of Songs* in 1621 (Gillespie 2008, 257).

the Roman Inquisition and the Thirty Years' War, the Scientific Revolution showed what could be accomplished when reason was freed from the shackles of authority. Viewed in this way, it becomes clear that earlier thinkers, even if not regarded as belonging to the Enlightenment *historically*, have every right to be tallied amongst its ranks (2). It is Hobbes, Descartes, and Locke, amongst others, who anticipate and promulgate the values of this new age, who make it possible. And uniting them above all is an unshakeable faith in the power of human reason to help us master the hostile forces of nature so that we might realize and enjoy the values of freedom, prosperity, and peace (Gillespie 2009, 258).

If Enlightenment is to be defined around a cluster of ideas, can we really say that it ever came to a decisive end?

Obviously, the answer is *no*. Many of the Enlightenment's most cherished ideas are romantically celebrated today, as if the Enlightenment project has not failed but instead has not been given enough time to deliver on its promises. To be sure, its ideals are admirable and its goals commendable,[8] but if these prove out of reach, as they did for Robespierre, it is only a matter of time before people become disillusioned and impatient, driving them to seek out another one of their own as a sacrificial lamb while ironically delaying the march of progress that demands we find an alternative.

Indeed, we must look elsewhere, for if history is any indication, we do not have all of the time in the world. If we wait too long, the contradictions inherent in Enlightenment thinking might prove fatal once more, irredeemably so.

---

[8] Steven Pinker, for example, argues in *Enlightenment Now* that this spirit is characterized by four themes: reason, science, humanism, and progress; and of these, he identifies faith in reason as central (Pinker 2018, 8).

0.4 | A Reasonable Defense of Violence

In 1651, Hobbes outlined a program for peace that would come to be known today as *deterrence theory*. "The passions that incline men to peace," he writes, "are fear of death, desire of such things as are necessary to commodious living, and a hope by their industry to obtain them" (Hobbes I.13.14). To him, and perhaps to many today, it is borderline trivial to declare that people have a drive for self-preservation. From this assumption, Hobbes derives the very reasonable conclusion that anybody would avoid those activities, criminal or otherwise, that jeopardize self-preservation. If you want criminal reform, then make the punishments *swifter* and more *severe*, and if you want foreign adversaries on the world stage to cooperate, then frighten them into believing that you can and will eliminate them when they cross a red line. Provided that everyone playing this game is a rational actor and shares the same beliefs, then they will surely derive the expected conclusions. This Enlightenment-inspired theory remains wildly popular to this day.

On November 28th, 1934, Winston Churchill delivered a speech before the House of Commons urging his government to increase its military defense spending out of fear that Germany was exercising a newly acquired vigor for rearmament, clearly in violation of the Treaty of Versailles. Churchill passionately advocated for his government to be proactive rather than reactive on this front, worrying that upgrading the military during a time of war would do too little, too late. Instead, he reasoned, if Germany could see that the United Kingdom was able to match them in terms of military might, it would deter the Germans from acting on any desire to attack in the first place. To the House of Commons, Churchill said:

> The fact remains that when all is said and done as regards defensive methods—and all that you can say now has been said already—pending some new

> discovery, the only direct measure of defence upon a great scale is the certainty of being able to inflict simultaneously upon the enemy as great damage as he can inflict upon ourselves. Do not let us under-value the efficacy of this procedure. It may well prove in practice—I admit you cannot prove it in theory—capable of giving complete immunity. If two Powers show themselves equally capable of inflicting damage upon each other by some particular process of war, so that neither gains an advantage from its adoption and both suffer the most hideous reciprocal injuries, it is not only possible but it seems to be probable that neither will employ that means. What would they gain by it? (HC Deb 28 November 1934 vol 295 c862)

Following Hobbes' rationale, it certainly seems sensible to believe that one could scare an aggressor into thinking twice about going to war. Clearly, we want to believe, one important motivating reason for going to war is the belief that somehow one possesses the upper-hand and has something to gain from it. If we can just make it look like this is not the case and show that going to war jeopardizes self-preservation, then it must follow that our opponent will no longer have any good reason to go to war in the first place. The logic is impeccable. Why would anyone act otherwise?

Although nuclear weapons were not yet a threat at the time Churchill delivered his speech, it would not be long before this same line of reasoning would be appropriated during the Cold War to establish the doctrine of Mutually Assured Destruction (MAD). On July 9th, 1962, the United States Secretary of Defense, Robert McNamara, delivered a commencement address to the University of Michigan wherein he discussed the international strategy for the US and NATO going forward in a world where adversarial foreign powers are in possession of novel weapons of mass destruction, nuclear missiles. Given that such weapons now existed, like Churchill before him, McNamara reached the conclusion that the only thing that could guarantee safety, much to everyone's disappointment, was building up and maintaining the nuclear arsenal in the hopes that it would discourage any enemies from using them as well. He explained:

> Let us look at the situation today. First, given the current balance of nuclear power, which we confidently expect to maintain in the years ahead, a surprise nuclear attack is simply not a rational act for any enemy. Nor would it be rational for an enemy to take the initiative in the use of nuclear weapons as an outgrowth of a limited engagement in Europe or elsewhere. I think we are entitled to conclude that either of these actions has been made highly unlikely. (McNamara 1962)

On the surface, deterrence seems like a perfectly reasonable strategy, and it is—if we were wholly or even mostly reasonable creatures in the way that the Enlightenment expects of us. We are called upon, once more, to live up to this Enlightenment ideal for the sake of our own security. Should we be concerned?

0.5 | Nearly Missing the Mark

To McNamara's credit, in spite of defending the doctrine of MAD, he was keenly aware that one can neither count on foreign leaders to act necessarily rationally nor to take all of the relevant variables into consideration when determining a course of action. He continued his speech to the University of Michigan:

> Second, and equally important, the mere fact that no nation could rationally take steps leading to a nuclear war does not guarantee that a nuclear war cannot take place. Not only do nations sometimes act in ways that are hard to explain on a rational basis, but even when acting in a "rational" way they sometimes, indeed disturbingly often, act on the basis of misunderstandings of the true facts of a situation. They misjudge the way others will react, and the way others will interpret what they are doing. We must hope, indeed I think we have good reason to hope, that all sides will understand this danger, and will refrain from steps that even raise the possibility of such a mutually disastrous misunderstanding. (McNamara 1962)

It can be described as nothing more than cruel irony that would create the conditions for such a "mutually disastrous misunderstanding" just three months later. During this grave situation, neither reason nor deterrence saved the day; it was luck.

The Cuban Missile Crisis began on October 16th, 1962. In response to the deployment of nuclear missile sites in Europe by the US, the Soviet Union returned the favor in kind by deploying their own sites in Cuba, creating an international boiling point in the process. During this tense standoff, a Soviet B-59, a submarine, had been patrolling the Caribbean waters after President John F. Kennedy had implemented a naval blockade of Cuba. Unbeknownst to the US military, however, this particular submarine had been equipped with a thermonuclear torpedo, carrying a payload comparable to the Little Boy atomic bomb dropped on Hiroshima during World War II.

Learning of the presence of the enemy submarine, the president and his council weighed their options, reasoning that intimidation was the best course of action. To accomplish their desired goal, they concluded that they could coerce the submersible into surfacing by dropping smaller depth charges nearby. These charges, they were reassured, are incapable of causing structural damage to it in the event of getting struck. By dropping these warning explosives, the intended message should be clear: their presence is known and not appreciated.

What instead ensued was mayhem inside the submarine.

One eyewitness to the account was Vadim Orlov, the onboard Intelligence Officer. According to Orlov, it was difficult to make out whether the depth charges were intended to be warnings or were signs that some sort of warfare had begun.[9] The submarine itself had lost contact with Moscow in the days leading up to this moment, operating on the captain's instruction to use their weapon only if Russia was in imminent

[9] For a recounting of the events mentioned in this paragraph, see Yasmeen Serhan, "When the World Lucked Out of a Nuclear War," *The Atlantic*, Oct. 31, 2017. https://www.theatlantic.com/international/archive/2017/10/when-the-world-lucked-out-of-nuclear-war/544198/.

danger. The protocol in place required that the senior officers onboard, captain included, reach a unanimous consensus in order to be authorized in using the weapon. In this case, three were present: Captain Vitali Savitsky, Deputy Political Officer Ivan Maslennikov, and Second-in-Command Vasili Arkhipov.

Tension was escalating inside the submarine with each tick of the clock and explosion outside the hull. Worse still, the air-conditioning had malfunctioned, causing temperatures in the cabin to reach more than 120°F in some areas, a known problem inside the vessels even when equipment was working optimally.[10] A decision now presented itself, one with enormous gravity. What in the world was going on? Was this the start of World War III? Was the US Navy *trying* to sink the submarine? Was this another invasion of Cuba? Or was this all some sort of terrible misunderstanding?

Convinced that it was foolish to take any chances, Savitsky had argued that the submarine needed to strike with its nuclear torpedo while it still had the chance, securing agreement from Maslennikov. That was two of the three affirmative votes needed for taking action. Fortunately for everyone today, Arkhipov, an otherwise quiet and soft-spoken person, firmly dissented. He believed that if the Navy *really* had been trying to sink their submarine, they *would have* sunk their submarine, erring on the side of caution that these explosions must be a barrage of warning shots intended to force them into complying with some further instruction.

In spite of the fact that this high-pressure, time-sensitive situation was infused with intense emotion, both sides in the submarine used good reasons in defense of their

[10] For a recounting of the events mentioned in this paragraph, see Robert Krulwich, "You (and Almost Everyone You Know) Owe Your Life to This Man," *National Geographic*, Mar. 25, 2016. https://news.nationalgeographic.com/2016/03/you-and-almost-everyone-you-know-owe-your-life-to-this-man/.

position. As a result, they reached a stalemate, neither willing to concede to the other. As far as they were concerned, there were no decisive arguments on the table that could settle the issue one way or the other for those involved. As if working through a Kantian antinomy, they each recognized that assumptions had to be made and taken for granted, and neither side had confidence in the other's initial assessment of what was happening. For Savitsky and Maslennikov, time was running out to aid, protect, and defend their country, and for Arkhipov, the cost of being wrong was too great to move into such sudden action, even if it turned out to be otherwise. Indeed, it was not reason that saved the day but the protocol in place that dictated what followed in the event of a stalemate, a protocol that *as a matter of chance* required unanimous agreement and not a simple majority to initiate the launch of a dangerous torpedo that had the potential to start a nuclear holocaust. Arkhipov's side did not emerge the victor in this debate except by a technicality, and it is only with the virtue of hindsight that one can see the matter much clearer than it was for those inside the submarine.

As a product of the Enlightenment faith in reason, one thing that deterrence advocates fail to appreciate is the contradiction inherent in the very notion of deterrence, a contradiction that was already present in the writings of Hobbes. The reason that peace is sought through the theory of deterrence in the first place is precisely *because* man can be anything but peaceful.[11] Much as Robespierre justified the use of terror for the sake of virtue, even though he firmly believed that people were virtuous by nature,[12] one finds in

[11] My use of the term "man" here and throughout is not intended to denote a specific gender but should be regarded as a truncated form of "humanity" and its various expressions in other parts of speech. The motivation is grounded in nothing more than a stylistic preference, as the brevity of the term strikes me as lending itself to more poetic opportunities within some of the contexts below than its lengthier parent term.

[12] In his February 1794 speech, "On the Principles of Political Morality," he proclaims:

any defense of deterrence theory that lying behind the assumption of man's rationality is the fear that he is not.

This is a fear that Hobbes explicitly acknowledges, and yet he somehow believes it to be consistent with his faith in man's ability to reason. He sees that we not only have desires, but, he worries, were we given the chance, we would do *anything* to satisfy them. This can only leave us in a natural state of constantly fearing for our security, anxious that others are just as ravenous in their pursuit of goods as we are. "Hereby it is manifest," writes Hobbes, "that during the time men live without a common power to keep them all in awe, they are in that condition which is called *war*, and such a war as is of every man against every man" (Hobbes I.13.8). Consequently, though we can all see through reasoning that having a little through peace is preferable to risking a lot through war, Hobbes believes, our inclination to war is driven in the first place by our natural passions and desires for personal gain, safety, and reputation (I.13.7). If we *think* we can have more, especially without consequence, then we *will* take more.

Whether or not Hobbes' pessimistic assessment of human nature in the absence of government is accurate, the contradiction underlying it is a product of the Enlightenment faith in reason. Rather than question that assumption—for the Enlightenment thinkers could envision no other way without forfeiting reason's fruits of scientific and political progress—Hobbes, along with his Enlightenment peers, choose instead to rationalize the absurdity or, sometimes, even just ignore it.

Is there something more to the essence of humanity than reason?

> Happily virtue is natural in the people, [despite] aristocratical prejudices. A nation is truly corrupt, when, after having, by degrees lost its character and liberty, it slides from democracy into aristocracy or monarchy; this is the death of the political body by decrepitude... (Robespierre 1794)

This is a question that they, and we, need to ask. It is a question that today can no longer be avoided. The paradox of Enlightenment thought is that the more rational we assume humanity to be, the less rational we appear. How did we get here? And where do we go?

0.6 | Fundamentally Flawed

During the crises of faith that followed from the re-emergence of Aristotelian natural philosophy in Europe and the recovery of Greek philosophical atomist literature—two pivotal events that sparked the Scientific Revolution—efforts were underway to understand man's place in this new world, a world no longer viewed as the literal and metaphorical center of the universe.[13] Though culminating in the Enlightenment exaltation of the power of human reason as a "natural light" belonging to and residing in every person, human nature was not always viewed as such. While people were regarded

[13] The intellectual development that leads up to the Scientific Revolution was by no means quick, straightforward, or clear. Philosopher William Wallace believes that the development of applied mathematics and experimentation—ideas inspired from misreadings of Aristotle's *Posterior Analytics*—that serve as the cornerstone for the Scientific Method started in the late 12th and early 13th centuries with Robert Grosseteste, Roger Bacon, and Albertus Magnus, among others (Wallace 1972, 10; 17). Stephen Gaukroger likewise argues for a similar historical development, noting that while Aristotelian philosophy paved the path forward for how an understanding of the natural world could be possible, it started to break down in light of theological criticisms. In particular, Christian intellectuals—such as Pietro Pomponazzi and Pierre Gassendi—argued that Aristotelianism provided inadequate solutions for understanding core doctrines like the personal immortality of the soul. As a result, many started to wonder "whether a philosophy completely different from Aristotle's might fit the bill" (Gaukroger 2008, 102–105). That philosophy would be Epicurean atomism. Philosopher Catherine Wilson explains:

> The experimentalists [like Gassendi and Robert Boyle] could further capitalize on the old complaint that Aristotelian philosophy was pagan through and through, by suggesting that Epicureanism was in fact easier to marry to Christianity than Aristotelianism. By repudiating Aristotle and the old philosophy of forms and virtues as heathen and idolatrous, the experimentalists established (barely and controversially) their Christian credentials and settled in their own minds the permissibility of their activities. God was given a new role as master and commander of the mindless atoms. (Wilson 2010, 67)

as rational, it was regarded as only one aspect of being human, and even still, reason itself was not valued as a perfect instrument.

According to the previous, predominantly Christian narrative in the West that was heavily influenced by Platonism and monastic practice, the human condition was that of a fractured existence, broken by the persistent, irrational appetites of the flesh. There was hope, however. It was believed that through the gift of reason and the revelations of faith, man could overcome these irrational desires, discovering in the process a sense of peace for which he longs. Historian Stephen Gaukroger explains:

> This appropriation of earlier [pagan] thought by Christianity made it possible for it to present itself as the final answer to what earlier philosophers were striving for, and we should not underestimate how successful it was in this respect. The main schools of Hellenistic philosophy had each sought to present a philosophy that transcended the flux and disorder of life and achieved peace of mind (*ataraxia* or *apatheia*). […] The Christian version of the search for peace and tranquility associates it with a state not fully achievable in this life—although monastic culture cultivated the idea of the power to be constant amid the flux and disorder of life, this was in the context of an attempt to separate oneself from the world through ascetism—but which is a reward for what one does in this life, and which relies as much upon sacramental as intellectual enlightenment. (Gaukroger 2008, 51)

The idea that the world is in a constant state of flux is one that predates Christianity, tracing back to Heraclitus. When embraced, it creates a very serious epistemological problem: if everything is in a constant state of change, how is it possible to know anything?

This is a problem that persists through Plato and Plotinus, both of whom present solutions that require positing a reality beyond the one that is experienced through sense-perception.[14] This same solution would be adopted by Christian philosophers, and this

[14] For example, in *Timaeus* 51d–52a, Socrates makes a distinction between being and becoming, arguing that intellect is oriented towards the changeless, everlasting realm of being, itself unperceived by the

new reality would be identified with God. In practical terms, the net result was an attitude of complacency toward the natural world and a focus on striving to live in accordance with Christian values. Gaukroger cites Ambrose's *Hexameron* 6.28 and Augustine's *Enchiridion* 3.9 as examples of this:

> Augustine's mentor, Ambrose of Milan, explained the absence of discussion of scientific matters in the Scriptures on the grounds that "there is no place in the words of the Holy Scripture for the vanity of perishable knowledge which deceives and deludes us in our attempt to explain the unexplainable," and Augustine himself took a similar approach:
>
> *When it is asked what we ought to believe in matters of religion, the answer is not to be sought in the exploration of the nature of things, after the manner of those whom the Greeks called 'physicists'....For the Christian, it is enough to believe that the cause of all created things, whether in heaven or on earth, whether visible or invisible, is nothing other than the goodness of the Creator.* (Gaukroger, 58–59)

And it is Augustine whom Gaukroger identifies as the central character responsible for acknowledging the value of reason for living well, so long as it is subordinated to faith (51).

In *Confessions* VII, xvii, Augustine laments that his bodily desires pull him away from his vision of God: "I wondered that, though now I loved you and not a phantasm in your place, I could not remain to enjoy my God, but I was torn away from you by my own weight, and I fell back with a sigh to these things here [*ipsa*]; and this weight was carnal habit" (Menn 2002, 198). Frustrated, he came to the conclusion that the way to overcome these passions is to attain wisdom, a culmination of intellectual, practical, and sacramental discipline. Stephen Menn elaborates that for Augustine:

> To acquire wisdom it is necessary, but not sufficient, that we have faith in the authority of Christ, that we apply the Platonist intellectual discipline to achieve an intellectual intuition of God, and that we recognize a properly theological

---

senses, while opinion is oriented towards becoming, a realm grasped by the senses and in a constant state of change.

> intellectual content in Christianity, irreducible to pagan philosophy; but beyond all this, we must also *act* in a particular way, performing outwardly visible actions as well as undergoing inward transformations, in order to make progress on the path to the *patria* [the land of God]. (203)

Thus, overcoming our irrational desires requires using reason rightly, that is, in the service of faith.[15] The upshot of this—and other premodern narratives—is that it elegantly accounts for the existence of these desires in the first place. The body is fundamentally flawed insofar as it is weak, constantly moved and tempted by the mere appearances of a world in flux. Although this is an important insight, within the premodern context it requires conceiving of the body and reason as distinct from one another. Before questioning this assumption, however, Modernity would first attempt to reject the narrative wholesale, and in so doing, elevate human reason to its highest status yet.

0.7 | Descartes' Dream

For the scientific worldview that was burgeoning, many so-called revelations, from proofs for the existence of God to the authority of Scripture itself, were challenged by experimental results and developments in natural philosophy or were scrutinized by critical reason. For instance, with the acceptance of a mechanical natural philosophy governed by necessity and determinism, it became a point of contention for

---

[15] According to Ernst Cassirer, the idea that reason needed to be tempered by and put into the service of faith even more generally applies to the *overall* spirit of the Middle Ages. In characterizing it, he writes:

> In medieval thought there remains, in theory as well as in practice, side by side with divine law a relatively independent sphere of natural law accessible to and dominated by human reason. But 'natural law' (*lex naturalis*) can never be more than a point of departure for 'divine law' (*lex divina*), which alone is capable of restoring the original knowledge lost through the fall of man. Reason is and remains the servant of revelation (*tanquam famula et ministra*); within the sphere of natural intellectual and psychological forces, reason leads toward, and prepares the ground for, revelation. (Cassirer 1932 / 1979, 40)

Enlightenment thinkers such as Hobbes, Locke, and David Hume, as to whether God could or would "change the course of nature" to perform a miracle, the supposed existence of which was traditionally used as evidence to ground belief in religion.[16] Similarly, Descartes' new theory of substance appeared to undermine the Catholic understanding of Eucharistic transubstantiation, a fear that posthumously landed his work on the Index of Forbidden Books (Buckle 2004, 251–52).

It was in this post-Renaissance intellectual milieu of the Enlightenment that a faith not in revelation but of reason began to take hold. Ernst Cassirer explains:

> "Reason" becomes the unifying and central point of this century, expressing all that it longs and strives for, and all that it achieves. […] The eighteenth century is imbued with a belief in the unity and immutability of reason. Reason is the same for all thinking subjects, all nations, all epochs, and all cultures. From the changeability of religious creeds, of moral maxims and convictions, of theoretical opinions and judgments, a firm and lasting element can be extracted which is permanent in itself, and which in this identity and permanence expresses the real essence of reason. (Cassirer 1932 / 1979, 5–6)

And inaugurating this new world is Descartes.

In a well-known anecdote recorded by biographer Adrien Baillet, it is said that on November 10th, 1619, Descartes had three dreams that were of personal importance for him. While each is interesting, it was the third dream that would prove most significant, paraphrased below:

> A book was lying on the table before him. When Descartes opened it, he discovered that an encyclopedia that catalogued the entirety of the world's practical knowledge. Though he thought it might be useful, he also took notice of a second book, *Corpus Poetarum*, an anthology of poems. Opening this one, he by chance turned to a verse by Ausonius that began, "What path shall I take in life?"

[16] See *Hume's Enlightenment Tract* (Buckle 2004, 239–41). See also: Hobbes, *Leviathan*, ch. XXXII, 257–258; Locke, *Essay Concerning Human Understanding*, IV.xvi.13; and Hume, *Enquiry Concerning Human Understanding*, 10.1.

> As he continued reading, a stranger entered his room and started to quote a different verse, beginning with the line, "It is and it is not." Delighted by the interruption, Descartes replied that he knew the poem well, identifying it as the *Idylls* by Ausonius before insisting that it should be included in the anthology he had just opened. Eager to find the poem to please the stranger, he began perusing the volume once more.
>
> While the stranger waited, he asked Descartes from where the book had come, but to his surprise, he was at a loss for words, struggling to produce an answer. Before he knew it, the anthology suddenly appeared at the opposite end of the table, side-by-side along with the encyclopedia. Although he was unable to locate the passage for the stranger, he leapt and exclaimed that he knew of an even better poem, one that began with the verse, "What path shall I take in life?"
>
> Upon reopening the book, Descartes took notice of some copperplate portraits. They seemed familiar, and as soon as he started recognizing them, the text disappeared along with the man.[17]

Mathematicians Philip Davis and Reuben Hersh say that Descartes believed "his third dream pointed to no less than the unification and the illumination of the whole of science, even the whole of knowledge, by one and the same method: the method of *reason*" (Davis and Hersh 1986, 4). This is also confirmed in Gaukroger's biography, wherein he elaborates that Descartes himself had actually interpreted the various elements of his dreams. He believed the encyclopedia represented the sciences and the anthology of poetry was a symbol for revelation and inspiration (Gaukroger 1995, 107–108). From Descartes' point of view, these dreams were a kind of divine inspiration in and of themselves, the genesis of his so-called *method*.

In 1637, eighteen years later, Descartes would go on to publish his *Discourse on the Method*, a treatise that described and defended how one should reason well; and in Part Five, he writes of his insights concerning the difference between man and animal:

> Now in just these two ways we can also know the difference between man and beast. For it is quite remarkable that there are no men so dull-witted or stupid—

[17] See Gaukroger's *Descartes: An Intellectual Biography* (Gaukroger 1995, 107). See also Davis and Hersh's *Descartes' Dream* (Davis and Hersh 1986, 4).

> and this includes even madmen—that they are incapable of arranging various words together and forming an utterance from them in order to make their thoughts understood; whereas there is no other animal, however perfect and well-endowed it may be, that can do the like. This does not happen because they lack the necessary organs, for we see that magpies and parrots can utter words as we do, and yet they cannot speak as we do: that is, they cannot show that they are thinking what they are saying. On the other hand, men born deaf and dumb, and thus deprived of speech-organs as much as the beasts or even more so, normally invent their own signs to make themselves understood by those who, being regularly in their company, have the time to learn their language. *This shows not merely that the beasts have less reason than men, but that they have no reason at all*. (Descartes AT VI 57–58; emphasis added)

The conclusion Descartes derives is that man must have a *rational* soul, one completely different in kind from whatever it is that animates the rest of the animal kingdom (AT VI 59). But he is not alone in this sentiment.

Although he holds very different metaphysical assumptions from Descartes, including a different understanding of reason itself, Hobbes reaches a similar conclusion. In *Leviathan*, he writes:

> I have said before that a man did excel all other animals in this faculty: that when he conceived anything whatsoever, he was apt to inquire the consequences of it, and what effects he could do with it. And now I add this other degree of the same excellence: that he can by words reduce the consequences he finds to general rules, called *theorems*, or *aphorisms*; that is, he can reason, or reckon, not only in number, but in all other things whereof one may be added unto or subtracted from another. (Hobbes I.5.6)

For Hobbes, like Descartes, man's ability to reason across so many diverse subjects is what sets him apart from all other creatures. He envisions reason as a kind of logical arithmetic, whereby one can add and subtract premises and conclusions (I.5.2). If reason goes astray, it is not because it is in any way faulty; it is because it was not used rightly.

On Hobbes' account, reason, much like a crude computer program, can only use what it is given, and in such cases, the problem is that the process started with bad input; or alternatively—as in cases of long chains of reasoning—the problem is memory,

specifically forgetfulness (Hobbes I.5.7–17). Believing reasoning to afford the same certainty as geometric demonstration, Hobbes confidently declares, "For all men by nature reason alike, and well, when they have good principles. For who is so stupid as both to mistake in geometry, and also to persist in it when another detects his error to him" (I.5.16)?

Locke likewise shares in this spirit, expressing it perhaps clearest of all. Although he rejects the Scholastic definition of man as "rational animal," in his *Essay Concerning Human Understanding*, he ultimately ends up defending a variation of it that makes reason even more central to our self-understanding.

First, Locke begins with what amounts to a conceptual analysis of *man*, arguing that the concept indeed refers to a particular kind of animal but does *not* necessarily refer to a *rational* one. He justifies this claim with a kind of thought experiment before supplying some empirical evidence to add weight:

> And whatever is talked of other definitions, ingenuous observation puts it past doubt, that the idea in our minds, of which the sound *man* in our mouths is the sign, is nothing else but of an animal of such a certain form: since I think I may be confident, that whoever should see a creature of his own shape and make, though it had no more reason all its life, than a *cat* or *parrot*, would call him still a *man*; or whoever should hear a *cat* or *parrot* discourse, reason, philosophise, would call or think it nothing but a *cat* or a *parrot*; and say, the one was a dull irrational *man*, and the other a very intelligent rational *parrot*. (Locke II.xxvii.8)

Although he does not supply an example of the "dull irrational *man*," it is not difficult to find contemporary cases of people struggling with—if not altogether lacking a propensity for—rational behavior, particularly in instances of brain trauma. In fact, one of the most widely known cases would occur a little more than a century and a half after the publication of Locke's *Essay*.

In 1848, an accident during a railroad expansion project in Vermont would send a tamping iron through the skull of Phineas Gage, a construction foreman. As the neurologist Antonio Damasio recounts, the physician who treated his injury, John Harlow, observed that Gage had become:

> Fitful, irreverent, indulging at times in the grossest profanity which was not previously his custom, manifesting but little deference for his fellows, impatient of restraint or advice when it conflicts with his desires, at times pertinaciously obstinate, yet capricious or vacillating, devising many plans of future operation, which are no sooner arranged than they are abandoned.…A child in his intellectual capacity and manifestations, he has the animal passions of a strong man. (Damasio 1994, 8)

And a story like Gage's is not the limit case, for modern medicine makes possible even more extreme instances of life support after massive brain damage, as in the widely publicized case of Terri Schiavo in 2005. These examples may not be exactly what Locke had in mind when he considered those who are dull or irrational, but they nonetheless support his argument: even without reason, a human animal is still a human animal.

What of a rational animal *other than* man? Locke is not only open to the possibility but believes there may have already been an example that sufficed for his purposes at the time of writing the *Essay*. Drawing upon an anecdote relayed in William Temple's *Memoirs of what pass'd in Christendom*, Locke quotes the story of Prince John Maurice's parrot in Brazil, according to which the bird appropriately and impressively responded to a series of questions asked it, as if they were being asked of any other man. Although Locke finds the story fairly credible, regardless of its accuracy, the moral still stands for him that such a parrot would be classified as a *rational animal* without being classified as a *man* (Locke II.xxvii.8).

Because he conceptually distinguishes between being *rational* and being a *man*, Locke looks to another metaphysical idea to account for rationality, *personhood*, to which he attributes many of the properties and capacities that had hitherto been assumed of the traditional concept *rational man*: "a thinking intelligent being, that has reason and reflection, and can consider itself as itself, the same thinking thing in different times and places" (Locke II.xxvii.9). Though this definition appears to place as much emphasis on reflection, self-identity, and thinking as it does reasoning, it is clear that Locke has in mind an inseparable unity between personhood and rationality, for he clarifies his use of "personal identity" as "the sameness of a rational being" two sentences later (II.xxvii.9). Furthermore, in Book IV, Locke asserts that reason "is necessary, and assisting to all our other intellectual faculties"—which presumably extends to mental activities like reflecting—and he identifies inference as an important feature of reason, explaining inference as "the perception of the connexion there is between the ideas, in each step of the deduction, whereby the mind comes to see, either the certain agreement or disagreement of any two ideas as in demonstration, in which it arrives at knowledge; or their probable connexion, on which it gives or withholds its assent, as in opinion" (IV.xvii.2). For Locke, it is through reason, particularly inference, that one can arrive at any judgments of certainty and probability.

In what is an epitome of the Enlightenment attitude, Locke's optimism about reason is so great that he expresses confidence that morality will one day be as demonstrable as mathematics (Locke IV.xii.8), contends that faith should be based on reason (IV.xvii.24; IV.xviii.1–11), and, like Hobbes, insists that reason generally only

fails us when it is not used properly, taking care to provide us with some guidelines for how to use it responsibly (IV.xvii.9–15).

Locke even brazenly argues that man is so inherently rational that formal logical training only *distracts* one from the natural ability to reason well. The target of his attack here is Aristotelian syllogistic reasoning. While he believes that such training is useful in helping us assess the validity of arguments, he insists that "the mind can perceive such connexion where it really is, as easily, nay, perhaps, better without it" (Locke IV.xvii.4). He muses:

> If syllogisms must be taken for the only proper instrument of reason and means of knowledge, it will follow, that before Aristotle there was not one man that did or could know anything by reason; and that since the invention of syllogisms, there is not one of ten thousand that doth.
>
> But God has not been so sparing to men to make them barely two-legged creatures, and left it to Aristotle to make them rational […] God has been more bountiful to mankind than so. He has given them a mind that can reason without being instructed in methods of syllogizing: the understanding is not taught to reason by these rules; it has a native faculty to perceive the coherence, or incoherence of ideas, and can range them right, without any such perplexing repetitions. […] Tell a country gentlewoman, that the wind is south-west, and the weather louring, and like to rain, and she will easily understand, 'tis not safe for her to go abroad thin clad, in such a day, after a fever: she clearly sees the probable connexion of all these, *viz*. south-west wind, and clouds, rain, wetting, taking cold, relapse, and danger of death, without tying them together in those artificial and cumbersome fetters of several syllogisms, that clog and hinder the mind, which proceeds from one part to another quicker and clearer without them: and the probability which she easily perceives in things thus in their native state, would be quite lost, if this argument were managed learnedly, and proposed in mode and figure. (IV.xvii.4)

Locke's confidence in our natural ability to reason well leads him to this bizarre mistrust of formal logic, believing that people are actually better off without any training. So convinced of his conclusion, he recalls having once "known a man unskillful in syllogism, who at first hearing could perceive the weakness and inconclusiveness of a

long artificial and plausible discourse, wherewith others better skilled in syllogism have been misled" (IV.xvii.4). To be clear, it does not escape Locke that people often *do* make the wrong inferences, but his diagnosis for such failure, similar to McNamara's, is twofold: (1) starting with faulty concepts and (2) working in haste.

Regarding the former, Locke insists that the strength of our reasoning extends only as far as the soundness of our ideas. If one reasons with empty or unclear terms, then our reasoning will inevitably be limited proportionately and fail (Locke IV.xvii.9). If his country gentlewoman, for instance, does not understand what rain is, then she will be unable to draw the inference between getting wet and catching fever were she to venture outside unprepared. Likewise, working in haste can produce similarly hazardous results by preventing one from seeing all of the relevant inferences that can be made (IV.xvii.4). Again, one can imagine the country gentlewoman, in a hurry to reach her destination, momentarily forgetting a leg of her journey and, as a result, innocently miscalculating the amount of time it will take for her to reach her goal. In either case, for Locke, it is not reason itself that is to blame. Were she to have clear ideas and enough time, reason cannot fail her.

0.8 | Children of the Enlightenment

By briefly entertaining the views of the Enlightenment thinkers above, the goal was not to suggest that it is confined to a particular time, place, or set of persons. Nor was it to suggest that Descartes', Hobbes', or Locke's views on reason itself is necessarily straightforward. Rather, the purpose of the above is to capture the spirit of the Enlightenment. Enlightenment, as those brilliant 18th century intellectuals wanted to believe, is an *attitude*, an *ideal*, a *worldview*, and it is a pervasive one. For as diverse as

the ideas of Enlightenment thinkers are, the one strand that connects them is a faith in reason. It is a spirit that is, to this day, alive and well. This faith, I have argued, is born out of an unquestioned, unexamined basic assumption that we are somehow fundamentally, necessarily rational, and it is this assumption that has had grave consequences for us.[18]

The assumption of Enlightenment creates a familiar problem for those who adopt this kind of position. How does one account for error? To answer this question is to get a little closer to understanding why so many Enlightenment thinkers trusted in the universality of reason while fearing the behavior of humankind when left to its own devices. If we are, by nature, virtuous and rational, and if reason is as powerful and perfect as we want to believe, then why do we still go astray?

One of the more extreme variations of the Enlightenment attitude appears in the form of a movement known as *psychologism*, according to which not only is the mind rational but even the principles of logic are abstractions from how the mind in fact psychologically functions.[19] On this view, a deductive inference such as a disjunctive

---

[18] Pinker briefly pushes back against this idea, arguing that it is a mischaracterization. He explains:

> Many writers today confuse the Enlightenment endorsement of reason with the implausible claim that humans are perfectly rational agents. Nothing could be further from historical reality. Thinkers such as Kant, Baruch Spinoza, Thomas Hobbes, David Hume and Adam Smith were inquisitive psychologists and all too aware of our irrational passions and foibles. They insisted that it was only by calling out the common sources of folly that we could hope to overcome them. The deliberate application of reason was necessary precisely because our common habits of thought are not particularly reasonable. (Pinker 2018, 8–9)

The problem, however, is that very few philosophers have *ever* considered humans to be *perfectly* rational agents, and so this amounts to a strawman of his opposition. In spite of acknowledging that people can tend to be unreasonable, Pinker still defends the Enlightenment idea that reason can overcome error and irrationality if one simply applies it *well enough*.

[19] As Martin Kusch explains, the term *psychologism* as it is used here follows from a vigorous debate in intellectual circles at the turn of the 20th century in Germany, the *Psychologismus-Streit* ("psychologism dispute"). Today, there are other uses of the term that depart from this understanding (Kusch 2015).

syllogism is merely one pattern of thinking that has been identified, abstracted, and formalized. The question then remains as to whether such thinking is indicative of how we *actually* think (*strong psychologism*) or how we *ideally* think (*weak psychologism*), a distinction that philosopher Susan Haack makes.[20]

The philosopher Gilbert Harman, for instance, openly identifies his own position on this matter as psychologistic and claims that "the valid principles of inference are those principles with which the mind works."[21] But as philosopher Martin Kusch worries, the standard response to this is that such a position "makes it impossible to account for invalid reasoning" (Kusch 1995, 9). To be sure, such poor reasoning not only occurs, but overwhelmingly so within certain contexts.

In addition to numerous psychological studies demonstrating the frequency of poor reasoning in the general population, cognitive scientists Hugo Mercier, Guy Politzer, and Dan Sperber have conducted their own research that highlights this as well.[22] What they discovered was a correlation between how a logical problem is *contextualized* and the number of correct responses, taking care to leave untouched the basic logical *structure* of the problem.[23] For example, when drawing upon the Pigeonhole

---

[20] See *Philosophy of Logics* (Haack 1978, 238).

[21] See *Thought* (Harman 1973, 18). See also Kusch, *Psychologism* (Kusch 1995, 9–10).

[22] See "What causes failure to apply the Pigeonhole Principle in simple reasoning problems?" in *Thinking & Reasoning* (Mercier, Politzer, and Sperber 2017). See also *The Enigma of Reason* (Mercier and Sperber 2017, 21–22).

[23] Anecdotally, I have made similar observations in my own courses. When lecturing on the fallacy of *denying the antecedent* to a class, for example, I will first present the following two logically equivalent arguments side-by-side:

Principle—a mathematical principle that states that whenever there are more objects than categories, if all objects are sorted into categories, then at least one category will contain more than one object—Mercier, Politzer, and Sperber found that only 30% of respondents answered correctly when the problem was presented primarily as a numerical one while 70% answered correctly when it became more contextualized.[24] If strong psychologism is correct, the mind should not be failing to draw the correct inference so easily, regardless of context.

Fortunately, for those who wish to defend Enlightenment, it could be argued that the thinkers above—Descartes, Hobbes, and Locke—have a lesson to teach and that one ought to adopt a position closer to weak psychologism. After all, those philosophers believe that the mind indeed thinks rationally but also understand that it obviously falls into error, and so it stands to reason that a discipline such as logic must codify an *idealized* form of thinking. The antidote to poor reasoning, as each suggests, is that you need only follow some tried and true guidelines, proceeding with caution while doing

---

P_1 If a team scores the touchdown, they will win the game
P_2 The team did not score a touchdown
C The team did not win the game

P_1 If you win the lottery, you will be wealthy
P_2 You did not win the lottery
C You are not wealthy

There is a consistent and significant difference between the number of students who believe that the sports-contextualized argument is *not* a fallacy while the wealth-contextualized argument *is*.

[24] Although in *The Enigma of Reason*, Mercier and Sperber do not clarify their observations, in the abstract of their published paper, "What causes failure to apply the Pigeonhole Principle in simple reasoning problems?," they speculate, "The failure to apply the Pigeonhole Principle might be due to the large numbers used, or to the cardinal rather than nomial presentation of these numbers" (Mercier, Politzer, and Sperber 2017). Of these two possibilities, the former is less likely given that one of the presentations of the logical problem asked a question about only twenty-two farmers, and respondents still tended to fail (Mercier and Sperber 2017, 23–24).

your best to control your nature or emotions so that you can think through matters with patience and determination.

Comparable recommendations in defense of Enlightenment rationality are even made today. Mercier and Sperber, for instance, identify *mental logicians* and *mental modelers* (Mercier and Sperber 2017, 25). The former believe that the mind has a set of logical rules that it uses to make relevant deductions while the latter believe that the mind instead takes advantage of something like logical blueprints or schemata, similar to how one might use and read a Venn diagram. Unlike strong psychologicists, both mental logicians and modelers allow for error whenever logical tasks increase in complexity. So, like the Enlightenment thinkers, mental logicians believe that error arises as the number of steps and rules increases to solve a problem, and the mental modelers similarly believe that error results more frequently as more mental blueprints are required for the task at hand (26). For as appealing as the three positions (weak psychologism, mental logic theory, and mental model theory) might at first glance seem, they still fail to address why there is a tendency to fall into error in the first place, and why, for that matter, fallacious thinking appears to be much more commonplace than sound reasoning.

Citing the work of cognitive psychologist Jonathan Evans and cognitive scientist Ruth Byrne, Mercier and Sperber highlight how regularly people struggle to make the correct inferences when presented with arguments that use conditional statements. While Evans' work demonstrates that most people are able to make a *modus ponens* inference when presented with a problem, he also discovered that good reasoning declined from there.[25] Mercier and Sperber summarize his results: "only two-thirds of the people, on

[25] See *Bias in Human Reasoning: Causes and Consequences* (Evans 1989).

average, draw the other valid inference, *modus tollens*, and about half of the people commit the two fallacies [of affirming the consequent and denying the antecedent]" (Mercier and Sperber 2017, 28). What is worse, Byrne shows how even *modus ponens* can prove difficult for most.[26] In her experiment, she found that while most participants were able to deduce the correct conclusion from a simple two-premise *modus ponens*, when a third premise was added, only 38% succeeded in making the inference (28–29).

Mercier and Sperber caution that while these studies raise interesting questions about the nature of rationality and the mind, it does not follow that the case is settled against weak psychologism, mental logicians, and mental modelers. For one, people tend to be highly sensitive to context, and so some of the studies might be indicative of the fact that, in order of priority, people tend to place more decision-making weight on what seems plausible as opposed to what logically follows (30). Objections can also be raised as to how these studies are conducted and how the problems are framed.

Setting aside these important questions, what these three positions have in common is their Enlightenment faith in reason. Each holds the conviction that the mind operates rationally, and each locates the source of error in something external to the mind (e.g. situations are too complex, ideas are too vague, emotions are too strong, scenarios are too implausible, problems are too involved, in short, anything other than the workings of the mind is to blame). It should come as no surprise then that such a conviction not only faces difficulty in accounting for error, but also struggles to explain seemingly irrational behavior.

[26] See "Suppressing valid inferences with conditionals" in *Cognition* (Byrne 1989).

0.9 | The Mighty Pen and the Marvelous Drink

Imagine for a moment that, while driving home from work, a brand new establishment catches your eye. Everything about it is appealing—the Goudy Old Style font of the sign out front that reads "Vertigo," unironically punctuated with a splash of playful confetti-like markings; the casual, yet sophisticated blend of vintage red brick and neon lighting, as if it had just materialized from the 1980s. Interest piqued, you decide to pay a visit.

As you waltz in, you notice a few patrons sitting at the bar as well as a couple engaged in a serious political conversation. "Must be on a first date," you think to yourself. You approach the bar and the bartender invites you to take a seat. "What'll you have?," he asks. As you feel the pressure to make a decision fast, he fortunately interrupts, explaining:

> This is a craft brewery, you know? We make all of our own. We've got a lot planned, but since we're new, there's only three offerings right now. There's our version of a west coast IPA, pretty hoppy and fruity, but we experimented a little to contrast those flavors with some citrus tang. We also got an Oktoberfest if you're looking for something on the maltier side. Smooth finish, velvety texture. It's like sipping on melted, dark caramel. And then there's the coffee stout. We use this quality cold brew from our friends up the street. It's dark, slightly bitter, in a cacao kind of way, and hits right at home. Here, let me pour you a quick sample.

Relieved that you will be making neither a spur-of-the-moment nor uninformed decision, you politely take him up on his offer.

After sampling each drink, you feel like you've made up your mind. "The IPA," you say. To make sure he understands, he double-checks, "The IPA?" You nod with a smile.

Are you certain of your decision? Obviously, as you sampled each drink, you sorted out the qualities that you liked and did not like, and while taking everything into

consideration, you reasoned that it was the IPA that satisfied your current desires most. Right?

Now imagine for a moment that just before the bartender pours your drink, he pauses, lifts his index finger near the side of his head, and says, "Wait! Just remembered!" He quickly scrambles to the back, disappearing briefly before returning with another small glass. "Try this!," he excitedly proclaims. He explains to you that this is their latest beer, a Vienna-style lager. They *just* finished it this morning and have not even had a chance to put it on the menu.

Given this situation, any rational person would surely find herself now deciding between the drink she originally selected and this new one. But, to the bartender's surprise, after a few more people saunter into the establishment and sit down, you look to your neighbor, smile, and say, "On second thought, let me get a glass of that Oktoberfest!"

How in the world could this possibly make sense? Surely this example is an outlier, right?

Psychologists Toshio Yamagishi, Hirofumi Hashimoto, and Joanna Schug had set out to explore the relationship between preferences and decision-making in a variety of different contexts, undertaking efforts to replicate the results of a previous experiment conducted by cultural psychologists Heejung Kim and Hazel Rose Markus.[27] In one of their studies, Yamagishi and colleagues presented participants with a scenario in which they had to decide between one unique pen and four common pens, the latter differing only in color but otherwise identical. The twist is that in one scenario, they were asked to

[27] See "Deviance or uniqueness, harmony or conformity? A cultural analysis" (Kim and Markus 1999).

imagine that they were the *first* of five people to select a pen, and in another, the *last* person to select one of five pens (Yamagishi, Hashimoto, and Schug 2008, 580–1).

So do people tend to like unique pens or common ones?

It depends on the context. When expected to make a selection first, about half of the participants selected a common pen, but when they were selecting last, over 70% showed a preference for the unique pen (Yamagishi, Hashimoto, and Schug 2008, 582). The takeaway, argues evolutionary psychologist Robert Kurzban, is not only that context affects decision-making, but that our decision-making is *highly sensitive* to changes in context (Kurzban 2010, 155). He continues:

> If I find you like this pen over that one [given the context], I can no longer say you have a preference over that set of choices. All I can say is that you have such a preference *in the context in which you did the choosing*. And it's even worse than *that*. What, exactly, *is* "the context"? Without a theory about which particular aspects of the choice matter, I can't even say what your context-specific preference is because I can't define the context. As Yamagishi and others have shown, the context can be any number of non-obvious factors, like the presence of a certain number of other pens. And with each contextual variable that matters, surely we're asking a great deal of our "preference books," specifying what is preferred to what across all contextual variables that surround a choice. (155)

As he contends, it is beginning to appear that our decision-making is so sensitive, in fact, that it raises serious questions as to what even counts as a context, threatening to throw into disarray whatever conceptions we have about preferences and reasonable behavior. If the Enlightenment assumption of the way in which reason operates and exercises its influence on us were correct, it should not be this complicated to sketch out a reliable theory of human action and choice. And yet, as Robespierre discovered the hard way, reality does not always conform to our expectations, and what we *want* to be true sometimes is not.

0.10 | Plumbing the Depths

The impulse of the day, as it was in 1793, is to deny that there is anything wrong with human reason or the Enlightenment understanding of human nature. If apparent irrational behavior is acknowledged at all, then it is viewed as a problem to be fixed through education, medication, or, as Hobbes and Robespierre suggested, even force.

Is this what is best? Should we, ironically, be resisting progress in order to get *back* to the Enlightenment? Should we, in other words, continue with our neurotic obsession in the Enlightenment ideal of human reason while explaining away (as reason is wont to do) its failures, thereby holding ourselves hostage to a movement that promised us prosperity, equality, and liberation? Is there no path beyond Enlightenment?

The Enlightenment is wise to sing the praises of our scientific institutions and what they have been able to accomplish in such an historically short amount of time. While it often hides its usefulness behind the success of the scientific enterprise, that enterprise itself has furnished us with inklings of what a world beyond Enlightenment might look like. To that end, this is one attempt to paint that picture, focusing primarily on how we might understand human nature in a way that strikes at the heart of the Enlightenment's most cherished assumption: the power and purity of human reason. The goal is not to recommend that we turn to the past, to whatever age before Enlightenment, except when such insights prove useful and complement what we know today regarding who we are. To that end, the result is thus a coordination of ideas that, hopefully, leave us with a new set of ideals that not only really can make our lives better but that prove attainable in practice as well.

In the service of that goal, this work is divided into three parts: action, irrationality, and theory of mind.

In an effort to loosen the grip that the Enlightenment's adoration of reason has on our understanding of action, part one proposes a theory that significantly deviates from the traditional philosophical conception of an action as a performance that is undertaken for a reason. Working from the perspective of a post-Darwinian world, it hopes to show that action must be understood far broader than this. Not only is it the case that non-human organisms have a right to be described as undertaking actions, I argue that human action fails to make sense otherwise, for then it appears to be an anomaly in the natural world.

To make this argument, I turn to some of our discoveries in ethology, cognitive science, and neurobiology, showing that there is good reason to suppose that the biological continuity that exists between organisms lends itself to the assumption that there are cognitive and behavioral similarities as well. Such discoveries should force us to reconsider how we think about human action. Reason, I argue, is simply not nearly as relevant a component for action as we have been lead to believe.

With a robust theory of action in place, part two turns to a series of studies on irrational thought and behavior. If the Enlightenment had struggled to account for error, it is all the more hopeless in the face of irrationality. Is such thought and behavior a mere aberration from our presumably otherwise rational lives, or is it far more common than we think? I advocate for the latter view, believing that the pre-Modern narrative deserves credit for getting this part of human nature right—even though it may have failed in other respects, such as what we can do about it. To hold an Enlightenment standard of human

nature high and expect us to conform to it always or even most of the time is as unreasonable and unrealistic as it gets. If the incorruptible among us do not wind up making the same decisions as Robespierre, the rest of us confused mortals will be stuck resenting ourselves for constantly falling short, an outcome that is neither healthy nor productive.

While an entire book could be written on closely analyzing the numerous candidates for inclusion amongst the pantheon of irrational thoughts and behavior, the selection taken here was more strategic in nature, focusing primarily on classic phenomena that are oft-discussed and oft-confused, specifically wishful thinking, self-deception, and *akrasia*, in chapters four, five, and six, respectively. In addition to establishing how frequently these phenomena affect human life, a secondary goal sets out to understand what makes irrational phenomena *irrational* in the first place.

The error of the Enlightenment on this topic has been to assume that irrationality is best understood, if at all, as a deviation from what is rational. On the contrary, I hope to show that such irrationality not only *can be* but *should be* understood on its own terms. If we are getting human nature wrong, then it behooves us to understand how and in what way, and subsuming what is potentially an integral part of who we are under the umbrella of a trivial deviation from rationality is hazardous. Furthermore, if we can begin to understand irrationality and its role in human nature on its own terms, we put ourselves in a better position to identify forms of it that have hitherto gone unnoticed. Chapter seven puts this to the test, focusing on a phenomenon known as *negation*.

If human action does not depend on reason and human behavior is often, in spite of our best efforts, irrational, what resources do we have available to explain any of this?

This is the topic of part three. It is here where I call upon some important developments in evolutionary and cognitive psychology, defending a theory of mind known as *modularity*. This, I believe, is the best theory that we have available for understanding the strangeness of human nature in all of its glory and frailty. It can account for a non-rational account of action as well as irrational behavior while proving itself to be importantly consistent with scientific discoveries.

It is this view of human nature, one tempered by scientific insights and philosophical ideas, that puts us in the best position to flourish and move forward once more, hopefully this time without our own version of *The Terror*.

Part I. Life Along the Biological Spectrum

It was an ingenious and yet relatively simple experiment for those with access to laboratory settings. An alcove was nearby whose sole function was to provide a subject with food, provided that the right conditions were met. Access to this food was triggered whenever a subject interrupted an infrared beam, not too dissimilar from how an electronic door operates. The entire process was automated, allowing experimenters to focus their attention on observing what happens. But there was a catch for the subject: the infrared beam was not always active. In order to arm it, there was a lever that the subject needed to press, not once, not twice, but some predetermined number of times, varying in range from four to twenty-four depending on the experimenters' preference. What made the task even more difficult was that if the subject had prematurely checked the alcove, the beam would remain disabled for ten seconds as a penalty, even if the lever had been pressed the correct number of times. There was no penalty if a subject pressed the lever more than was necessary (Gallistel and Gelman 2005, 580).

This experiment was first devised by the psychologist Francis Mechner while in pursuit of his PhD at Columbia University in 1958. Building on Mechner's work in 1971, psychologists John Platt and David Johnson ran their own variation of the experiment with rats and took note of something rather fascinating. Once they had a chance to figure out what was required of them, the rats could consistently press the lever approximately the correct number of times needed to activate the infrared beam, usually hitting the target number but sometimes going slightly over (Gallistel and Gelman 2005, 580). The type of numerical recall required to perform this activity is what psychologists refer to as *remembered numerosity*, suggesting that these types of creatures had some way of

cognitively working with numbers. What is more, the researchers also observed a curve. The larger the number that needed to be remembered for the lever, the greater the variability of total presses, a phenomenon known as *scalar variability*. A rat that figured out that the lever needed to be pressed four times might press it four or five times, but if it needed to be pressed twenty times, the total pressings could be anywhere between twenty and twenty-seven. Different versions of this experiment have been repeated with pigeons and monkeys as well (580).

So what is happening here? Remarkably, according to psychologists Charles Gallistel and Rochel Gelman, these nonverbal creatures have a way of thinking about and acting on number; that is, they are capable of a form of *mathematical cognition*. And yet, this form of thought, which influences their behavior, somehow does not depend on language but rather something else to work with the quantities in question (Gallistel and Gelman 2005, 580).

In philosophy, the topic of agency is often linked to questions of personhood, responsibility (such as epistemic, moral, or legal), and self-identity. Within the battleground of the field known as action theory, metaphysicians, ethicists, and philosophers of mind debate whether there is a *who* that underlies actions, what it means to be held accountable for having done something, and how we can understand subjectivity, consciousness, or even selfhood. But even more fundamentally, perhaps *the* central question that motivates action theory is this: what does it mean to *do* something, something free of external compulsion and the blind, brute mechanisms of determinism? The sipping of one's tea seems to be something very different from the falling of the boulder down the hillside. What distinguishes the one from the other? Whether action

theory can provide any insight into the problems of metaphysics, ethics, or philosophy of mind depends entirely on how one develops an answer to this question.

In his essay, "Animal Minds," John Searle argues that animals have consciousness, intentionality, and thought processes on the bases of behavioral inference and biological continuity (Searle 1994, 206; 208). That some kinds of animals can act in ways similar to humans—at least in certain situations—is clear (216–17). When someone surprises a pet, such as a cat or dog, its eyes often grow wide before jumping or running away, much as one might expect from a person who was startled. Those who believe that animals have psychological states and experiences might argue that being able to identify these behavioral analogues warrants making the inference that the animal in question is having a similar experience to human beings. But critics are dismissive, believing animals to be little more than sophisticated machines, programmed to act by the mechanisms of operant conditioning and instinct.

Searle concedes that behavioral inference is uninteresting on its own and insufficient to make any convincing claims regarding the capacities of animals, but he is not convinced that nonhuman animals are unable to experience their environments. Instead, he argues that, in addition to inferences from behavior, the structural and neurobiological continuity between humans and the rest of the animal kingdom gives us every reason to believe that the causal capacities and experiences of animals are at least similar to our own in important ways (Searle 1994, 217). He explains, "I know that my dog has a certain inner causal structure that is relevantly similar to my own. I know that my dog has eyes, ears, skin, etc., and these form part of the causal bases of his mental life, just as similar structures form part of the causal bases of my mental life" (217).

Searle's biological continuity argument is not without support elsewhere. Cognitive scientist Douglas Hofstadter, for instance, has made comparable arguments of his own. In trying to understand consciousness or "I"-ness, Hofstadter proposes that, rather than treat consciousness as a categorical property that creatures either have entirely or lack, it is best understood as a matter of degree, as something proportionate to cognitive sophistication. At one end of the spectrum are mosquitoes, whose behavior can be understood largely in terms of reflexes and whose neurological system is unlikely to be complex enough to symbolize anything, and though human beings occupy the other end of the spectrum, Hofstadter insists that there is plenty of room in the middle (Hofstadter 2008, 77–9). He explains:

> I think it's obvious, or nearly so, that mosquitoes have no conscience and likewise no consciousness, hence nothing meriting the word "soul". These flying, buzzing, blood-sucking automata are more like miniature heat-seeking missiles than like soulful beings. Can you imagine a mosquito experiencing mercy or pity or friendship? 'Nough said. Next!
>
> What about, say, lions—the very prototype of the notion of carnivore? Lions stalk, pounce on, rip into, and devour giraffes and zebras that are still kicking and braying, and they do so without the slightest mercy or pity, which suggests a complete lack of compassion, and yet they seem to care a great deal about their own young, nuzzling them, nurturing them, protecting them, teaching them. This is quite unmosquito-like behavior! Moreover, I suspect that lions can easily come to care for certain beasts of other species (such as humans). (348)

Hofstadter suggests that this selective display of compassion in some cases and not others is indicative of a mental life that is rich enough to symbolize some creatures *as* prey and others *as* kin, and he even references an anecdote of a vegetarian lion named "Little Tyke" who seemed to take joy in playfully engaging with creatures that typically make for tasty meals, such as lambs and chickens (348).

While the focus over the next three chapters is not on consciousness, these observations and arguments by Searle and Hofstadter inspired the following analysis of agency. If one assumes the importance of biological continuity in trying to understand what it means to *do* something, then how does that impact an analysis of agency itself?

What emerges is a richer and fuller conception of agency that respects our biological kinship with the rest of the animal kingdom by viewing agency, too, as a matter of degree. Such a view can elevate the ontological status of our fellow creatures, but it need not do so by diminishing or trivializing what it means to be human. Supposing then that such biological continuity exists—a supposition that looks increasingly plausible in light of ethological, genetic, and neurobiological research—the following account in chapter one explores what is conceptually necessary to be able to be counted as doing anything free from external compulsion. If biological continuity is correct, then one should expect that such capacities increase in sophistication as creatures increase in biological complexity.

But what is wrong with action theory in the first place? In what way has the tradition lead us astray? Answering these questions requires a quick detour through the branch of philosophy known as contemporary metaphysics, which E.J. Lowe eloquently describes as a discipline concerned with "*the fundamental structure of reality as a whole*" (Lowe 2002, 2–3). To say that metaphysics is concerned with the fundamental structure of reality is not to suggest that the scope of metaphysical inquiry is coextensive with physics or limited to investigating the nature of *empirical* reality, which consists of everything from pencils and books to atoms and fields. Metaphysics extends to other

forms of reality as well, such as *mental* or *social* realities, and so metaphysicians might investigate things such as *psychological states* or the nature of *causation*.

Broadly speaking, contemporary metaphysicians want to know three things. One, what is the best way to understand reality as a whole? For example, is there *just* empirical reality, everything reducible to it, or is something like mental reality indispensable for a complete understanding of the world? Two, what is the best way to classify the things and properties that exist? In other words, does something like time only belong to empirical reality, mental reality, or both? Three, how do these things relate to one another? Much that falls under the third question belongs to an analysis of causation. By asking these questions, metaphysicians develop and refine ideas that help us better understand reality, and when it comes to action theory, it has furnished the field with three important concepts: *actions*, *happenings*, and *events*.

Chapter 1: The Thin, Biological Red Line

> It will not help to try to imagine that one has webbing on one's arms,
> which enables one to fly around at dusk and dawn catching insects in one's mouth;
> that one has very poor vision,
> and perceives the surrounding world by a system of reflected high-frequency sound signals;
> and that one spends the day hanging upside down by one's feet in an attic.
> In so far as I can imagine this (which is not very far),
> it tells me only what it would be like for *me* to behave as a bat behaves.
> But that is not the question.
> I want to know what it is like for a *bat* to be a bat.
>
> — Thomas Nagel, "What Is It Like to Be a Bat?"

1.1 | A Prolegomenon to Any Current Metaphysics Needed to Understand Action Theory

It is commonly believed that actions are a particular kind of *event*, which is a fundamental metaphysical concept that helps theorists identify and analyze changes in the world around us. Everything from the picking up of a glass to the smashing of an atom counts as events. Any change in the world that can be isolated and identified can be singled out as an event. Indeed, though events typically involve interactions between things, *the* feature of reality that defines them is that of change, whether a change in property, kind, time, or even location. Event ontologists—metaphysicians who study the nature and reality of events—strive to understand primarily how this change comes to be by studying causation itself.

Even though all actions are events, it does not follow that all events are actions. While the picking up of a glass by Tom almost certainly qualifies as an action, the falling of a tree by an earthquake almost certainly does *not*. To help keep these two types of events distinct from one another, there is an additional metaphysical concept known as a

happening. Whereas the picking up of a glass is an event that is caused by a human being with a goal in mind of doing precisely that—picking up the glass—the falling of a tree just happens, occurring as a result of necessity or accident. Both happenings and actions thus comprise two exclusive metaphysical categories into which metaphysicians and action theorists sort events, and what marks the difference is both whether an *agent* is involved and whether the agent played a pivotal role in causing the event to occur.

Jane's picking up of the glass, for example, cannot happen without her reaching out, grasping the glass, and lifting it. By contrast, the falling of a tree can occur without any creature doing anything at all. Everything from a strong wind to a bad case of rot could have caused it to happen. What these two examples help illustrate is that actions are precisely those events that could not have occurred without an agent contributing something of her own that ultimately causes it. In other words, actions are events that centrally feature agents as *doing* or *causing* something to happen. But what is it that an agent contributes? How we are to understand her role in causing the event to occur? And above all, what does it mean to be an agent? These are some of the most important questions in action theory.

1.2 | The Pitcher and the Crow

There has been a tendency, an Enlightenment holdover, to associate the very concept of action with any performance by an agent that is done *for a reason*, otherwise known as *intentional action*. One of the more contemporary defenses for this position can be found in Davidson's essay, "Agency," in which he remarks that "a man is the agent of an act if

what he does can be described under an aspect that makes it intentional."[28] If agency is regarded this narrowly, however, it seems to ignore the space between happenings and actions. How does one, for example, adequately account for performances by non-rational subjects? Should they be regarded simply as happenings?

In a well-known fable attributed to the ancient storyteller Aesop, a thirsty crow had come upon a pitcher with just a small amount of water inside. The opening of the pitcher—too tall and too narrow—proved an obstacle for the crow, desperate for a drink. After trying and failing multiple times to take a sip of water straight from the pitcher, it occurred to the crow that the volume of the water inside could be displaced by dropping small pebbles into it. With each pebble dropped into the jug, the water level began to rise until the crow was finally able to quench its thirst.

Though "The Crow and the Pitcher" was often thought to be a mere moral, exhorting listeners to never give up, scientists took the story one step further, developing the now widely used "Aesop's Fable Paradigm," an experiment designed to test causal understanding and goal-oriented behavior. In 2009, zoologist Christopher Bird and biologist Nathan Emery explained how they presented four rooks—a species of birds known as corvids, which comprise the crow family—with a tall, narrow cylinder. Inside was a shallow pool of water, and floating on top was a worm. Placing a handful of stones nearby, Bird and Emery wanted to test how the rooks would react, running three different experiments to control for important variables. The first introduced the cylinders with varying amounts of water; the second introduced two cylinders, one with a worm and one without; and the third introduced two cylinders, each with their own worms except one

[28] See "Agency," in *Essays on Actions and Events* (Davidson 1971, 46). I discuss the idea of intention in more detail in section 1.9, but also address Davidson's view specifically throughout chapter two.

used sawdust in the place of the water. To their surprise, after an initial period of experimentation, the birds succeeded in every version of the task, using exactly the number of stones needed to reach the worm, preferring larger stones to smaller ones, and losing interest in the sawdust rather quickly (Bird and Emery 2009, 1410–11). As they explain:

> The results of these experiments provide the first empirical evidence that a species of corvid is capable of the remarkable problem-solving ability described more than two thousand years ago by Aesop. What was once thought to be a fictional account of the solution by a bird appears to have been based on a cognitive reality. (1411)

Now, if theorists determine that actions must be wed so closely to reasons, then it follows that the behavior exhibited by these rooks must, by definition, be excluded as actions, or else one has to be prepared to attribute reasons to birds. At the same time, it fails to do justice to their creativity and innovative tool-use to regard such behavior as a mere happening, as if it were mechanically determined by external forces like the rotting of a tree. This kind of steadfast division between happenings and intentional actions can skew debates about the nature of agency, such as whether animals or infants have agency, forcing us to choose between a position that takes the subject in question to be rational or one that takes the subject to be something like a biological machine whose behavior can be explained in terms as simple as a combination of reflexes and brute conditioning. The latter possibility is actually wholly rejected by Bird and Emery:

> There is some reason to suggest that the behavior was not *solely* a conditioned action: multiple acts of stone dropping were necessary for success (in previous experiments, one stone had been necessary for success), and subjects did not try to reach for the reward after dropping each stone. In addition, they reached for the worm from the top of the tube (see Movie S1) rather than checking at the base (in previous experiments, the worm was accessible below the tube). (1411; emphasis added)

Something more is happening in these kinds of cases, and these findings have been replicated with Caledonian crows, young children, and even raccoons (but not with Western scrub-jays, who were instead uninterested in participating).[29] How is one to understand the cognitive capabilities of animals such as these, exhibiting behavior too complex to be regarded as mere automated responses? If agency is a categorical concept that turns on reason, then it does not seem possible, but if instead it is treated as a matter of *degree*, these issues can be resolved. By proposing the thesis of biological continuity, Searle pointed us in the right direction to understand how this is possible.

So if there is biological continuity, then what sorts of features might make agency possible? Such features should be found, presumably in rather diminished and minimal senses, in some of the simplest organisms to help account for how they navigate their environments. Where then does one begin?

1.3 | Clipping Pinocchio's Strings

One of the philosophical purposes of determining agency is to help make sense of responsibility. How does one assign responsibility to agents for undertaking the performances that they do? Responsibility itself is a rather complex idea. While there are fuller notions of it, such as *moral* or *legal* responsibility, the kind most appropriate for a rudimentary form of agency is simply *causal*, which seeks to determine the primary cause of an event. Causal responsibility, after all, is a precondition for the other, fuller kinds. Concepts such as praise and blame, relevant to moral responsibility, presuppose that one

[29] See "Adaptation of the Aesop's Fable paradigm for use with raccoons (*Procyon lotor*)" (Stanton et al. 2017). See also: "Western scrub-jays do not appear to attend to functionality in Aesop's Fable experiments" (Logan et al. 2016); "How do children solve Aesop's Fable?" (Cheke et al. 2012); "Using the Aesop's Fable paradigm to investigate causal understanding of water displacement by new Caledonian crows" (Jelbert et al. 2014).

can identify a subject deserving of the praise or blame by sufficiently causing the event in question to occur in the first place. The same is true of guilt and innocence within legal contexts. Thus, sorting out the prerequisites for a specific kind of causal responsibility is needed.

When seeking what is needed for causal responsibility appropriate to agency, a good starting point is to begin with the question, "Who (or what) did this?" The goal is not to identify anything that counts as a person—something that emerges much further down the spectrum of agency—but rather, the goal is to locate and understand where a happening ends and an agent begins. This agent is precisely the kind of subject that falls short of intentional agency while also being insufficiently accounted for by brute mechanical processes. In other words, this subject must be acting free from external compulsion. But what is it that enables subjects to do this?

Assigning agency requires more than just identifying the mere cause of an event's occurrence. In domino effect fashion, subjects sometimes cause an event to occur incidentally. For example, the drunkard who is thrown out of the saloon is, in a sense, the cause of the breaking of the post outside, but it was the inertia from being tossed by a surly bartender that is responsible for the drunkard's breaking of the post. The breaking was not the result of anything internal to the drunkard. In order to distinguish incidental causings from actual causings (the kind required for agency), it is a precondition when assigning causal responsibility to the subject in question that she not be externally forced to act in any obvious way. Rather than act as a puppet does on marionette strings, it should be the case that, as far as an observer can tell, the cause is internal to the subject.

Starting with an obvious and advanced case, imagine that Smith reaches for his remote control, looks at the "power" button, presses it, and the television screen lights up. Unless one is committed to the idea that every event in the universe proceeds necessarily and invariably from some prior event—a variation of determinism known as *event causation* that denies the existence of agents—very few would deny that Smith counts as an agent when he turns on his television. After all, it did not power itself on, and, counterfactually, it would not have powered on had Smith not picked up the remote control and pressed the "power" button. There was nobody else present who forced Smith to do this. Indeed, for this event to have occurred at all, Smith needed to participate in a particular button-pressing activity that caused it to happen. Did this performance proceed from external compulsion? Not in any obvious way.

In the Serengeti, a lion carefully watches from a nearby shrub as a half-dozen zebras trot less than a hundred yards away. Seizing the opportunity, the apex predator quickly emerges from its hiding spot, ferociously galloping in pursuit of its prey. Alerted by the sound, the startled zebras attempt to flee, some managing to do so with greater ease than others who clumsily bump into one another. The lion keeps pace, as if assessing the many opportunities before it by calculating what might make for the greatest meal with the least effort. It now spots its victim, momentarily exerting maximum effort to close the distance before leaping on the zebra's back. The zebra struggles to free itself, bucking and kicking, but the strength of the lion is too much to overcome. The lion has its meal and knows it.

It is beyond dispute that the lion was the cause of the killing of the zebra. What is less clear is how this differs from both Smith's powering of the television and the rotting

of the tree. Could the killing of the zebra have occurred without the lion participating in the activity of hunting? Most certainly not. Furthermore, the lion, like Smith, seemed to be acting free from external compulsion. And yet, though there are some, few action theorists would be willing to count the lion as an agent.

More controversially still, imagine walking outside on a warm, humid summer's night. For a brief moment, like a higher-pitched dentist's drill, an annoying buzzing can be heard in your right ear. You swat in the air near your head, and the sound goes away. A few seconds later, you happen to look down at your watch, concerned about the time. Out of the corner of your eye, you notice a small, winged insect on your upper forearm. It raises its two back legs, lifts its head for a brief instant, and then plunges its spear-like proboscis into your skin right before you feel a slight, sharp pinch. You swat once more, this time aiming for your forearm, and as you pull your hand back, all that is left is a tiny, irritated patch of red skin. The mosquito has successfully drawn your blood.

It is true that the drinking of your blood was an event caused by the mosquito, and yet this too seems different from Smith, the lion, and the rotting tree. While the tree is not responsible for its own rotting—after all, it could not in principle have done otherwise—each of Smith, the lion, and the mosquito *are* responsible for causing the respective events. And like Smith and the lion before it, the mosquito appears to be acting free from external compulsion.

In the passage of Hofstadter's quoted in the introduction to part one, he likened mosquitoes to automata and heat-seeking missiles, believing that their behavior can be explained with reference to little more than complex reflexes, but he also conceded that something noticeably richer is happening in the case of lions. Nonetheless, unlike an

actual heat-seeking missile, the mosquito does have a neurological system, and as such, it can be placed (along with Smith and the lion) somewhere along the neurobiological spectrum. In other words, there are some important similarities between these three. This is not to suggest that mosquitoes ought to be necessarily classified as agents *per se*, but studying their behavior in relation to lions, human beings, and passive events like tree-rottings or volcano-eruptings can help determine more precisely what distinguishes an action from a happening, *viz*., where agency *begins*.

When assigning causal responsibility, there is no need to speak of reasons, intentions, or even desires. Who or what was responsible for the powering of the television? It was Smith. Who or what was responsible for the killing of the zebra? It was the lion. Who or what was responsible for the drawing of the blood? It was the mosquito. How might these cases be different from, say, motorized fans or storm clouds? When the motorized fan suddenly short-circuits and shuts off, is it causally responsible for its shutting down? When it begins violently storming on your wedding day, are the clouds causally responsible for the ruining of your outdoor wedding? There is no obvious external compulsion in the latter two cases. Where, then, is the line to be drawn?

1.4 | It's Alive! It's Alive!

One important feature that Smith, the lion, and the mosquito have that motorized fans and storm clouds lack is *cognition*. There is sometimes ambiguity surrounding the term, as people tend to think of cognition as implying psychological states, thought processes, or some other internal correlate that requires conscious awareness of mental goings-on. This is true to an extent, but for cognitive scientists, these kinds of activities properly belong to *higher-level cognition*. There are other cognitive activities, however, performed by all

living organisms with neurological systems. Some of the most rudimentary forms of cognition include perception, attention, and memory, but neuroscientist Joseph LeDoux, for example, has persuasively advocated for including even emotion amongst cognitive abilities (LeDoux 1996, 35).

When reviewing how desert ants (*Cataglyphis fortis*) forage for food in the Sahara Desert and successfully find their way back to their nests, Hugo Mercier and Dan Sperber explain how cognition plays a role and discuss why it may have evolved in living organisms:

> Cognition is first and foremost a means for organisms that can move around to react appropriately to risks and opportunities presented by their environment. Cognition didn't evolve in plants, which stay put, but in animals that are capable of locomotion. Cognition without locomotion would be wasteful. Locomotion without cognition would be fatal. Desert ants in particular moving out of their nest would, without their cognitive skills, quickly fry in the sun. (Mercier and Sperber 2017, 56)

Cognition extends all the way down to even desert ants and mosquitoes. By using processes involved in attention, perception, and memory, these creatures are able to navigate their environments, taking advantage of opportunities while avoiding threats. If agency, in the sense of acting free from external compulsion, is to have its origins anywhere, cognition would be the best place to start. It is cognition, after all, that makes *learning* possible, the same learning that equips organisms with the adaptive flexibility needed to have the best chance of succeeding in complex and ever-changing environments. Learning, however, presupposes an even more fundamental cognitive capacity: representation.

1.5 | Re-presenting Representation

Perception is perhaps the simplest cognitive activity, and yet it is still quite complex. As Hofstadter suggests, if one were to connect a camera to a television such that the television started to display the incoming information from the lens, sadly (for the television) no perception would be taking place. The television's activity is better characterized as *reception*, for all that is displayed on the screen is the input received from the lens and nothing more (Hofstadter 2007, 75). So what more is needed for perception?

It is true that perception, too, requires some kind of input, but the cognitive process of perception enables an organism to go *beyond* the input by sorting, categorizing, highlighting, and ignoring information received. This is done, explains Hofstadter, through *symbolization*. From a neurobiological point-of-view, what this means is that when an organism perceives its environment, its sensory input activates highly specific neural pathways that become consistently correlated with environmental features. Whichever neural pattern is activated when you think of "hamburger" *is* the neural pattern *for* "hamburger"—it *symbolizes* it—and that pattern is a very different one from, say, "nothingburger." "Symbols in a brain," says Hofstadter, "are the neurobiological entities that correspond to concepts, just as genes are the chemical entities that correspond to hereditary traits" (Hofstadter 2007, 75–6). What symbols ultimately do is help organisms *represent* their environments, abstracting from the sensory data the information that is relevant to whatever an organism is trying to do; they enable organisms to map concepts like *threat* or *food* on to the environment they are sensing.

There are two issues that need to be addressed when invoking the notion of *representation*. First, one of the chief problems with theoretical accounts of representation occurs when one assumes that a representation must be *identical* with the thing represented. This is a naïve account of representation, and as such, it should be discarded.

In engineering and physics, it is possible to represent all sorts of things through mathematical formulas and diagrams, from the equations that symbolize the effects of gravity on an object to the abstract blueprint of a manufacturing facility. Even something as simple as a roadmap is a representation of the surrounding environment. Though these sorts of things are not faithful replicas of what they represent, they are useful for helping people interact with their environments in particular and meaningful ways. If one wishes to represent a building that one would like to evaluate for remodeling, a blueprint that includes information about electrical circuits, dimensions of rooms, materials used, and other relevant information would be necessary, but this would not be an ideal way to understand heating efficiency. For that, one might use a thermographic image to check for heat retention and heat loss. To represent the economic value of the project though, one would need further still to hire an appraiser to assess the property and determine the right monetary figure that reflects that value. Each of these representations is very different from the other, and yet they represent the same object, highlighting or isolating various aspects of it for different purposes.

The neurologist Antonio Damasio makes a similar suggestion in favor of a more robust understanding of representation when discussing mental images and neural

patterns, both of which he counts as representations.[30] In clarifying his use of the term, he writes:

> [Representation] simply means "pattern that is consistently related to something," whether with respect to a mental image or to a set of neural activities within a specific brain region. The problem with the term representation is not its ambiguity, since everyone can guess what it means, but the implication that, somehow, the mental image or the neural pattern *represents*, in mind and in brain, with some degree of fidelity, the object to which the representation refers, as if the structure of the object were replicated in the representation. When I use the word representation, I make no such suggestion. I do not have any idea about how faithful neural patterns and mental images are, relative to the objects to which they refer. (Damasio 1999, 320)

For Damasio, the purpose of a representation is precisely to enable organisms to interact with their respective environments in meaningful ways, and what is demanded of the representation is not fidelity to what it represents but consistency with it. If one uses a mnemonic device to represent and remember the items in a room, its use is only good insofar as the same letters stand for the same things.

Damasio further suggests that it would be a mistake to think of representations in purely *visual* terms; instead, he importantly broadens the concept to include more than just visual images, explaining that *any* sensory modality can be the subject of a representation, such as "touch, muscular, temperature, pain, visceral, and vestibular" (Damasio 1999, 318). When one recalls in memory and imagination the feeling of accidentally touching a hot stove or the piercing discomfort from a dentist drill, she is drawing upon a representation of those experiences. In addition, Damasio argues, there is

[30] It is worth noting that Damasio also accepts the thesis of biological continuity. In the opening chapter of *The Feeling of What Happens*, he details some of his ideas about consciousness, revealing his sympathy for the thesis. For instance, he believes that the simplest form of consciousness is *core consciousness*, which he describes as "a sense of self about one moment—now—and about one place—here" (Damasio 1999, 16). It is not rich enough to provide for any stable sense of identity, is not unique to humans, and does not depend on processes like memory, reasoning, or language; it simply enables an organism to interact with the objects in its environment and serves as the foundation for higher, more advanced forms of consciousness (16).

no requirement that a representation must be a "static" image, for we can have "sound images such as those caused by music or the wind" (318). One can, for instance, mentally trace the acoustic contours of Chopin's Op. 9 No. 1 Nocturne as if she were hearing it in real time, an observation that researchers have empirically verified in laboratory settings with test subjects as well.[31]

What is more, like Hofstadter, Damasio recognizes that neural patterns are consistently mapped to different behaviors, perceptions, and experiences of a subject; as such, they also qualify as representations in his sense of the term. This thus makes them a form of *nonconscious* representations, for a subject is not able to access or become aware of the neural pattern underlying the production of her experiences. The topic of nonconscious representation introduces the second issue with representational theory, which is the idea that representations must necessarily be "mental" in the sense that they are only products of *conscious* thought processes. It is obviously true that we can experience these neural patterns with the help of fMRI imaging and other diagnostic tools, but Damasio's point is that we cannot have a first-person experience of these things unaided by technology (Damasio 1999, 318). But this is not the only form of nonconscious representation he acknowledges.

Not only does Damasio believe that neural patterns are types of nonconscious representations, he also believes that there are nonconscious mental images. For instance, there are mental images that are accessible to conscious thought *in principle* although

---

[31] In the August 2014 *Nautilus* article, "This Is Your Brain on Silence," author Daniel Gross explains, "Imagine, for example, you're listening to Simon and Garfunkel's 'The Sound of Silence,' when the radio abruptly cuts out. Neurologists have found that if you know the song well, your brain's auditory cortex remains active, as if the music is still playing," citing his conversation with a researcher of human auditory processing at Dartmouth, David Kraemer (Gross 2014). See: http://nautil.us/issue/16/nothingness/this-is-your-brain-on-silence.

they are not being attended to at some particular moment, such as some kinds of feelings, emotions, and memories, each of which have a representational character (Damasio 1999, 51–2; 332). The emotions generated by the nostalgia when one recalls walking a beach at sunset are not brought into conscious thought until one draws upon that memory, and yet those emotions *are* representations that one has, even when not currently remembered, for they are stored somewhere. There are also mental images that are *inaccessible* to conscious thought, as in cases of blindsight where those affected can successfully point at objects in a meaningful, statistically significant way even though they cannot see as a result of cortical blindness (268). Such subjects behave *as if* they are aware of the relevant objects in spite of being unable to *consciously* experience them.

Following Damasio, I am taking representation in a similar sense. More specifically, within the context of action theory, it is best to think of representation as that towards which a subject can orient herself such that she is able to interact with her environment in ways that serve her interests. It is something that goes beyond perception. Anytime some subject treats an object in its environment *as* a *something*—source of pleasure, source of pain, source of food, source of aid, etc.—it is representing aspects of its environment to itself by pursuing, avoiding, or ignoring those things accordingly. This is more of a behaviorist understanding of representation, but the reason for understanding it in this way is that simpler organisms can nonconsciously interact with their environments in sufficiently behaviorally complex ways that they must count as having representations of their environment even if it is the case that they lack the neurological capacity to form mental images. Extracting information from perception and behaving

accordingly is precisely what it means to use representations, regardless of whether that information is imagistic in nature or otherwise.

When it comes to the rats exhibiting mathematical cognition in Johnson and Platt's experiment, for example, the psychologists Charles Gallistel and Rochel Gelman argue that these nonverbal animals must be representing some kind of numerical quantities to themselves in the absence of the tools necessary, namely language, for discrete representation (Gallistel and Gelman 2005, 580). When a person adds seven to three, the numerical property of *seven* is represented by a distinct word "SEVEN" or symbol "7," and this is also true when one turns to the numerical properties of *three* and *ten*. Given the scalar variability in lever pressings, the rats appear incapable of *discrete* representation when it comes to number. For human beings, being able to use highly specific words or symbols in place of these numerical properties—discrete representation—helps us keep more accurate track of what we are doing. The rats, however, must be doing this some other way. Is the vague representation they use a mental image? Is the rat conscious of it? Might the rats be using *nonconscious* mental images? Though these are important questions that need to be settled, no matter the answer, a behaviorist understanding of representation captures the salient feature of what it means to represent something—using one thing to symbolize something else—without depending on the presence of images as a requirement. On a behaviorist understanding, the rats are acting as if they are representing, and therefore they are representing. Without the language of representation, it is not clear how else to talk about what it means to act on the basis of information that goes beyond the immediate perception of the environment as it is.

1.6 | Integrating Ideas

Clearly, Hofstadter is skeptical that organisms like mosquitoes are cognitively complex enough to engage in learning. He argues (rightly) that being able to symbolize is integral to perception, and he even uses the language of representation to explain symbolization, describing symbols as having a "representational quality" (Hofstadter 2007, 75). But mosquitoes, he thinks, almost certainly do not even make the cut for perception, one of the most basic cognitive activities. He writes:

> What kinds of [mental] categories, then, does a mosquito need to have? Something like "potential source of food" (a "goodie", for short) and "potential place to land" (a "port", for short) seem about as rich as I expect its category system to be. It may also be dimly aware of something that we humans would call a "potential threat"—a certain kind of rapidly moving shadow or visual contrast (a "baddie", for short). But then again, "aware", even with the modifier "dimly", may be too strong a word. The key issue here is whether a mosquito has *symbols* for such categories, or could instead get away with a simpler type of machinery not involving any kind of perceptual cascade of signals that culminates in the triggering of symbols.
>
> […] I would be quite happy to compare a mosquito's inner life to that of a flush toilet or a thermostat, but that's about as far as I personally would go. Mosquito behavior strikes me as perfectly comprehensible without recourse to anything that deserves the name "symbol". In other words, a mosquito's wordless and conceptless danger-feeling behavior may be less like perception as we humans know it, and more like the wordless and conceptless hammer-fleeing behavior of your knee when the doctor's hammer hits it and you reflexively kick. Does a mosquito have more of an inner life than your knee does? (78–9)

If mosquitoes are nothing more than a bundle of reflex arcs, is it fair to say that they have *any* cognition, at least in any meaningful sense of the word? Or are they more like automata? To answer these questions, it is important to understand some key differences between reflexes and the cognitive equipment required for learning.

It was mentioned in section 1.4 that learning presupposes the capacity for representation. The reason for this has to do with something known as *cognitive*

integration. Cognitive integration occurs whenever an organism's internal states begin interacting with one another, using the informational output of one cognitive mechanism as input for another, creating a unified system that functions together. When one talks of a creature adapting to its changing environment, it is cognitive integration that makes it possible. As philosopher José Luis Bermúdez explains it:

> The behavior of organisms that are suitably flexible and plastic in their responses to the environment tends to be the result of complex interactions between internal states. Organisms respond flexibly and plastically to their environments in virtue of the fact that their representational states respond flexibly and plastically to each other, most obviously through the influence of stored representations on present representations. The possibility of learning and adaptation depends on past representations contributing to the determination of present responses, and hence interacting with them. (Bermúdez 2003, 9)

Indeed, representation is required for memory to be able to do what it does—store information so that it can be manipulated or recalled—and memory is presupposed by so many other cognitive activities, such as causal understanding, arithmetical reasoning, pattern recognition, and assessments of similarity, to name a few. It would not be incorrect to say, metaphorically speaking, that representation is the mental glue of cognition, and that it, along with memory, is critical for learning.

One of the concerns after the initial experiments with the Aesop's Fable paradigm is that the birds may have had a pre-existing preference to manipulate objects like stones and sticks. Did they *really* exhibit causal understanding of water displacement or were they simply acting on the basis of reward reinforcement and prior training? One study set out to test exactly this hypothesis.

Prior to exposing new Caledonian crows to the Aesop's Fable test, researchers trained them to prefer non-functional objects, such as hollow, wire-frame constructs and floating objects, both of which would have a minimal effect on water displacement and

risk making the reward in the tube inaccessible (Miller et al. 2016, 11). The experiment validated those initial concerns. During the initial trials, the crows showed a bias for manipulating the non-functional objects they had been trained to use, dropping them into the tube to access the reward. Floating objects, for instance, were preferred in 76% of their selections. What was surprising, however, is that over repeated trials the crows eventually learned to prefer the *functional* objects to displace the water and access the reward. By the end of the 5th trial, the preference for non-functional objects had dropped to just 48%, and by the end of the 30th trial, it made up a meager 23% of their selections (12). While this experiment shows that associative processes and reward reinforcement plays a greater role in initial approaches to the problem than the exercise of causal understanding, that the crows were able to learn to solve the problem over time in the first place demonstrates the work of at least *some* degree of causal understanding (16).

What is more important is that this experiment (and others) shows that crows demonstrate the ability to flexibly respond to novel problems. This is evidenced by the fact that the researchers above used crows who had previously encountered the Aesop's Fable test as a control group, and from the outset, they had selected the correct object 63% of the time (Miller et al. 2016, 12). Unlike the untested crows, the control group expressed familiarity with the problem, drawing from their stored memories to solve it. But being able to solve the problem at all requires forming a connection between object-dropping and reward-raising, associating the two and storing it into memory for applications in subsequent tasks. It is not clear how this is possible in the absence of representations.

When it comes to reflexes, this type of plasticity in behavior is not observed. Unless it has been conditioned otherwise, the same response is invariably elicited in the presence of the relevant stimulus. The classic example of an invariant response is the patellar reflex, when the knee automatically jerks forward after being struck by the physician's hammer. But even when a response is conditioned, it still exhibits this same kind of inflexible connection between stimulus and response. As noted by Bermúdez, "Tropistic and classically conditioned behavior can be explained without reference to representational perceptual states because the response is invariant once the creature in question has registered the relevant stimuli" (Bermúdez 2003, 9).

There also exists what might be thought of as a more advanced form of the reflex: the *fixed action pattern*, discovered by ethologists Konrad Lorenz and Niko Tinbergen. Rather than a simple reflex in response to a stimulus, a fixed action pattern instinctively initiates a chained sequence of movements through the work of something known as an *innate releasing mechanism* (Bermúdez 2003, 7). Like the reflex, fixed action patterns are characterized by the same type of invariant behavior when deployed in the presence of the relevant stimulus. As Tinbergen discovered in the 1950s, newly hatched herring gulls will peck at anything placed in front of them that resembles the adult herring gull's bill.[32] Bermúdez argues that due to the invariant nature of their response, it is better to think of this behavior in mechanical terms (8).

If Hofstadter is correct in his assessment of the mosquito, then it should follow that it is too cognitively simple for any learning to take place, that all of its behavior should be able to be described in terms of reflexes and fixed action patterns.

---

[32] See *Thinking Without Words* (Bermúdez 2003, 8). See also *The Herring Gull's World* (Tinbergen 1961).

1.7 | Learning on the Fly

Ecologists Thomas Ings and Lars Chittka set out to understand more about learning and predator avoidance by studying how bumblebees altered their foraging habits in the presence of robotic crab spiders. Knowing fairly well how the bumblebee perceives colors, they used sixteen artificial white and yellow flowers for their experiment, placing nonlethal robotic spiders on four yellow flowers at any given time. Some of the spiders were *cryptic*, matching the color of the flower to simulate the predatory strategies of actual crab spiders in the wild, strategies that force bees to rely on shape cues rather than color. Other spiders were *conspicuous*, appearing white in color because, from the bumblebee's perspective, this made for a stark visual contrast with the yellow (Ings and Chittka 2008, 1520).

At the beginning of the experiment, they found that bees in both groups, cryptic and conspicuous, were caught by the robotic spiders at a rate fairly close to random, approximately one out of four times. This was important because it suggests that the bees lack "an innate response to the visual appearance of the spiders," ruling out the likelihood of more mechanical responses driven by instincts (Ings and Chittka 2008, 1520). What the researchers expected to happen next was to find that the bees in the conspicuous group would begin avoiding predation attempts much faster than the bees in the cryptic group, but to their surprise, they discovered that the bees in *both* groups learned to avoid their robotic predators at the *same rates* (1520). How was this possible? Was there any learning taking place or was this the result of some instinctive behavioral defense mechanisms issuing in fixed action patterns?

Using three-dimensional video tracking software to better monitor bumblebee behavior, it turns out that the bumblebees encountering the cryptic spiders started slowing down their inspection flights to more carefully check for the presence of possible predators. In fact, the difference in inspection time between the two groups only continued growing over time with increased visits, "suggesting," as the researchers say, "that bees became increasingly cautious as they learned the meadow contained cryptic predators" (Ings and Chittka 2008, 1520). This modified behavior persisted even after 12 and 24 hours had elapsed, demonstrating memory retention. As Ings and Chittka explain, "This speed-accuracy tradeoff is all the more interesting because it appears selective: Bees do not alter their flight behavior when they have learned that spiders are easy to detect" (1521). They expand on the significance of this:

> A common assumption about memory is that learned associations and responses tend to fade over time without further reinforcement. However, memory can be highly durable in insects, and in some animals, memories (or responses to past events) can actually intensify over time despite the absence of new learning trials. Indeed, we found that the learned predator-avoidance of bumblebees subjected to simulated predation attempts at flowers harboring either conspicuous or cryptic spiders was persistent over at least 24 hr. (1521)

Even more fascinating is that this learned behavior began resulting in an increasing number of false alarms for those bees encountering the cryptic spiders, something noticed only after the 24 hr memory tests. Ings and Chittka believe that the "increased rate of false alarms indicates that bees are extending their perception of danger to all yellow flowers rather than just those with cryptic spiders" (1521). The importance of reacting to false alarms indicates that this is not simple behavior that the bees are exhibiting. What a false alarm indicates is that there is a *disruption* between a stimulus and a response, reacting *as if* a stimulus were present when it is in fact otherwise. Reflexes and fixed

action patterns, by contrast, consistently demonstrate invariant responses in the presence of the relevant stimuli. The only thing that makes such disruptions possible, explains Bermúdez, is *misrepresentation* (Bermúdez 2003, 9). And so these bees must be using representations of their environment and making mistakes.

That bumblebees are able to make assessments of their environments and adapt their behavior accordingly suggest that they possess a degree of cognitive integration, warranting an attribution of a particular form of cognition to them, one that utilizes representations. After all, they can perceive their environments in sufficiently rich ways to alter foraging habits based on their experiences, and even more striking, they can make mistakes from exercising too much caution. However, bumblebees also have approximately one million neurons in comparison to the mosquito's one hundred thousand. Even if bumblebees (surprisingly) make the cognition cut, there is still quite the neuronal gap for mosquitoes to fill. But that gap may be narrower than one expects.

A team of researchers wanted to know whether—and if so, how—mosquitoes respond to aversive learning. As they point out, life as a mosquito is rather difficult, for the host is both prey and predator, often swiftly killing the mosquito right in the act of drawing blood (Vinauger et al. 2018, 333). It would make sense, then, if they were able to adjust their behavior accordingly in response to defensive strategies employed by hosts. For their experiment, the researchers first isolated a group of mosquitoes in a small chamber where they were conditioned to associate a particular type of human odor with an aversive stimulus consisting of nonlethal mechanical perturbations. Once they were conditioned, the researchers waited 24 hours before setting them free in a small Y-shaped maze where they had to choose between the conditioned stimulus (CS) odor and the

control. The result was that the trained mosquitoes showed significantly less preference for the CS odor than the untrained mosquitoes (334). This same experiment was repeated with rat and chicken odors as well, and when it came to the chicken odors, there was no difference between the trained mosquitoes and the untrained ones, suggesting that not all chemical odors are equal as far as mosquitoes are concerned (334).

Next, the researchers observed whether the CS odor had an effect on biting preferences by setting up two feeders of bovine blood, one scented with the CS odor and the other a control. While the trained mosquitoes showed a preference for the control, those that did land at the CS feeder probed the bovine blood with the same statistical frequency as the untrained mosquitoes, demonstrating that the conditioning affected only their flight behavior but not their biting behavior (Vinauger et al. 2018, 334).

Finally, understanding the role that the neurotransmitter dopamine plays in both memory formation and learning outcomes, the researchers manipulated the dopamine receptors of another group of mosquitoes. Compared to control groups, the manipulated mosquitoes "showed significant deficits in their learning abilities" even though it had no effect on the motor skills involved in flight or their innate preferences for human odors (Vinauger et al. 2018, 336–7).

While classical conditioning was used in the mosquito experiments, there are several observations worth nothing. First, the mosquitoes demonstrated their conditioned behavior after a 24 hour waiting period, reflecting that the associations made during the conditioning were stored in memory. Second, the conditioning only affected their flight preferences, changing neither their feeding preferences nor their motor skills. This reveals that several independent cognitive mechanisms are working together to produce

feeding behavior. This seems to point to a degree of cognitive integration. And third, in vertebrates, dopamine signaling plays an important role in creating a reward system whereby creatures learn to pursue environmental stimuli based on the reinforcement from discharges of dopamine in their brains, discharges that elicit feelings of pleasure (Sirigu and Duhamel 2016, 47). Parallel cognitive mechanisms have been discovered in insects (Perry and Barron 2013, 543). Given that interfering with dopamine signaling disrupted the flight preferences of the mosquitoes, it is likely that some kind of primitive reward system plays an important part in their feeding behavior.

It turns out that even with their scant amount of one hundred thousand neurons, mosquitoes are a little more cognitively complex than one might expect. Granted, it is certainly possible that reflexes and fixed action patterns alone can account for mosquito behavior. After all, they did not seem to make any obvious mistakes (like the bumblebees) and much of their learning can be accounted for by classical conditioning. Still, even in the mosquitoes one can find diminished analogues of the same kinds of cognitive processes used in more sophisticated organisms. Whether or not they manage to make the cognitive cutoff, it is clear that representational cognition plays an important role in acting vis-à-vis learning and responding to environmental changes.

This is not to suggest that mosquitoes and bumblebees are agents in any traditional sense of the word; rather, just as mosquitoes have a primitive reward system, the goal is to recognize a primitive sense of agency that scales up with the complexity of an organism. This crude form of agency is what I will call *basic agency*, and it serves the purpose of assigning causal responsibility to creatures that are capable of acting free from external compulsion. While representational cognition—the kind of cognition that unifies

an organism through cognitive integration—plays a key role in performing actions as a basic agent, it is not the only piece to the puzzle.

1.8 | If You do the Locomotion, You Have to do it Well

In section 1.4, it was mentioned that Mercier and Sperber believe that cognition evolved with locomotion in organisms. While locomotion is certainly important, when it comes to the kind of causal responsibility relevant for basic agency, organisms need something more than *just* locomotion. Having the cognitive capacity to cause events to occur is not by itself sufficient for qualifying as an agent if a subject is not able to do so in a causally relevant way. While this certainly involves locomotive activity, it is not just locomotion that comprises this requirement but also means and opportunity.

Consider that mosquitoes have a sharp, spear-like proboscis that enables them to pierce through multiple layers of skin in order to draw blood from their host; by contrast, crane flies lack the mouth-parts to draw blood. It is not uncommon for people to mistakenly identify a crane fly as a large mosquito, fearing that its bite will be proportionately more terrifying than the smaller, more common mosquitoes they encounter. However, it is utterly impossible for the crane fly to be responsible for an insect bite because it lacks the means to harm a human being in this way. While it does have a proboscis, it is significantly shorter and adapted for consuming decaying vegetative matter, a diet only known to be followed during the immature life stages. Once they are adults, most species actually stop feeding altogether.[33]

[33] As is the case with most things in nature, there are exceptions to the rule. Some species of *Elephantomyia* and *Limonia* do possess significantly longer probosces, but they are adapted for drinking nectar from flowers.

This additional requirement for basic agency—means and opportunity, or *causal relevance*—is one of the cornerstones for judicial systems where criminality is proved by examining whether one had the means to break the law at a given place and time, in short, whether one can be *responsible* for the alleged crime. While this might seem self-evident and trivial, its importance lies in the fact that it further limits what can count as an agent in both particular and general cases by determining whether it is even possible for a given subject to have caused an event to occur.

For example, an ardent believer of witchcraft, Betty, might be convinced beyond a shadow of a doubt that her neighbor, Jack, has cursed per. As far as Betty is concerned, Jack is *the cause* of her present misfortunes. Fortunately, such hypotheses can be tested and counterexamples are plentiful. For instance, one might ask: Did Jack *actually* perform such a ritual or is Betty just imagining that he did? If Jack did so, has he cursed others? Is it possible to prove that there is a consistent, lawlike connection between the act of cursing and the beginning of misfortunes? Has cursing ever been demonstrated in a laboratory setting with strict controls? Using Occam's razor, might there be more immediate, sound explanations for Betty's misfortunes? Whether or not Jack even performed such a ritual, after a careful consideration of the evidence, it is clear that whatever he did, if anything, it was not causally relevant to Betty's misfortunes.

Not only can the causal relevance requirement adjudicate cases involving particular attributions of responsibility, it can also aid in determining whether a prospective subject can count as an agent in principle. Suppose that a young girl, Charlene, has an imaginary friend, firmly believing that he is a real person. This imaginary friend, named Alasdair, keeps Charlene entertained. One day, Charlene's

mother was fixing dinner when Charlene noticed that just two place settings were prepared. Feeling as though Alasdair has been slighted, Charlene requested a third for her friend. Her mom humored her, not only preparing the third place setting but even serving a small amount of food with it. When the meal finished, Charlene's mother started cleaning up when she noticed that Alasdair's meal was gone. Curious, she inquired of Charlene, "Where'd the food go?" Charlene smiled, pointed her finger at an empty chair, and said with an ornery voice, "Alasdair!" While one could certainly run Charlene through the same battery of questions posed to Jack and Betty, even more fundamentally, there is simply no way Alasdair could have eaten the food or have done, well, *anything*. The reason is simple: he does not *actually* exist. It is thus impossible for Alasdair to count as an agent in any meaningful sense of the word; the very condition of his existence (i.e. *imaginary*) rules out the chance that anything he does will ever be causally relevant.

1.9 | There's More than One Way to…Respond to the Environment

While cognition and causal relevance are extremely important criteria for basic agency, there is a third requirement that must also be taken into consideration: *teleological behavior*, or acting in the service of some kind of goal. As Bermúdez argues, any organism that acts on the basis of a reflex or a fixed action pattern cannot be construed as acting with a sense of purpose. Discussing Tinbergen's herring gulls, Bermúdez explains that the newly hatched gulls are not exhibiting goal-directed behavior with their fixed action response of pecking at relevant stimuli because even though it might satisfy a desire by resulting in a feeding by the adult gull, the pecking behavior is "an invariant response to the appropriate stimuli" (Bermúdez 2003, 8). Although the response is not entirely coincidental due to the workings of natural selection, the satisfaction of the desire

for food in the young gull is more like a convenient side-effect of its preprogrammed mechanical response than the plastic behavior that follows from more advanced forms of cognition.

Because teleological behavior is in the service of a goal or desire, Bermúdez associates it with the presence of psychological states since it requires that a subject represent something *as* a goal or desire (Bermúdez 2003, 8). However, because representations can be nonconscious, there is no need to impose an additional psychological requirement. All that is required to demonstrate that behavior is teleological is that a subject act with a sense of purpose, regardless of whether she is consciously aware of her goals. It is not uncommon for even human beings to engage in this kind of behavior, such as when one fidgets with a nearby object for the sake of managing an anxious experience. In such situations, people often report not even realizing that they were fidgeting, but upon noticing it and reflecting on it, they explain that they were feeling nervous.

Speaking of behavior in terms of goals and desires might tempt one to think about teleological behavior in terms of *intention*—or the capacity to commit to doing something—but there are important differences that reveal that teleological behavior can be kept distinct from intentional explanations.[34] For one, it is possible to act with a purpose without intending to do it. This is clearest in the case of certain kinds of reactions that neither qualify as reflexes nor fixed action patterns but that still issue in appropriate responses to an environmental stimulus in the absence of the formation of an intention.

[34] The account given here is condensed for the purposes of conveying that it is not necessary to invoke intentional explanations for teleological behavior. Chapter three provides a fuller explanation of my understanding of intention.

Someone who intends to do no harm, for example, might find herself suddenly striking an aggressor without any forethought for the sake of escaping a dangerous situation. Similarly, one might intend to taste an extremely spicy hot sauce but pull away at the last second for the sake of avoiding the anticipated discomfort. There is no reason why committing to do something necessitates following through with it, and there is no reason why acting with a purpose requires committing to do it. In addition, habits might be started by acting with intentions to do something, but once such actions become habits, it is often the case that one acts without an intention. One might, for instance, form the intention to tap on a table with a pencil when thinking through a problem, finding that the rhythm and the sound brings a sense of comfort and promotes the ability to focus, but at some point, whether it is months or years later, this action can evolve into the kind of habit where one no longer realizes that she is even doing it, let alone forming an intention to do it in the first place.

Secondly, forming intentions requires mental states that can serve as the objects of the intention, while teleological behavior merely requires sensitivity to features in the environment. To commit to doing something, what is needed is some degree of self-conscious reflexivity such that one can entertain thoughts, beliefs, desires, courses of action, a mental *something* to which one can attach her commitment. Whereas in the case of acting with a sense of purpose, one merely needs to have a representation of the environment such that one can behaviorally react to the various features of it *as* good, bad, desirable, undesirable, pleasurable, painful, etc. Furthermore, responding to a feature as good or bad, for example, does *not* require that one form a mental *idea* of it as good or bad. Consider the case of touching a hot object: one does not need to mentally conceive

of the object *as* hot in order to pull one's hand away and react to it *as* harmful. The representation of a harmful object and the teleological response of avoiding what is harmful are embedded in the behavioral orientation that the subject has to such things rather than in some mediating thoughts in its mind.

1.10 | Conclusion

When the biological continuity thesis is embraced, the question of where a happening ends and an action begins forces one to look a little further down the biological spectrum for answers. Research from the areas of cognitive science, biology, neurology, ethology, and psychology suggest that the feature of cognition makes for the best starting place, equipping action theorists with a relatively clear cutoff between happenings and actions. No cognition? No action.

But cognition alone is not sufficient for an organism to be counted as acting in the crudest sense of the word. The organism must be cognitively integrated, acting as a unity, and this is made possible through a form of cognition known as *representational cognition*, which comprises the cognitive capacity to go beyond perception by symbolizing features of the environment in ways that promote learning. This is something that can only be accomplished when otherwise independent cognitive mechanisms begin interacting. In addition to this, the organism must not only be in possession of locomotion (which tends to be coextensive with cognition anyway), but must be in a position to act in causally relevant ways, having both the means and opportunity to be responsible for an event's occurrence. Finally, in order to distinguish reflexes and fixed action patterns from the flexible behavior required for acting, the behavior in question must be teleological or goal-driven. Otherwise, something as

automated as a patellar reflex in a primate would have to be counted as an action, which is obviously problematic.

These conditions together comprise what is needed for *basic agency*, a minimal form that does not go beyond mere causal responsibility but nonetheless underwrites more advanced forms of agency. Not only does introducing the idea of basic agency begin to loosen some of the Enlightenment grip that rationality has on the concept of action, it also helps justify an intuition concerning animals that seems increasingly plausible in light of emerging research—that their behavior is best explained by imputing a sense of agency to what they do—without explaining their behavior in terms of the possession of reasons. So what does life look like along the biological spectrum of agency? What makes more advanced forms of agency possible? And what role does reason play in any of this? To answer these questions, it is important to understand how reason became entangled with action in the first place.

Chapter 2: Deconstructing Donald

My dog can want me to take him for a walk
but he cannot want me to get my income tax returns in on time for the 1993 tax year.

— John Searle, "Animal Minds"

2.1 | Intending to Build a Squirrel Academy

It is becoming increasingly difficult to deny that animals partake in a range of complex cognitive activities, and yet there is still a surprising amount of reluctance to attribute higher-level cognitive functions to them. Although some concessions have been made, these are usually not without some semantic gymnastics. As philosopher Hans-Johann Glock has observed, few today will deny that animals possess at least some degree of *intelligence*, but many will insist that animals do not and cannot have *reason* (Glock 2013, 385). These self-appointed defenders of the Enlightenment crown will argue that intelligence is merely a problem-solving skill that helps creatures adapt to unforeseen circumstances, but reason (either literally or metaphorically) is as a divinely dispensed gift, unique to humans who were made in God's image. There is a sense in which this is true, but not for the reasons that the Enlightenment defenders have come to believe.[35] More pressing for present purposes, however, is that if animals cannot reason, then according to traditional action theory, they cannot act. Why might this be so? To answer this, let us turn to a traditional account of intentional agency, one defended by philosopher Donald Davidson.

Davidson begins one of his essays on action theory, "Intending," with a straightforward analysis of a standard human action. He invites us to imagine someone—whom I shall call Sally—who intends to build a squirrel house, and as a result, she

[35] This is addressed in chapter ten, specifically, sections 10.8—11.

undertakes the action of meticulously nailing the boards together in order to accomplish the task. At first glance, it seems apparent that the intention to build the squirrel house is what *caused* the action. Nobody manipulated Sally's hands; nobody demanded under threat of violence that she build the squirrel house; Sally's nailing of the boards was not the response of some reflex or fixed action pattern; in short, there is nothing to lead us to believe that Sally's building of the squirrel house occurred as a result of external compulsion. Better still, the virtue of invoking intention as part of the explanation is that it has the added bonus of coherently explaining just *why* the action of her building the squirrel house was performed without requiring any further analysis to make sense of it. It is a perfectly acceptable form of explanation to state, "The squirrel house was built *because* Sally *intended* to build it." As Davidson puts it:

> We are able to explain what goes on in such a case without assuming or postulating any odd or special events, episodes, attitudes or acts. Here is how we may explain it. *Someone who acts with a certain intention acts for a reason*; he has something in mind that he wants to promote or accomplish. (Davidson 1978, 83; emphasis added)

Thus, intentions simply *are* the reasons for undertaking an action, and as such, they are the *causes* of said action. This seems like a straightforward enough account of action.

As a philosopher, however, part of Davidson's job is to seek out problems, and in this particular case, he finds two. He observes first that it is possible for someone to build a squirrel house without an intention,[36] and second, it is also possible for someone to intend to build a squirrel house but never get around to doing so (Davidson 1978, 83). These concerns appear to create some problems for understanding how intention relates

[36] Though Davidson does not clarify, presumably he has in mind that someone could intend to build a *bird* house that squirrels happen to inhabit.

to action after all. In the first case, if it is possible for someone to cause something to happen without intending to do it, then intention is not *necessary* for an action; and in the second case, if it is possible to intend to do something without actually doing so, then intention is not *sufficient* for an action. How then are we to understand intention? What does this say about its relation to action? It certainly seems as though there is *some* kind of relationship between the two.[37] Where does one go from here?

2.2 | The Purest of Intentions

Davidson's solution to the first problem of whether intention is *necessary* for action is especially clever. Because actions can be described in many different ways, it is possible for an action to be intentional when presented under one description but not another. Whenever an interpreter finds a description of an action that makes it appear intentional, the interpreter has *rationalized* the action. A *rationalization* is an explanation of an action that identifies the agent's reason for undertaking the performance in the first place. As he explains in his essay, "Actions, Reasons, and Causes," "A reason rationalizes an action only if it leads us to see something the agent saw, or thought he saw, in his action—some feature, consequence, or aspect of the action the agent wanted, desired, prized, held dear, thought dutiful, beneficial, obligatory, or agreeable" (Davidson 1963, 3). Because Davidson believes that reasons (as intentions) play a *causal* role when it comes to actions, he takes rationalizations to be a form of causal explanation (3).

Knowing that actions can be described in different ways, Davidson poses the following litmus test for actions: if an interpreter can identify at least one description

[37] Note that this is only a serious problem if one understands agency in terms of *intentional* agency. Introducing basic agency, for example, resolves any difficulties here.

under which a prospective action appears intentional, then it is an action. Suppose, for example, that Sally's friend, Ricardo, found himself recently inspired to build his own animal habitat after visiting Sally. As soon as he returned home, he started to build a bird house, but unfortunately for Ricardo, he ended up with a squirrel house instead. Even though Ricardo performed the action of building a *squirrel* house, it came to be as a result of his intention to build a *bird* house, and so for Davidson's purposes, this illustrates that Ricardo's intention was sufficient in one way (it caused the action), but insufficient in another (it was supposed to be a bird house). This is clearest when the following observations are made explicit:

- Ricardo built a structure with the intention of building a bird house.
- Ricardo built this structure for the following reasons: he desires to build a bird house and he believes that this structure is a bird house.

The second bullet point supplies the rationalization of Ricardo's action: he *intentionally* built a *structure* on the bases of a desire and a belief. While those reasons cannot rationalize why this structure happened to be a *squirrel* house—for Ricardo had no intention to build one at all—they do suffice for explaining the building of the structure in the first place. Explaining why the structure is a squirrel house requires one to look elsewhere, particularly at a nonintentional explanation. In this case, it just so happens that, on account of the fact that squirrels inhabit the structure and not birds, the structure is functionally better suited for squirrels rather than birds. Even though intention cannot explain Ricardo's action from *all* points of view, there is at least one from which the action appears intentional.

The second problem for intentional action (of whether intention is *sufficient* for action) is somewhat of a non-issue for Davidson because he accepts this as true; intention

is insufficient for action. As far as he is concerned, all that he needed to do was prove that the idea of intention is inseparable from action. By instead demonstrating the existence of a necessary relationship between the two, he manages to do precisely that. This does not mean, however, that the second problem is not interesting. Because intention does not always issue in action, it occurs to Davidson that analyzing this phenomenon, which he calls *pure intending*, can be instructive regarding the nature of intention itself (Davidson 1978, 88). How is it possible to have an intention without acting on it? Does this not undermine the supposed intimate metaphysical relationship between intention and action? If there are times where having an intention does *not* cause Sally to build the squirrel house, then what good is having an intention at all?

2.3 | Getting around to It Later

Even though the idea of pure intending seems strange, the phenomenon is experienced quite routinely, especially in those cases where one intends to act at some point in the future. Ricardo might say to Sally, for example, "I intend to purchase some groceries on Thursday." Even though this utterance expresses an intention that Ricardo has currently, it does not terminate in a here-and-now action, for the purchasing of groceries will not be undertaken until Thursday. According to Davidson, the phenomenon of pure intending reveals an interesting fact about the nature of intention itself. Whatever it is, it is distinct from actions. "Intending," he says, "is a state or event separate from the intended action or the reasons that prompted the action" (Davidson 1978, 89). Thus, by focusing on pure intending, unadulterated by action, the hope is that an action theorist can develop a clearer picture of precisely what an intention is. This is exactly what Davidson does, and he first focuses on what pure intending is *not*.

First, Davidson believes that pure intending is *not* a state of commitment because there are standards that apply with commitments that do not apply with intentions. He says of commitments that "if the deed does not follow, it is appropriate to ask for an explanation," but there is no such requirement that extends to all cases of intention (90). If Ricardo states *publicly* that he intends to shop for groceries on Thursday, then Sally might inquire why he failed to do so. It is thus the *public expression* of the intention that obliges Ricardo to follow through with it. This, in turn, makes him accountable to others and responsible for acting accordingly. This is no longer *just* a matter of causal responsibility, but at a minimum, it involves *normative* responsibility—keeping in line with what one ought to do as a matter of social conventions—and possibly even *moral* responsibility. In the deceptively simple social interaction between Ricardo and Sally, promises are made, expectations are set, and inferences are drawn. Knowing what Ricardo's plans are for Thursday can have a very real impact on Sally's plans for the same day.

When it comes to intentions, however, there is no requirement that they be public expressions at all. People intend to do all sorts of things in the private recesses of their minds, sometimes acting on them and sometimes not. Davidson thus believes that the dividing line between intentions and commitments is that the latter entail obligations to follow through and carry consequences for failing to do so. He explains that "if an agent does not do what he intended to do, he does not normally owe himself an explanation or apology, especially if he simply changed his mind," but when it comes to a commitment, by contrast, "this is just the case that calls for explanation or apology when a promise has been made to another and broken" (Davidson 1978, 90).

In section 1.9, I stated that intention is a form of commitment, and if Davidson is correct, this seems to contradict that view. Commitment, though, can be understood in at least two different ways. For example, the *Oxford English Dictionary* recognizes that commitment can be taken as (1) being *dedicated to something* or as (2) being *obligated* (in the sense that it involves restricting one's freedom by imposing the force of a promise). The following two examples suffice to illustrate that they are not the same: Michel can be *dedicated* to living a healthier lifestyle without being *obligated* to do so, for he is not breaking any promises if he happens to enjoy a milkshake one evening; conversely, Chuck can be *obligated* by company policy not to take office supplies home without being *dedicated* to upholding the obligation. When Davidson argues that pure intending is not a commitment, he is using it in this second sense—evident from the fact that he uses the term *promising* synonymously, and even says of it, "Promising involves assuming an obligation"—whereas my use of commitment is more along the lines of dedication (Davidson 1978, 90).

Besides arguing that intention does not involve commitment, Davidson also thinks that intention does not imply a *belief* that one *will* act. This might sound odd at first. After all, when one utters an expression of the sort, "I intend to do *X*," it invariably seems to imply a belief that one plans to do it (91). If after dropping off a passenger, Matilda were to say, "I intend to park the car," then she almost certainly believes that she will in fact act in such a way so as to park the car. If intention does not imply the belief that one will act accordingly, then why do the two so frequently coincide?

As it turns out, the answer is not because the two are somehow inextricably linked but because one's awareness of the present situation engenders the belief. Matilda, for

example, is very much aware of her present circumstances—she just dropped off her guest, she is at a shopping plaza, she is driving a car, she is to join her passenger shortly, there is a nearby parking lot with plenty of vacant spaces, etc.—and as a result, she has a sense for the likelihood that her intended action will come to fruition in light of her circumstances. Conversely, suppose that the likelihood of Matilda's action were to come into doubt—the vacant parking lot suddenly begins filling, she is caught behind a delivery truck unloading its supplies, her daughter calls with an urgent request. In this situation, there is no reason why Matilda cannot still intend to park the car even while being aware that succeeding does not look very good. In other words, there is no reason why one cannot form intentions in the face of uncertainty.

As a final consideration, Davidson denies that intending is equivalent to *wanting* to act. In action theory, *wanting* is a technical term that denotes something known as a *pro attitude*. A pro attitude describes a kind of mental state that urges or compels a creature to behave in such a way so as to satisfy the conditions of the state in question. The most basic and paradigmatic pro attitude, of which wanting is a type, is known as *desire*. Thus if Harry desires a drink of water, he is experiencing some kind of mental state that inclines him to drink water. What Davidson is denying, then, is that intentions are equivalent to desires. It is true that intending implies desiring to some degree—if Harry *intends* to drink the water, then he *desires* the water—but the converse does not hold. There are all sorts of things that people can want or desire that never rise to the level of intentions. Davidson explains:

> What we intend to do we want, in some very broad sense of want, to do. But this does not mean that intending is a form of wanting. For consider the actions that I want to perform and that are consistent with what I believe. Among them are the actions I intend to perform, and many more. I want to go to London next week,

> but I do not intend to, not because I think I cannot, but because it would interfere with other things I want more. (Davidson 1978, 101–102)

It is true that not all of our wants and desires will rise to the level of intentions. We can and do find multiple, possible courses of action desirable, and yet, we only intend to act on some of them. Even in a matter as trivial as visiting a restaurant, one might find oneself desiring several different food offerings on the menu, but in general, one typically orders less than the total number of food offerings that one originally desired.

Now, if intention is not the same as desire, then what is it? A consideration of pure intending suggests that intention is what causes us to act on the basis of some desires but not others. How is this possible?

2.4 | Sometimes You Have to be Judgmental

Judgment is notoriously a philosophically problematic term. Is a judgment linguistic or non-linguistic in nature? Is a judgment something embedded in a logical structure or merely analyzable with formal logic? Are animals or infants capable of judging? What about insects? When a software program crunches data and then executes one of a number of functions contingent on the result of its initial analysis, did it make a judgment? Is perception a form of judgment or a precondition for it? Is there just one kind of judgment or many?

Unfortunately, settling all of these questions goes far beyond the scope of this project, but for present purposes, it is important to note that Davidson believes the following about judgments: (1) they are products of reasoning, and (2) they can be either *cognitive* or *practical*, issuing respectively in the formation of beliefs or the evaluation of possible courses of action.

A cognitive judgment takes the form of a *proposition*, a semantic entity whose meaning is expressed in the form of a *sentence*. In studies of formal logic, it is said that propositions have the job of relating two concepts with one another, a *subject* to a *predicate*, and this can be observed whenever one utters simple declarative statements. For example, uttering a statement like, "The grass is green," allows one to express the proposition that coordinates some predicate (e.g. the color <green>) with some subject (e.g. the thing <grass>) such that <green> is a property that belongs to <grass>.

Why not dispense with propositions altogether and just talk about sentences? There are a handful of reasons, but the most important ones have to do with explaining *meaning* and *truth-value*. Both inter- and intra-linguistically, it is possible to intend the same meaning with different words and phrases. When Francis says, "The grass is green," and Franz replies, "Ja, das Gras ist grün," even though one is speaking English and the other German, the two speakers are intending the same meaning. Suppose now for a moment that both men are wearing sunglasses, and as a result of this, they are terribly mistaken in their judgments because the grass is actually yellow. About what are the two men speaking? It cannot be the grass because that is yellow, not green. Is it their *perception* of the grass? If so, both Francis and Franz have their own perceptions, and so their sentences could not possibly mean the same thing since their unique perceptions, generated by their respective physiologies, have nothing directly in common. What is more, their statements would have to be referring to their *perceptions*, and so their statements could not be false. For each, the grass really does *appear* green. Positing the existence of theoretical entities like propositions helps resolve many of these issues. Both Francis and Franz intend the same proposition <the grass is green> with their utterances,

and the fact that the grass is yellow entails that that particular proposition is false. Propositions thus are a way of referring to and talking about the meanings of sentences.

What cognitive judgments do is help determine whether some proposition is true or false, resulting in a belief about that proposition, and this is accomplished through acts of reasoning. When Harry, for instance, sees that it is raining, he may infer that it is not sunny outside. One way of making this inference would be through a chain of reasoning that results in the cognitive judgment, "I believe that it is not sunny outside," a judgment that expresses the fact that Harry believes the proposition <it is not sunny outside> is a true one. Such a chain of reasoning would look something like this:

P_1 If it is raining, then it is not sunny outside
P_2 It is raining
C It is not sunny outside

Here, after a consideration of the reasons, the conclusion <it is not sunny outside> is taken up as a belief that Harry holds true.

Whereas a cognitive judgment focuses on whether a proposition is true or false, a practical judgment, by contrast, assesses the desirability of some course of action. This too is decided by a chain of reasoning, one that takes into consideration preferences, wants, and desires, and it can also be expressed in the form of propositions. While walking through the market, for instance, Theresa will be evaluating whether she would like to purchase apples, grapes, oranges, or pears, if anything at all. In this case, she may find herself reasoning as follows:

P_1 If I want something tasty, then I will purchase Gala apples
P_2 I want something tasty
C I will purchase Gala apples

Although a practical judgment is evaluative in character, there is no requirement that it be anything more than an expression of preference.[38] Theresa might feel strongly about the current state of Gala apple farming, making her choice a moral one, or she might simply be thinking in terms of what will maximize pleasure. In both cases, the practical judgment is at bottom an assessment of what she desires; the difference is in specifying on what grounds the judgment has been made (moral or hedonic). Because practical judgments sort and evaluate possible courses of actions in terms of their desirability, Davidson believes that intending is a special kind of practical judgment, and as such, it must be distinguished from others, namely, *prima facie* (*pf*) judgments and *all-things-considered* (*atc*) judgments.

2.5 | They are Practically a Family

Both *pf* and *atc* judgments are what he calls *conditional*, judgments made on account of particular considerations. *Pf* judgments, for instance, hold "that actions are desirable in so far as they have a certain attribute," that is, that actions are desirable *on the condition that* they have such-and-such an attribute (Davidson 1978, 98). If Theresa is concerned about workers' rights at apple orchards, then she is judging that Gala apples are desirable or undesirable in light of those considerations, making it a *pf moral* judgment. Similarly, desiring Gala apples in light of whether it will maximize her pleasure relative to, say, Granny Smith or McIntosh apples, is a *pf hedonic* judgment.

---

[38] Davidson issues a similar warning / reminder to his readers:

> No weight should be given the word 'judgement'. I am considering here the *form* of propositions that express desires and other attitudes. I do not suppose that someone who wants to eat something sweet necessarily *judges* that it would be good to eat something sweet; perhaps we can say he *holds* that his eating something sweet has some positive characteristic. (Davidson 1978, 97, note 7)

Although it bears a close resemblance to *pf* judgments, *atc* judgments go a step further and hold that actions are desirable *on the condition that* they are evaluated to be the best in light of all available reasons. While Theresa might desire to maximize her pleasure, her moral considerations could outweigh her hedonic ones, and so in light of *everything*, she will make an *atc* judgment that Gala apples are not desirable.

What makes intentions different is that they are *all-out* practical judgments. That is to say, to intend an action is to judge that it is desirable *tout court*, that it is a judgment worth putting into action (Davidson 1978, 101). While a *pf* or *atc* judgment might give someone like Theresa a *reason* to act by determining that apple-purchasing is something desirable in such-and-such ways, the act of *actually* purchasing them "represents a further judgement that the desirable characteristic was enough to act on—that other considerations did not outweigh it" (98).

Consider the following example:

> Jones, upon finishing his meal at a restaurant, is asked by his server if he has any interest in dessert. Upon learning which desserts are available, Jones forms a couple of *pf* judgments. One is that eating a piece of cherry pie is desirable for its tartness, and another is that eating a piece of peanut butter pie is desirable for its savoriness. At the same time, Jones recalls his visit with his primary care doctor last month, and he has been trying to watch what he eats for the sake of good health. He thus forms an *atc* judgment that foregoing dessert altogether is desirable, given his high cholesterol and family history of deadly cardiac disease. While three different judgments are crossing his mind, each for very different reasons, Jones' server returns and asks if he has made a decision. In that moment, he forms the *all-out* judgment (the intention) that eating a piece of peanut butter pie right now is desirable enough that, of his three options, *that* is the one worth acting on, giving preference to the peanut butter pie judgment over the others.

As an all-out judgment, an intention thus functions as a kind of executive, meta-judgment that enables the agent to act on a desire. It is important, however, not to think that this implies that the all-out judgment is something *distinct* from the action. As mentioned

above, when it comes to practical judgments, the action can *be* the judgment, but it can also be "the formation of an intention to do something in the future," as one finds in cases of pure intending (96).

As judgments that express the preferences of a subject, Davidson's understanding of intention bears a close resemblance to my understanding of commitment, provided that one takes a preference to be an inclination to act in such-and-such a manner. However, given his reliance on the idea of propositions (i.e. semantic entities), what is not clear is whether he takes these commitments to be linguistic in nature. If one judges or holds that eating something sweet is desirable and hence worth acting on, must one *necessarily* be a language-user? Or is it that language is merely a tool that can *express* some kind of pre- or non-linguistic thoughts or feelings? For Davidson, it appears that such judgments really are necessarily linguistic.

2.6 | Intending, or My Life as a Linguistic Fascist

Lingualism is the philosophical position that thought requires language, that the two are so intimately bound up that they cannot be distinguished from one another. The implications for such a position are clear: any non-linguistic subjects are incapable of thinking. From this it follows that there is no reasoning, no intending, and hence, no action—only behavior, whatever that is. Although he does not refer to the position by name, Searle summarizes the argument well in his essay "Animal Minds":

> However let us turn to the actual arguments against the possibility of animal thinking. The form of the arguments is and has to be the same: humans satisfy a necessary condition on thinking which animals do not and cannot satisfy. Given what we know of the similarities and differences between human and animal capacities, the alleged crucial difference between humans and animals, in all of the arguments I know, is the same: the human possession of language makes

> human thought possible and the absence of language in animals makes animal thought impossible. (Searle 1994, 209)

Might Davidson be advocating for a form of lingualism?

In his essay, "Thought and Talk," Davidson argues that "belief is central to all kinds of thoughts," and though he does not mention intention specifically, he does single out a handful of other psychological states: emotional feeling (gladness), noticing, remembering, knowing, wondering, and considering. As he explains, "If someone is glad that, or notices that, or remembers that, or knows that, the gun is loaded, then he must believe that the gun is loaded" (Davidson 1975, 156–157). Any thoughts or feelings regarding the gun thus require that one be in possession of a cluster of gun-related beliefs, such as: there is a gun, it is loaded, loaded implies it has ammunition, bullets count as ammunition, such-and-such object is a gun, etc. Without these kinds of beliefs up-and-running in advance, one cannot have any meaningful thoughts or feelings, for they derive their meaning precisely from these beliefs.

The preceding section explained how, for Davidson, intention is a special form of practical judgment regarding the desirability of some course of action, and so taking that into consideration, there is no reason why Davidson would not extend this same assumption about belief and thought to cover desire and intention. It so far appears consistent to attribute to him the idea that if someone *desires that* or *intends that* the gun is loaded, then it must be the case that she *believes that* the gun is loaded, among other relevant beliefs. And in fact, he gives two other reasons for supposing that his theory of intention turns on the centrality of belief.

Davidson acknowledges elsewhere that beliefs are capable of conditioning intentions in the sense that, if one intends something regarding the future, other relevant

beliefs that one holds about the future can influence the formation of that intention (Davidson 1978, 99–100). For example, Jay might entertain the idea of going to the restaurant on Tuesday until he remembers that his friend, Mark, had talked about inviting him over for dinner at his place that day. Provided that there are no countervailing reasons and that he takes Mark to have made a sincere offer, Jay will intend to have dinner at Mark's place. It is clear that this same analysis holds for present, here-and-now intentions as well. Whether Jay is making up his mind for dinner on Tuesday or dinner *right now*, his beliefs regarding Mark and himself will influence the intention he forms. Thus, for Davidson, belief is indeed a prerequisite for intention. But does lingualism necessarily follow from the assumption of the centrality of belief?

If his position really is lingualist, then there should be some indication that language is bound up with thinking, and on this topic, Davidson cannot be any more explicit. Again in his essay, "Thought and Talk," he comments on exactly this, writing:

> The assumption [that language or thought—each relative to the other—is easy to understand on its own terms] is, I think, false: neither language nor thinking can be fully explained in terms of the other, and neither has conceptual priority. The two are, indeed, linked, in the sense that each requires the other in order to be understood. (Davidson 1975, 156)

Given this interdependence between thought and language as well as the idea that belief is not only a kind of thought but central to all thought (including intentions), it is undeniable that Davidson's position amounts to a form of lingualism, the consequence of which is that only language-users are capable of having intentions and performing actions.[39]

[39] It is thus no coincidence that Davidson describes the central thesis of his essay thusly: "a creature cannot have thoughts unless it is an interpreter of the speech of another" (Davidson 1975, 157). To be an interpreter of speech, he thinks that one must be a member of a speech community, but because language and beliefs are intertwined, one must also have beliefs. It is at this stage in his argument where he raises the

2.7 | Did I Just Do That?

What is wrong with being a lingualist about intentions? Although Davidson's theory of action precludes the idea that there are basic actions, is it not possible to salvage his insights by conceding that basic actions exist while maintaining that there are also higher forms of action (*intentional* action) that *do* require language in the way that he supposes? This is a step in the right direction, but it does not go far enough.

First, it would be a mistake to assume that the potential for performing a basic action is limited to cognitively simple organisms, that somehow more advanced creatures not only can but even just perform more advanced actions as an extension of what they are. Although it is true that organisms scale in complexity, it is still possible for them to possess the same features that define their more primitive genetic kin. For example, one of the simplest cognitive functions, the reflex, is found in everything from earthworms to human beings.[40] Are there correlates to basic agency in human beings as well? Consider for a moment the phenomenon of *nonconscious experience*.

Drawing upon the work of philosopher Gilbert Ryle, Hans-Johann Glock believes that "even the most rational performances need not be accompanied (and *a fortiori* need not be caused) by *conscious* processes of reasoning" (Glock 2013, 385; emphasis original). Though he does not provide an example of this, the most striking one might be

---

question, "Can a creature have a belief if it does not have the concept of belief? It seems to me it cannot" (170).

The reason for this rests on his assumption that the concept of belief, that is, taking something to be true, requires being able to grasp truth from falsehood. This in turn requires membership in a community of language-users, for there needs to be a point of contrast with one's own beliefs in order to see beliefs *as* true or false, something that other members of the community can provide. This is also how Searle interprets Davidson's argument (Searle 1994, 211).

[40] See "Escape reflexes in earthworms and other annelids" (Drewes 1984).

that of the nonconscious experience. This type of experience produces behavior that, to an observer, appears complex, goal-driven, and sensitive to environmental cues; and yet, for the person who is the subject of the experience, there does not seem to be any phenomenological quality to it, i.e., there is nothing it is like to have that experience.

Philosopher Peter Carruthers provides several familiar examples of nonconscious experience in his notoriously critical article, "Brute Experience," in which he claims that many animals simply have no experience—a thesis that is supposed to defend a slightly different lingualist position that holds that *conscious* thought and language depend on one another (Carruthers 1989, 511). Carruthers writes:

> While driving the car over a route I know well, my conscious attention may be wholly abstracted from my surroundings. I may be thinking deeply about a current piece of writing of mine, or phantasizing about my next summer's holiday, to the extent of being unaware of what I am doing on the road. It is common in such cases that one may suddenly "come to," returning one's attention to the task at hand with a startled realization that one has not the faintest idea what one has been doing or seeing for some minutes past. Yet there is a clear sense in which I must have been seeing, or I should have crashed the car. My passenger sitting next to me may correctly report that I had seen the lorry double parked by the side of the road, since I had deftly steered the car around it. But I was not aware of seeing that lorry, either at the time or later in memory.
>
> Another example: when washing up dishes I generally put on music to help pass the time. If it is a piece that I love particularly well, I may become totally absorbed, ceasing to be conscious of what I am doing at the sink. Yet someone observing me position a glass neatly on the rack to dry between two coffee mugs would correctly say that I must have seen those mugs were already there, or I should not have placed the glass where I did. Yet I was not aware of seeing those mugs, or of placing the glass between them. At the time I was swept up in the finale of Schubert's "Arpeggione Sonata," and if asked even a moment later I should have been unable to recall at what I had been looking. (505–6)

Carruthers' examples illustrate how, from the perspective of an observer, one can engage in seemingly rational behavior while at the same time being wholly unaware of what one is doing. Given that such nonconscious rational behavior is possible in human beings, this

raises a question of how to understand rational behavior. If it is possible to *appear* rational without entertaining a *conscious* process of reasoning, then what does it mean to act rationally?

2.8 | She Came Home… either in a Flood of Tears or a Sedan-chair

In 1949, Ryle published his *Concept of Mind*, in which he argued that we had inherited a faulty theoretical framework from Descartes. The Scientific Revolution with its mechanical natural philosophy seemed to entail that the material world was governed by deterministic laws. Every event that occurs, occurs *necessarily*, as effect from some preceding cause; it cannot be otherwise. To the disappointment of everyone intent on taking extreme measures to disprove determinism, Baron d'Holbach (intent on crushing their dreams) explains in 1770:

> "Am I not the master of throwing myself out of the window?" […] There is, in point of fact, no difference between the man that is cast out of the window by another, and the man who throws himself out of it, except that the impulse in the first instance comes immediately from without whilst that which determines the fall in the second case, springs from within his own peculiar machine, having its more remote cause also exterior. (Holbach 1770 / 2001, 106; n.73)

It should come as no surprise that such a worldview created an intense amount of anxiety for theologians, philosophers, scientists, and other intellectuals who believed that this jeopardized everything from the doctrine of the immortality of the soul to the possibility for genuine agency. The solution advanced by Descartes was to separate the mind from the body, as if the two were entirely different metaphysical things. Ryle summarizes (and to an extent, caricatures) the Cartesian position:

> The difference between the human behaviours which we describe as intelligent and those which we describe as unintelligent must be a difference in their causation; so, while some movements of human tongues and limbs are the effects of mechanical causes, others must be the effects of non-mechanical causes, i.e.

> some issue from movements of particles of matter, others from workings of the mind. (Ryle 1949 / 2009, 8–9)

By saving rational behavior from the bogeyman of determinism, an unintended side-effect of Descartes' remedy was that he rendered the minds of everyone but oneself mysterious and inaccessible. Interpreting another's behavior becomes little more than a guess. Does it *look* irrational? "Then it must spring from mechanical causes." Does it *look* rational? "There, that must be the mind in action." As Ryle observes, when the body and the mind are distinguished from one another, a mountain of absurdities grows between them: "It would have to be conceded, for example, that, for all that we can tell, the inner lives of persons who are classed as idiots or lunatics are as rational as those of anyone else," and, "Perhaps, too, some of those who are classed as sane are really idiots" (10). Herein lies the genesis of the contemporary philosophical concern with *zombies*, or people who *appear* rational but who lack inner mental lives.

By accepting that the ideas of *mind* and *body* somehow belong at odds in the same conversation, as if it makes sense to propose them together as contrasting pairs or otherwise in fundamental opposition to one another, Ryle argues that we, like Descartes, are making a *category mistake*. From his perspective, it makes no more sense to talk about the causal powers of *mind-or-body* than it does to talk about whether Charles Dickens' Miss Bolo came home in a flood of tears *or else* a sedan-chair (Ryle 1949 / 2009, 11). Instead of fretting about whether mental states are reducible to physical states or whether they denote two different modes of existence, Ryle exhorts his readers to entertain the idea that *mental* and *physical* are simply two kinds of description that are compatible with one another, two ways of describing the same thing.

It is astonishing that Davidson—who otherwise accepts Ryle's insights by forwarding the view that some behavior can be identified as an action whenever an interpreter locates at least one description that rationalizes it—allows himself to be hampered by Enlightenment thinking by implicitly accepting a variation of the Cartesian metaphysical framework and thereby sliding into lingualism. Much of the tension in contemporary action theory turns on a set of problems that descends from Descartes' category mistake: the assumption that behavior that *appears* rational implies that one is acting *with* reasons, and the assumption that non-rational behavior must be mechanical. Chapter one targeted the second assumption by explaining the role of representational cognition in generating causally relevant teleological behavior; the rest of this chapter will target the first.

2.9 | Monkey Business

Imagine being locked in a room, but you are not trapped—at least, you do not *feel* trapped. As far as you are aware, this is where and how you live. Each day, you awake to find a coin in your room, and whenever you are hungry, you walk up to the front of your room to exchange the coin for some food, delivered to you by a chef. Though prepared by a chef, the offering is usually the same each day, consisting of a medley of raw vegetables with no seasoning or sauces, vegetables such as broccoli, carrots, and cucumber slices. To help pass the time, there is also a window in your room that allows you to see your neighbor, a portly fellow whom you know well and whom happens to follow the same lifestyle. Now imagine one day, shortly after receiving your vegetable medley, you happen to notice that your neighbor receives a slice of cherry pie, which

happens to be one of your favorite desserts. You cannot recall a single time where the chef delivered a slice of cherry pie to him but not you. How would you feel?

Most of us would imagine that this must have been some sort of a mistake, and when this continues to happen day after day, we might start to get angry with the chef—asking questions, demanding answers, perhaps even throwing his vegetables back at him or abstaining from eating as a form of protest. It is reasonable to behave this way when you detect that you are being treated unfairly, and this is exactly how *Cebus apella* (the brown Capuchin monkey) behaved in primatologist Frans de Waal's groundbreaking experiments that sought to explore whether animals might have a sense of morality, looking specifically at whether and how they might perceive fairness and inequity distribution (Brosnan and de Waal 2003, 297).

The behavior of the monkeys cannot be explained in terms of invariant fixed action patterns, for not all of them responded the same way. Some monkeys over time started to reject their cucumber slices altogether, throwing them out of their cages or even in the direction of the researchers; some would violently shake their cages, as if in loud protest of their treatment; some would refuse to eat; and others seemed to gradually accept that this unfair cucumber life was going to be their new normal (Brosnan and de Waal 2003, 298). If the monkeys are not behaving mechanically, are they behaving rationally?

When it comes to the matter of interpreting behavior, a *subjectivist* is someone who would argue that acting rationally is equivalent to acting *for* a reason. For the subjectivist, the behavior of a subject must be interpreted by reference to the mental states that motivated the behavior, usually and paradigmatically beliefs and desires

(Glock 2013, 387). For example, considering the nonconscious examples above, the subjectivist would argue that Carruthers acted rationally only if it is possible to attribute the appropriate mental states to him, possibly asking the following questions: Did Carruthers have a *desire* to switch lanes? Did he have the *beliefs* that using a signal and turning the wheel could satisfy that desire? Did he *intend* to do those things? If the answer to these questions is no, then Carruthers could not have acted rationally. When it comes to the brown Capuchin monkeys, the question of whether they acted rationally turns on whether it is possible—and if so, how—to attribute beliefs and desires to them in the absence of language, a problem analogous to how a Cartesian can attribute mental states to others without fearing that they are well-disguised zombies.

The Rylian-inspired alternative is known as *objectivism*, which holds that acting rationally is to be understood as acting *in the light of reasons*, that is, in light of how the situation appears from the subject's point-of-view (Glock 2013, 387). The objectivist might inquire, "Given the facts and states of affairs surrounding Carruthers' circumstances while driving and rinsing dishes, does his behavior make sense?" Similarly, she could ask whether the monkeys are acting in a manner consistent with their situations. In both cases, it seems that the answer is a resounding yes—Carruthers and the monkeys act consistent with the facts and states of affairs that govern their circumstances. Glock, an objectivist, puts the point this way: "My reason for taking an umbrella is that it *is* raining, not that I *believe* that it is raining; for it is the weather rather than my own mental state that makes taking an umbrella good or bad in my eyes" (387; emphases added). For the objectivist, the perception of how things are determines how one behaves, and when one behaves in a manner consistent with how things really are, one is acting

rationally. But is this *really* what it means to act rationally, or is this an abuse of the term? And if subjectivism, by contrast, distances itself from lingualism, what is wrong with interpreting non-linguistic subjects as having non-linguistic, belief-like representations?

2.10 | Another Category Mistake

Subjectivism traditionally relies on the Cartesian dualistic classification of the world into rational and non-rational metaphysical *things*, and so objectivism is a welcome alternative that embraces a much older idea, *monism*, or the metaphysical assumption that reality is one, continuous, and whole. What defines rationality is not so much whether something counts as a rational or non-rational *thing*, but whether it can be described as exhibiting rational or non-rational *behavior*.

When action is defined in terms of reason, it is easy to see the appeal of objectivism. For Glock, who believes that non-human animals can count as agents, the fear is that subjectivism will preclude or at least make it extraordinarily difficult to account for agency in non-linguistic subjects. While objectivism theoretically moves us in the right direction, the problem with the subjectivism / objectivism distinction in the first place is that both positions rely on making another category mistake: assuming that the category that makes up behavior can be bifurcated into rational actions and non-rational, mechanical behavior. This war is waged over precisely what is it that counts as a reason for acting, and in this battle, both positions get something right and something wrong.

There are two different ways in which we can talk about *reasons*. On the one hand, a reason can be that which makes intelligible the phenomenon in question, usually

by relating the subject of the interpretation to a goal or purpose. This is the sense of reason employed when we say that the reason the plant turns towards the sun is for the sake of initiating photosynthesis. It does not imply that the plant *has* a reason for doing this; it instead implies that what the plant does makes sense with respect to plant-like functions and needs, i.e., what it does is *intelligible* to an interpreter. This is also the sense of reason recruited by objectivists, and it only serves to undermine their project. Even though what the plant does is reasonable, its behavior is invariant, and thus mechanical. Ironically, understanding reason in terms of intelligibility still embraces the same Cartesian metaphysic the objectivist had set out to avoid. By dropping out any appeal to subjective states, it is not clear how the objectivist can distinguish the mechanical from the non-mechanical.

On the other hand, a reason can be anything that can be recruited in a *chain of reasoning* and can be made explicit as premises in arguments. This is precisely what the subjectivist has in mind. Used in this way, the idea is that acting rationally is a matter of engaging in practical reasoning. If Tino wants to get his friend's attention, he might reason that raising an arm is a form of signaling and signaling is how one gets attention. He would thus conclude that he must raise his arm in order to get his friend's attention. For the subjectivist, this means that Tino *has* a reason—a psychological state wherein one is aware of the reason and uses it in an argument. Now, an objectivist might argue that Tino raised his arm *for the sake of* getting attention, but this merely highlights the fact that such chains of practical reasoning often issue in *intelligible* behavior.[41] Like the

[41] The exceptions that I have in mind are those chains of reasoning that people can construct which are nonsensical or incoherent, failing to count as arguments at all. For example, were someone to ask Bruno why he is raising *his* arm, he might respond, "Because last year during the holidays, someone decorated a

objectivist, however, the subjectivist *also* accepts the Cartesian metaphysic, and though she has a way of distinguishing the mechanical from the non-mechanical (by appealing to internal reasons) her position is subject to the same, familiar Cartesian criticisms. She can know when *she* is acting for reasons, but how will she know if anyone else is, *viz*., how will she know that others are not zombies? Does it follow that all behavior that does not proceed from reason is mechanical? What is more, it is not at all clear that human beings engage in the reasoning required for the traditional sense of action some or even most of the time.

2.11 | The Opacity of Self-Understanding, or Against the Subjectivists

In 1931, working out of the University of Chicago, the experimental psychologist Norman Maier wanted to learn more about how people perceive themselves to be solving a problem. To study this, he furnished a large room with a number of objects, such as pliers, poles, tables, chairs, etc. Also in the room were two cords hanging straight down from the ceiling, just barely above the floor. The participants were instructed to find a way to tie the two cords together, and the first thing that they discovered was that it was impossible to reach one cord while holding on to the other; they were just out of reach (Maier 1931, 182).

Whenever a participant discovered a solution, the researchers instructed her to find another, patiently waiting for the participant to find the "difficult solution," after which it would end. The difficult solution involved fixing a weighted object to the end of one of the cords to create a pendulum effect, and then pulling the other cord as close as

---

tree in Phoenix, and my uncle purchased a pony." Bruno might sincerely take these to be somehow involved in the raising of his arm, but they fail to make his behavior intelligible.

possible until the weighted cord could catch it during a swing. Once this solution was discovered, the researchers ended the experiment to ask the participants about it, wanting to know from where the idea came and what was crossing their mind when the solution occurred to them. If the difficult solution had not been discovered before a participant was ready to quit, Maier's team would offer two hints: one implicit and one explicit. The implicit hint was so subtle that it may as well be regarded as subliminal. The experimenter would simply brush past one of the cords, gently setting it in motion to create a slight pendulum effect. Whereas when the explicit hint was given, the experimenter would hand over a pair of pliers and say, "With the aid of this and no other object there is another way of solving the problem" (Maier 1931, 182–3).

At first glance, the results of the experiment were about what one might expect. Almost 40% of the participants (group one) were able to find the difficult solution on their own, and another 40% (group two) solved it only after the hints were given.[42] When participants in group two were asked whether they noticed the experimenter set the cord in motion (the implicit hint), most of them denied having seen it, some even expressing skepticism that it had any impact on their ability to solve the problem. In fact, one participant went so far as to insist that she had the idea for the solution from the beginning (Maier 1931, 186).

Curious of the results, Maier then wondered whether the issue was one of time. Maybe if the participants could experiment longer, they would have arrived at the solution on their own? Maybe the giving of the hint coincided with a natural progression in their problem-solving abilities, that it really did have no impact on them? A second

[42] The remaining participants were unable to solve the problem and had quit.

experiment was thus conducted, giving participants thirty minutes before they were permitted to quit. Over 80% of those who discovered the difficult solution without any hints did so within ten minutes. Even more striking, however, is the fact that *after* the implicit hint had been given to those unable to solve the problem, almost 80% of *these* participants solved it on average within 42 seconds! As Maier observed, "the effectiveness of [the implicit hint] can hardly be doubted" (Maier 1931, 187).

Just because a subliminal effect can *influence* how people solve problems, surely they still engage in practical reasoning, right? At least not those in Maier's experiments. Whether they belonged to group one or group two, nearly all of the subjects reported a "Eureka!" moment, unaware of how it developed, and rather than detail any arguments with premises and conclusions that crossed their minds, they instead referred to imagining analogous scenarios like a violent swinging of the cord, with one participant even describing "imagery of monkeys swinging from trees" (Maier 1931, 188–9). Given both—the suddenness and completeness of mentally apprehending the solution in addition to the failure to report the influence from the implicit hint—the results of Maier's experiments suggest that people tend to be out of touch with the operations of their own minds. How can a subjectivist possibly make sense of this? Are these participants simply not performing actions? Is their problem-solving behavior to be understood as mechanical?

When up against the ropes, the subjectivist will often invoke the existence of *implicit reasoning*, unconscious thought processes that otherwise resemble our conscious reasoning of drawing conclusions from premises. But this is little more than a *deus ex*

machina. What does it mean to be able to implicitly reason?[43] How can one distinguish implicit reasoning from other unconscious cognitive processes? If one cannot consciously exercise control over the reasoning process, how does implicit reasoning differ from a mechanical process? How do we distinguish implicit reasoning in humans from non-reasoning in animals? What does it mean to implicitly act for reasons?[44] Such a proposal creates more difficulties than it resolves. If actions are restricted to acting for reasons, then Maier's experiment invites the subjectivist to concede the possibility that much of human behavior is more mechanical than previously acknowledged.[45]

---

[43] Consider the following case involving infants. Dan Sperber along with two other researchers, Luca Surian and Stefana Caldi, repeatedly showed thirteen-month-old infants a video of an experimenter placing a piece of cheese behind a screen on the left and an apple behind one on the right. Next, they observed a caterpillar make its way to the piece of cheese on the left every single time. On the very last viewing of the video, the experimenter switched sides. The cheese was now on the right and the apple on the left. The caterpillar was then placed down, itself seemingly unaware that the switch had occurred. By noting the amount of time that the infants looked at the screen, the researchers were able to infer their expectations regarding what was happening, and in this last case, they expected the caterpillar to search for the cheese on the left. It did precisely that (Mercier and Sperber 2017, 94–6). The researchers concluded that their findings showed not only that "infants are capable of distinguishing between their own visual perspective and that of other individuals" but also that "infants are capable of attributing true beliefs to agents" (Surian, Caldi, and Sperber 2007, 583; 584).

What are we to make of this? As Mercier and Sperber put it, "Should we conclude that the infants in this study have a mental representation of a general psychological fact—that agents form beliefs and intentions rationally—and that the agents use this psychological fact as a premise in inference" (Mercier and Sperber 2017, 96)? Do infants implicitly understand Davidsonian intentional agency? Are they implicitly reasoning and forming expectations regarding how the caterpillar will act? Are caterpillars intentional agents since we can attribute beliefs and intentions to them?

[44] Mercier and Sperber express clearly their own skepticism of implicit reasons when they consider how the idea is used in the philosophical and psychological literature they have reviewed:

> On the other hand, when psychologists or philosophers talk of implicit reasons, they might mean either that these reasons are represented unconsciously or that they aren't represented at all (while somehow still being relevant). Often, the ambiguity is left unresolved, and talk of "implicit reasons" is little more than a way to endorse the commonsense view that people's thought and action must in some way be based on reasons without committing to any positive view of the psychological reality and actual role of such reasons. (Mercier and Sperber 2017, 118)

[45] This possibility is discussed in a little more detail in section 3.1.

2.12 | The Clarity of Behavior, or Against the Objectivists

The difficulty that faces the objectivist is finding a way to understand rational behavior that enables them to distinguish mechanical from non-mechanical behavior. By appealing to behavior that is intelligible in light of the environmental context, it is not clear on what grounds the objectivist can exclude the operations of alarm systems and thermostats from being included in the category of rational behavior. The alarm system reliably signals when an intruder has entered the property, and a thermostat reliably warms the home when the temperature becomes uncomfortable. Is this not precisely the behavior that we would expect from human beings given the same jobs?

Rather than understand behavior in terms of intelligibility *per se*, the objectivist's goal of pivoting away from Cartesian metaphysics would be better served by using the criterion of *teleological* behavior as her litmus test for the non-mechanical.[46] How does teleological behavior sort out alarm systems and thermostats from agents in a way that intelligibility does not?

As explained in section 1.9, invariant responses to stimuli (such as the behavior that stems from reflexes and fixed action patterns) cannot be construed as goal-oriented,

---

[46] Another problem with objectivism is that one must flesh out the behavioral norms that count as rational. Objectivists have adopted several approaches to this, defending anything form social norms to Darwinian norms of biological fitness. Glock defends *expected utility*, an economic approach that argues that rational behavior is the kind that pursues the most obviously advantageous course of action.

Given all possible actions at any moment, expected utility is determined by considering which of those actions has the highest probability of yielding the best outcome, all things considered. This makes for something much stronger than teleological behavior while remaining firmly committed to construing the behavior as rational. Glock writes, "Unlike machines or plants, animals can act for purposes, adopt purposes of their own, and adapt their behavior to circumstances in pursuit of these purposes" (Glock 2013, 389).

While expected utility theory attempts to overcome the mechanical divide, it still conflates *intelligible* behavior with *rational* behavior, and I am unconvinced that it successfully rules out machines and plants without an expected utility theorist adding a few auxiliary assumptions.

for these kinds of responses exclude the possibility for adapting to changes and adjusting behavior accordingly, for *learning*. No matter how the environment presents itself, invariant responses will always belong to the category of preprogrammed, mechanical behavior, hardly different from a chemical reaction such as rusting. There is nothing purposeful about it. The movement of the plant towards the sun is *reasonable* insofar as it makes sense to an interpreter, but it is not *teleological*.

The subjectivist is correct on this point: to appear to act rationally is precisely that, to *appear* to act rationally. Appearances alone do not give an interpreter justification to suppose that one is acting *with* reasons or *for* reasons. Interpreting performances in terms of teleological behavior accounts for the intelligibility of the behavior just fine. And of course it would appear intelligible! Against the backdrop of evolutionary theory, where organisms and other living things grow, develop, and evolve to adapt to their environments, why would the behavior appear otherwise? Rational behavior, by contrast, is better reserved for subjects who act *with* reasons. The brown Capuchin monkeys respond *teleologically* to the awareness of being treated unfairly, but they do not respond *rationally*. This comes at the cost of abandoning ascriptions of rationality to non-linguistic subjects, but as Maier's experiment shows, when it comes to action, there is reason to doubt whether human beings are even as rational, in the subjectivist sense, as we want to believe. The mistakes that the subjectivist makes are assuming that the *modus operandi* of human beings is rational and that acting intentionally is equivalent to acting rationally.

2.13 | Conclusion

In spite of Ryle's warnings, much of 20th century intellectual thought inherited Descartes' metaphysical framework and the Enlightenment faith in reason. Contemporary philosophers and action theorists, for example, have been working with faulty assumptions and grappling with the problems that extend from those assumptions. The gravest of these was the idea defended by Davidson: that action must be understood in terms of reasons. Reason, he rightly insisted, can function as a form of *causation*, but he also implicitly assumed that the only other form of causation was mechanical.

Because of this, there is a tendency to view subjectivist and objectivist approaches to interpreting rational behavior as incompatible. If one is to understand agency solely in terms of rationality, then how could it be otherwise? However, when agency is understood as a matter of *degree*, then it is plausible that both approaches offer valid forms of interpretation, even proving to be complementary. In the case of more primitive, basic agency—which follows from representational cognition as outlined in the preceding chapter—an objectivist interpretation will prove incredibly useful in understanding a subject's behavior; whereas, when more sophisticated forms of agency are exhibited, it becomes more useful to complement objectivist interpretations with subjectivist ones. This proposal proves to be more Rylian in spirit than the version of objectivism that was intended to rival subjectivism. After all, mentation exerts at least *some* influence on action, and quality of mentation hinges on cognitive sophistication. The error of the subjectivist lies not with positing the existence and influence of psychological states; it ultimately lies in its alliance with lingualism.

Is this a concession that reasons really do influence at least some actions? If so, how are we to understand action when reasons *do* intervene? And what does this mean for intention? When the brown Capuchin monkeys act on the basis of fairness, are they intending? How can one intend without reasons? To these questions we now turn.

Chapter 3: Reining in the Rational Horses

An ant is crawling on a patch of sand. As it crawls, it traces a line in the sand.
By pure chance the line that it traces curves and recrosses itself in such a way that it ends up looking like a recognizable caricature of Winston Churchill.

Has the ant traced a picture of Winston Churchill, a picture that *depicts* Churchill?
Most people would say, on a little reflection, that it has not.
The ant, after all, has never seen Churchill, or even a picture of Churchill,
and it had no intention of depicting Churchill.
It simply traced a line (and even *that* was unintentional), a line that *we* can 'see as' a picture of Churchill.

[…]

Suppose the ant had seen Winston Churchill,
and suppose that it had the intelligence and skill to draw a picture of him.
Suppose it produced the caricature *intentionally*.
Then the line would have represented Churchill.

— Hilary Putnam, *Reason, Truth, and History*

3.1 | The Language of Intentional Mysteries

Though Davidson believes that intention necessarily involves language, need this be the case?[47] What if, contrary to his view, language can help express mental objects, such as representations, but these objects themselves are not necessarily linguistic in nature? And what if these non-linguistic mental objects can influence actions?

To begin, it seems that we can and do perform actions in the absence of intentions. This is strikingly clear in cases of cognitive *dysfunction*. For example, the neurologist Iftah Biran has published work on a bizarre condition known as *Alien Hand*

[47] Here and throughout, no distinction will be made between *language*-use and *symbol*-use. I believe that both draw from the same cognitive well and serve the same functions. Whether one has in mind the idiosyncratic expressions used by a speech community, the mathematical notation of a calculus, or even a logo on the side of a building to signal ownership and property, I am taking all of these to fall under the broad umbrella of language. This is not to say that such a distinction between language and symbols is unhelpful as much as it is to say that, for present purposes, it does not contribute to or change this particular discussion.

Syndrome (AHS) which can affect patients with brain damage.[48] Such patients discover that their hand seems to act of its own accord, often mischievously. As the evolutionary psychologist Robert Kurzban summarizes a case, "the hand would wake the patient up, interfere with eating, and un-tuck shirts previously tucked in by the other hand" (Kurzban 2010, 10–11). Equally fascinating is a well-known case of *apperceptive agnosia*, which is a condition characterized by some kind of visual impairment coupled with an inability to recognize or name what is perceived. A neurological patient named D.F. is unable to do things like recognize a ball or determine the orientation of a mailbox slit, and yet she can still catch the ball and "deftly post a letter into the slit."[49] The behavioral phenomena that result from these conditions lead neuropsychologists to posit the existence of *zombie agents*, brain processes that circumvent conscious awareness and that usually involve "stereotypical tasks, such as shifting the eyes or positioning the hand" (Koch 2004, 3). But the idea of zombie agency is not restricted to cases of dysfunction; it is involved, for instance, even in those mundane situations where you "grab a pencil before you actually see it roll off the table" (213).

Habits are another example of actions in the absence of intentions. For instance, after washing my hands, I turn to the towel to dry them. This is something I do without any forethought, or any thought at all for that matter. If you stop to ask me what I am doing, I can certainly use language to provide a description, but no language is used in the performing of the action itself, *as* it happens. Whether they are habits or zombie

---

[48] See "Alien hand syndrome" (Biran and Chatterjee 2004). See also "The alien hand syndrome: What makes the alien hand alien?" (Biran et al. 2006).

[49] See "Perception and action in 'visual form agnosia'" (Milner et al. 1991); *Sight Unseen* (Goodale and Milner 2004); and *The Quest for Consciousness* (Koch 2004, 3; 217–220).

agents, actions without intentions make up the broader category defended in chapter one of *basic agency*, and so although the mosquito, desert ant, and bumblebee have neuroanatomies sufficient for a degree of representational cognition that makes basic agency possible, it is unlikely that they are cognitively complex enough to form intentions.[50] In other words, what they do does not require an intentional explanation.

But this is not the extent of the separation between language, reason, and action because, secondly, it appears that we can even *intend* without reasoning or language. Consider, for instance, what happens when one is engaged in a grueling exercise program. Once fatigue starts, it is not uncommon to *intend* to finish an exercise, such as a race, in spite of the fatigue. While it is true that in some cases a person might say to herself something to the effect, "I must get through this," expressing the resolve to continue, it is also true that this can happen in the absence of any thought, registering as a surge of energy and determination to complete the task at hand. As in the example of hand washing and drying, one can use language to offer a description of the circumstances and accompanying mental states—states like intense emotions or motivational imaginings of oneself or another finishing the task—but these linguistic expressions need not be used in either thought or speech *as the action is performed*. These observations were made in Maier's experiments as well.[51] Not only did most of the participants confess that they were unable to explain what thought processes lead up to the discovery of the difficult solution, one even *imagined* (which is a non-linguistic form

[50] The seeds for intention may be present, but having the cognitive capacity to entertain alternatives is a prerequisite for exercising intention in any meaningful sense of the word. This is discussed in section 3.4 below.

[51] See section 2.11.

of representation) an analogous scenario of swinging monkeys that enabled him to solve the problem.

Now return for a moment to Platt and Johnson's lab.[52] The rats in this experiment end up figuring out approximately how many times they needed to press a lever in order to access food, suggesting that they can remember and represent numerosity. Are these level pressings intentional? What about drying my hands after washing them? In my own case, I have no thoughts consciously in mind as I use the towel to pat my hands; the rats, by contrast, have some number-like thing somewhere in their minds. Which behavior is more complex? Which calls for an intentional explanation?

There are no easy answers to these questions, and perhaps one might suppose that neither count as intentional. If that is true, then theorists will quickly discover that human beings seldom act intentionally, as much of our behavior is of the hand-drying type, from the driving of automobiles to the manipulation of our electronic devices. As neurobiologist Christof Koch explains, the entire "point of training is to teach your body to quickly execute a complex series of movements—returning a serve, evading a punch, or tying shoelaces—without thinking about it" (Koch 2004, 3). Even some of the most vaunted displays of human excellence in contact sports occur when athletes report being in a state of *free flow*, which psychiatrist Robert Cloninger describes as a state where "people typically feel happy, alert, in effortless control, not self-conscious, and at the peak of their abilities" (Cloninger 2004, 83). Mihalyi Csikszentmihalyi, co-founder of positive psychology, goes so far as to describe it as a state where "people become so involved in what they are doing that the activity becomes spontaneous, almost automatic;

[52] See the introduction to part one, as well as section 1.5.

they stop being aware of themselves as separate from the actions they are performing" (Csikszentmihalyi 1991, 53). During such rapidly moving athletic competitions,[53] there is little time to think or reason, and when one does, it often proves itself to be a detriment to performance. And yet surely, such performances go beyond the requirements for basic agency.

If it is possible to intend without linguistic thought, then how are we to understand intentional action? What would distinguish it from a basic action? And what does this mean for rational actions?

3.2 | Intentionality Then

As he relaxes on a beach and sips his cup of coffee, Diego begins thinking about how nice it would be to have a slice of freshly baked cherry pie as well. His thoughts are *not* about slices of *frozen* cherry pie, slices of peanut butter pies, or even waves, grains of sand, phlogiston, or quantum loop gravity; it is simply a belief that a slice of freshly baked cherry pie would be the perfect complement to his otherwise extremely pleasant experience. What differentiates Diego's thought about cherry pie from, say, a thought about his loved ones, yogurt, or checkout time? Why are his thoughts about anything at all, and why *this* rather than *that*?

In philosophy of mind, there is a feature of mental states sometimes referred to as *intentionality*, and it is often invoked to explain how mental states can be differentiated from one another or be *about* things, that is, it seeks to explain how mental states can

[53] And it is not limited to this. Psychologist Jonathan Haidt explains that while it often occurs during activities that involve physical movement, even driving, it can even occur during "solitary creative activities, such as painting, writing, or photography" (Haidt 2006, 95). Perhaps many non-conscious experiences are those that occur during a state of peak flow?

refer to or represent things.[54] The fact that Diego's thought is directed towards slices of cherry pie rather than slices of peanut butter pie is due to Diego intending to think about the former rather than the latter. To help avoid confusion, it is important to keep in mind that *intentionality*—or, the directedness of the mind towards particular mental objects, like thoughts of loved ones, pets, or work-related matters—is not the same as *intention*—the ability to commit to something, paradigmatically an action.[55] So what is intentionality and how is it different?

Contemporary interest in the topic of intentionality was revived after the German philosopher and psychologist Franz Brentano published *Psychology from an Empirical Standpoint* in 1874. Among the topics discussed in his work, Brentano sought to make sense of the difference between physical phenomena and mental phenomena. *Phenomenon*, as he uses the term, is a philosophical word that refers to *appearances*, and as such, it is often contrasted with *reality*. For a current understanding of the difference, reflect on what happens during a visual experience. When you look at a rose and notice that it is red, your experience leads you to believe that the rose is *actually* red-colored; however, according to our best understanding of optics and physics, color is created by a complex interaction that results from electromagnetic waves reflecting off of the surface of objects such that when they strike the retina, they activate special cells in the eyes that transmit electrical signals down the optical nerve before they are processed in an area of

---

[54] My use of the term *mental state* and any variations on the idea will heretofore refer to *conscious* mental states, that is, the internal awareness of what appears in thought. Koch similarly advocates for using *consciousness* and *awareness* synonymously as well (Koch 2004, 2, n.2).

[55] In general, the noun *intentionality* and adjective *intentional* usually denote the directedness of our thoughts towards mental objects (i.e. *intentionality*), while the noun *intention* and the verb *intend* (along with its conjugations) denote the ability to commit to an action.

the brain known as the visual cortex.[56] Color is thus properly a psychological qualitative experience, or as it is called in philosophy of mind, a *quale*.

Depending on one's neurophysiology, objects can appear to an organism in all sorts of different shades and colors. As Koch explains:

> The much-cherished sense of color is a construct of the nervous system, computed by comparing activity in the different cone classes. There are no "red" or "blue" objects in the world. […] Color is not a direct physical quantity, as is depth or wavelength, but a synthetic one. Different species have fewer or more cone types, and therefore quite different colors for the same objects. For example, some shrimp even have eleven cone classes. Their world must be a riot of colors! (Koch 2004, 52)

From Brentano's philosophical perspective, our experience of colors would properly be classified as an experience of phenomena, and what *causes* that phenomena to appear, reality, is imperceptible, never appearing to conscious experience.

The mystery for Brentano was thus discovering how to distinguish between a mental phenomenon and a physical one. Surely the experience of thinking *about* the Louvre in Paris is qualitatively different from the experience of *actually seeing* the Louvre in person. The former appears to the imagination while the latter appears during sense-perception. How does one explain this? What criteria can be used to help distinguish them?

After surveying and criticizing the opinions of his peers on the matter, Brentano draws inspiration from the Scholastics and argues that one important difference comes down to something known as *intentional inexistence*, which would come to be known as *intentionality* (Brentano 1874 / 2005, 245). The idea is that mental phenomena have the

[56] This is a bit of an oversimplification to convey a point about phenomena. For a fuller and fascinating account of how visual perception works, see chapters three and four of *The Quest for Consciousness* (Koch 2004, 49–86).

unique property of referring to things *other than themselves*. When Diego thinks about the cherry pie, the idea of the cherry pie is *about* something and it metaphorically points the mind in the direction of it, as if to say, "Although I exist as an *idea*, that's not what's important. Don't look at me; look at the cherry pie!" By contrast, when Diego perceives the cup of coffee in his hand, the referent *is* the thing in his hand; for him to grasp what his sense-perception is *about*, his mind does not need to be directed to anything other than what directly appears in his sensory experience. If it were not for intentionality, our ideas would merely exist *as* ideas, indistinguishable from one another and lacking in content, making for a rather impoverished inner life. Intentionality, in other words, describes the fact that we can consciously perceive *inner representations* in the form of *mental objects*, such as ideas, imaginings, concepts, etc.

So what are these perceptions like? Do they even exist, or do we merely *think* they do? Are they visual? Might the lingualist be correct in assuming that human thought and, by extension, intentional action is fundamentally grounded in linguistic thought?

3.3 | Intentionality Now

While consciousness is still a mystery, what is less so are the neurobiological mechanisms involved in constructing some of our experiences. In 1994 and 1995, the electrophysiologist Nikos Logothetis along with his colleagues were able to train macaque monkeys to recognize a uniquely twisted paper clip such that they could reliably identify it not only at a particular angle but also amongst a group of misleading twisted paper clips that were similar in appearance.[57] While the monkeys were presented with the

[57] For more on this, see "View-dependent object recognition by monkeys" (Logothetis et al. 1994). See also "Psychophysical and physiological evidence for viewer-centered object representations in the primate" (Logothetis and Pauls 1995).

different paper clips and angles, the experimenters measured the firing activity of a set of neurons located in an area of the brain associated with the perception of visual objects (Koch 2004, 26).

What the researchers discovered was that some of these neurons, the same ones every time, showed intense activity when the right paper clip was presented at the right angle. As that angle shifted, the activity from these neurons decreased, and in the presence of the misleading clips, there was little to no neural excitations (Koch 2004, 26–7). For the monkeys, this meant that a particular set of neurons had become correlated with the perception of a particular paper clip, aiding them in picking that one out from others. Koch advises to keep in mind that "the monkeys were not born with paper clip cells. Rather, as the animals were trained to distinguish one bent wire frame from distracting ones, cortical synapses rearranged themselves to carry out this task" (27, n.10). Through frequent experiences, the brain thus begins creating consistent associations between a set of neurons and recurring aspects of those experiences, which neuroscientists call *neural representations*. It does not stop here, however.

Aspects of experience can not only be represented by sets of neurons, it is even possible for a *single neuron* to become associated with just one object or concept.[58] These are known as *grandmother* or *gnostic* neurons, so named because they could become correlated with and activate "every time you saw your grandmother, but not when you looked at your grandfather or some random elderly woman" (Koch 2004, 28). Koch, for instance, relays a case involving a neurological patient who was presented with a series of pictures while measuring the firing activity of a single neuron in the *amygdala*, which

[58] See *Integrative Activity of the Brain* (Konorski 1967).

is a pair of neuron clusters through which activity from the visual cortex and other areas can pass.[59] This neuron fired three times, each to a very different image of Bill Clinton (a line drawing, a portrait, and a group photo), but did not fire when the patient looked at pictures of others (29). While such neurons exist, this does not mean that a single neuron becomes correlated with every object that we experience; that would be highly inefficient. Rather, it seems that these types of neurons are reserved for the things that we encounter most often (29–30). For the rest of the objects of our experience, things are a little different.

As opposed to single neurons that represent highly specific persons and objects, like grandmothers, many other neurons become associated with broader features, such as facial identities, facial expressions, angles of objects, etc. When they fire together to form a pattern, they can create a *distributed representation* (Koch 2004, 30). Rather than have a single neuron associate with a blonde woman wearing glasses while sitting on the beach, our brains are furnished with sets of neurons that can come to represent broader features, such as colors, angles, directions, gender, hair, and so on and so forth. When the relevant neurons activate *together*, they jointly contribute to the experience of the blonde woman with glasses. The flexibility this affords over having a brain composed of nothing but grandmother neurons is enormous. As Koch points out, if one needed to recognize a few thousand faces, one would need a few thousand grandmother neurons correlated with each face, but through the magic of distributed representations, the recognition of a few

[59] See: "Category-specific visual responses of single neurons in the human medial temporal lobe" (Kreiman, Koch, and Fried 2000a); "Imagery neurons in the human brain" (Kreiman, Koch, and Fried 2000b); and "Single-neuron correlates of subjective vision in the human medial temporal lobe" (Kreiman, Fried, and Koch 2002). See also *On the neuronal activity in the human brain during visual recognition, imagery and binocular rivalry* (Kreiman 2001).

thousand faces is made possible through the number of potential patterns that can be created between neurons (32). For instance, mathematically speaking, there is the potential for creating over thirty million patterns with just the basic on/off firings of a scant twenty-five neurons!

All of the different features of our experiences are thus correlated with their own firing patterns. Interestingly, when it comes to *imagining* these things that we have experienced, *the same neural firing patterns activate*, not in a *perfect* but in an *approximate* way. This is why when you recall the image of the blonde woman in your mind, it is less detailed, fleeting, and probably somewhat inaccurate (Damasio 1994, 101; 105). Most of the neurons in that firing pattern activate, but not all of them and not to the same degree of intensity. This empirical observation also accounts for why damage to areas of the brain responsible for sense-perception can also impact the imagination. Damasio discusses one kind of damage in particular:

> In the condition known as achromatopsia […] local damage in the early visual cortices causes not only loss of color perception but also loss of color imagery. If you are achromatopsic, you can no longer *imagine* color in your mind. If I ask you to imagine a banana, you will be able to picture its shape but not its color; you will see it in shades of gray. (101)

So through a combination of neural firing patterns, memory, experience, and consciousness, we are able to represent to ourselves all sorts of mental objects in a variety of ways, and not just visual either but also auditory, somatosensory, gustatory, etc. The distributed neural representations that underlie these intentional mental objects are always being modified, strengthened, and weakened by new experiences.[60]

[60] Koch also advocates for recasting our traditional understanding of intentionality in neuroscientific terms. Such an approach has even become a new field of study, *neurosemantics*, which, he says, "focuses on how meaning arises out of brains shaped by evolution" (Koch 2004, 239).

The important takeaway from this brief neurobiological digression is that our best scientific understanding makes clear that our conscious thought processes are not merely linguistic in nature. Far from it. They are better characterized as *imagistic*, like imperfect copies of what we experience, provided that one does not restrict this adjective to just *visual* images. In fact, Damasio argues that imagistic thought actually *precedes* linguistic thought, that the latter is made possible by the former. He writes:

> Most of the words we use in our inner speech, before speaking or writing a sentence, exist as auditory or visual images in our consciousness. If they did not become images, however fleetingly, they would not be anything we could know. (Damasio 1994, 106)

This has important implications for our understanding of conscious mental acts, whether acts of reasoning, judging, affirming, loving, hating, desiring, etc. As Brentano was acutely aware, many of these acts are frequently directed towards mental objects with intentional inexistence, reminding us that "in judging something is affirmed or denied, in love [something is] loved, in hate [something] hated, in desire [something] desired, etc." (Brentano 1874 / 2005, 245). Diego's love for cherry pie does not consist solely in a mental experience of loving; it is a loving that is directed towards an idea about cherry pie. Considering this in light of contemporary neurobiology, it follows that these kinds of conscious mental acts (including intending) are directed primarily towards *imagistic mental objects*. To put it another way, *intention requires intentionality*.

So what could this mean for our understanding of intentional action?

3.4 | Ratting Out Intentional Action

Primates have evolved to have relatively similar brain structures to one another, so similar, Koch explains, that "it takes an expert to distinguish a cubic millimeter of

monkey brain tissue from the corresponding chunk of human brain tissue" (Koch 2004, 13). This leads a neurobiologist like Koch to believe that neither consciousness nor qualia are unique to human beings. He writes:

> It is plausible that some species of animals—mammals, in particular—possess some, but not necessarily all, of the features of consciousness; that they see, hear, smell, and otherwise experience the world. Of course, each species has its own unique sensorium, matched to its ecological niche. But I assume that these animals have feelings, have subjective states. To believe otherwise is presumptuous and flies in the face of all experimental evidence for the continuity of behaviors between animals and humans. (12–3)

This is certainly consistent with the assumption of biological continuity presented in chapter one. But even if animals do have conscious experiences and are capable of inwardly perceiving, what is the point of having such conscious mental states at all? Can these states affect a subject's ability to act? The answer is *yes*.

Koch along with his mentor, Francis Crick, speculate that the function of consciousness is to provide an organism with an executive summary that represents its present circumstances, highlighting only the most relevant details in its environment. The purpose of this, they argue, is to help with planning and determining courses of action, making for extraordinarily plastic behavior that can adapt to a variety of circumstances (Koch 2004, 233–5). If this is correct, when it comes to action, it is easy to see why it would be important for an organism to have *mental* representations as well, as they would serve as a point of contrast between the present and any possible alternatives, which would aid in both planning and problem-solving.

For instance, whenever one course of action fails or proves excessively difficult, having the ability to inwardly perceive mental representations would make it possible for an organism to use the imagination and memory to recall what was done and how it

might be improved upon. Similarly, rather than follow through with the first inclination that an organism has, which would make for terrible survival odds, mental representations would allow it to anticipate and virtually work through a scenario, drawing upon memory to see if the present bears any similarity to previous experiences and perhaps even testing it in the imagination before implementing it into action. Mental representations open up a world of alternatives, equipping an organism with the possibility of following through with one course of action rather than another as well as the ability to alter course mid-stream. This is precisely what it means to intentionally act; it is committing to the implementation of a mental representation, trying to make real what is conceived.

It was asked in section 3.1 whether the rats in Platt and Johnson's lab were acting intentionally whenever they pressed the lever the correct number of times to access their food or whether the drying of my hands after washing them could count as acting intentionally. We are now in a position to provide a hypothetical response. With respect to either case, it was an intentional action if and only if the action was guided by a conscious mental representation. Remember the participant in Maier's experiment who found the difficult solution by imagining monkeys swinging in trees?[61] To the extent that that mental image guided his action, what he did was intentional.

This is not to suggest that intentional action is an all-or-nothing matter. It is possible to perform an action guided by a mental representation that fails to result in the expected outcome; likewise, one might fail to act according to the mental representation—say, one gets surprised or distracted during the performance—and yet

[61] See section 2.11.

one might still manage to bring about the expected outcome, perhaps accidentally. In these cases, it would be best to make more fine-grained distinctions. For example, imagine that Robin is playing a game of billiards and is ready to sink the 8-ball into the corner pocket to win the game. Her intention is to bank the shot off of the side rail, but as she strokes the cue stick, her hand slips. The 8-ball indeed bounces off of the side rail (further down the table than she planned) before proceeding to slowly roll into the corner pocket. It is true that Robin intended to sink the 8-ball into the corner pocket with a bank shot, just not exactly in the way that it actually occurred.

3.5 | I Met a Linguistic Representation

The lingualists are correct to believe that the complexities of human language distinguish us from the rest of the animal kingdom. Even though Koch, for instance, disagrees with their assertion that language is a prerequisite for conscious experience, he has no reservations acknowledging that human language is importantly unique:

> Of course, humans do differ fundamentally from all other organisms in their ability to talk. True language enables *homo sapiens* to represent and disseminate arbitrarily complex concepts. Language leads to writing, representative democracy, general relativity, and the Macintosh computer, activities and inventions that are beyond the capabilities of our animal friends. (Koch 2004, 13)

So if intentional action is defined in terms of undertaking performances under the guidance of mental representations, what implications does this have for our understanding of the roles that language and reason play in acting?

The argument in section 3.3 is that the conscious experience of our inner lives is primarily imagistic. The images are types of representations that distill information to us, information that proves extraordinarily helpful in interacting with the environment, especially when it comes to planning. In a way, using mental representations is akin to

accessing navigational software. If you are traveling from Pittsburgh, PA to Washington, DC, one option is to jump in the car and hit the road without any forethought whatsoever. But is this the most efficient way to do it? There could be accidents, construction, or any other number of obstacles impeding your journey. There is even a chance that you get lost. By accessing navigational software, you are presented with an overall summary of the current conditions as well as a number of alternative routes that you can travel, complete with an estimation of how much time each would take and what sorts of obstacles to expect along the way. Mental representations accomplish this same feat for us.

But now a slight modification to the thesis that conscious mental life is primarily imagistic is in order; it is *exclusively* imagistic, governed wholly by mental representations. How is this possible? As it turns out, language is just another form of representation, but it is unique on two counts: it allows for greater representational flexibility and it allows for the representation of much more precise information.

Imagine walking along the beach at sunset. You can picture the shades of pinks, blues, and reds decorating the sky; you can feel the slight breeze of the warm air and the crunchy granules of sand below your feet; you might even be experiencing tranquility, satisfaction, or even pleasure. No matter how much you picture this, however, describing it in words improves the accuracy exponentially. Was that positive feeling one of joy, gratitude, amusement, awe, relief, or some combination? Are the shades of pink more along the lines of salmon, shrimp, coral, blush, or some other shades? It is language that enables us to make these fine-grained distinctions, and even more important, *communicate* them to others. Consider how difficult it would be to convey the beach

scene without the use of language. One would have to recreate it with tools, like paints, cameras, or graphic design software.

Unfortunately, there *is* a tradeoff when language is used to represent. What we gain in accuracy, we lose in quality. We can pick out those particular shades of pink with *words* at the cost of forfeiting the *image* of those shades. This is why it is notoriously impossible to describe basic sensations in the absence of analogies. How do we explain what it is like to see red to someone who is blind? We can pinpoint the electromagnetic wavelength that corresponds to that color—even poetically relate it to objects like flowers, sensations like heat, or feelings like love—but we cannot recreate what it is like to visualize in perception or imagination the image of redness.

This view of language explains why Platt and Johnson's rats discussed in the introduction to part one exhibited scalar variance in their pressings of the lever in proportion to the total number required to access the food. Although the rats are capable of mathematical cognition, their mental representation of number is one that does not rely on language, leading to an imprecise representation of it. While *something-like-ten* is certainly different from *something-like-one* and *something-like-twenty*, for the rats, it becomes increasingly difficult to separate *something-like-ten* from numbers that are closer in quantity, a difficulty that worsens as the numbers grow larger. In the case of humans who have the luxury of discretely representing number with symbols and words, there is no question that 10 is different from 11 as well as every other number.[62]

[62] For more on this, see "Mathematical cognition" (Gallistel and Gelman 2005, 562 – 565). It also discusses experiments where animals demonstrate abilities to add, subtract, multiply, and divide these fuzzy mental numbers, as well as the fact that human beings appear to still use fuzzy numerical representations as well, particularly in those cases where we are expected to process and recall number quickly (566–9; 571–2). For instance, if someone were to rapidly press a button too quickly for us to count, when asked how many times it was pressed, most of us will recall a vague number, such as "*about* twenty times."

Recall now that Brentano believed that the peculiar feature of mental objects is that they are necessarily about things other than themselves. The idea of a beach is never about *the idea*; it is always about *the beach*. Language adds another layer to this representational cake. Whereas an imagined scene of a beach is a kind of idea that represents a beach, all of its contents can be summarized and represented with a linguistic symbol, B-E-A-C-H. Though our imagined beach might include sights, sounds, and motions, all of this information is compressed into our single word for this, ready to be extracted by any other language-user, including ourselves. What is more, the representational capacity of language seems limitless. We can represent not just visual perceptions and ideas with words but *any* sensory modality. We can represent numbers, relations, emotional states, possibilities, and states of affairs. But perhaps the greatest and most fascinating capacity is the ability to hold multiple linguistic representations together in a higher linguistic representation known as a *sentence*. It is one thing to be able to represent the beach scene with the symbol B-E-A-C-H, but by being able to construct these higher representations, we can refine and make precise the linguistic representations themselves with phrases such as, "the beach at sunrise" as opposed to "the beach at sunset," or sentences such as, "I walked the beach in Clearwater" as opposed to "I walked the beach in Nags Head."

Mercier and Sperber use the term *metarepresentation* to describe representations that can represent other representations, much like how a linguistic symbol can compress the information from sense-perception, imagination, or memory (Mercier and Sperber 2017, 92; 154). There is not a better description of the representational power of

language. To see what this will mean for reason, it is first necessary to take a large step back.

3.6 | Foraging for Reasons

It is well-known that the ancestors to modern human beings were foragers, and this lifestyle had a profound impact on how humans evolved. The life of a forager is nomadic, a wandering of the environment in search of areas where food is bountiful, but a forager does not travel alone. Instead, they form close-knit communities consisting of anywhere between 20 to 50 members (Simler and Hanson 2018, 46). A size of a foraging community is no accident. The odds for survival increase when communities grow larger, finding themselves better able to defend against predatory and environmental threats, but that also comes at the cost of traveling efficiency. It is much more cumbersome for a group of 250 to coordinate and travel through a harsh environment than it is for a group of 25. In a land where food can become scarce, it is important to be mobile.

As a result of having a smaller community in conjunction with living under less-than-ideal circumstances, foragers come to rely on one another in important ways, ways that can very literally mean the difference between life or death if a fellow forager fails in her duties. Author Kevin Simler and economist Robin Hanson explain:

> To be without a band for more than a short time is effectively a death sentence. Everyone is expected to try to provide for themselves and to pitch in and help each other as they're able (no freeloading), but they can reasonably expect help from the rest of the band if they fall on hard times. (Simler and Hanson 2018, 46)

This mutual interdependence gives way to a "fierce egalitarianism" within the community, such that those who are regarded as wise do not compete with one another but rather "relate to each other as peers and equals" (47). The purpose of such

egalitarianism is to both keep watch for and serve as a check on individual ambition and desire for dominance, factors that might increase the benefits for one at the expense of the rest of the group (48).

To enforce the will of the community and correct any problems with individuals, foragers use an effective mechanism: ostracism. Whenever a group believes that a fellow member of the community is not holding up her end of the bargain or has become a threat to their well-being, she is expelled, a fate that may as well be equivalent to death. Given that life alone as a forager is close to impossible for any significant period of time, it is thus of the utmost necessity that a forager does her best to keep her community happy, and she accomplishes this by managing her *reputation*, a social phenomenon that serves as a representation of an individual's worth to the community. Psychologist Patrik Söderberg and anthropologist Douglas Fry explain:

> For the great majority of nomadic foragers the threat of ostracism is more than enough to make them conform to unspoken social norms or to yield to group opinion and reform their ways (Marshall, 1961). This is clearly illustrated in Yahgan society, where anyone threatened with ostracism or even with earning a bad reputation "promptly hastens back to the path of virtue" and with "exaggerated zeal" tries to change any bad opinions about himself (Gusinde, 1937, p. 930). (Söderberg and Fry 2017, 267)

Might this anthropological context help provide an evolutionary explanation for the role of reason? Perhaps. While the western intellectual tradition has been persuaded throughout much of its history that the function of reason is primarily to track truth, it has been well-known amongst psychologists that it does a poor job when put to this task in experimental settings (Mercier and Sperber 2017, 4). Worse still, other experiments suggest that reason is far more interested in *explaining* actions than *causing* them, pointing in the direction of a very different function. It was suggested in section 3.1 that a

close look at cognitive dysfunction can sometimes illuminate cognitive function. This turns out to be true in the case of reason as well.

3.7 | Fabulous Confabulations

In the middle of the brain lies a bundle of neural fibers known as the *corpus callosum*. These fibers make cross-communication possible between two brain hemispheres that otherwise lack a way to effectively transmit information between one another.[63] It is important to keep in mind that even under normal conditions, each hemisphere regulates and controls the functions on the opposite side of the body. So the *left* hemisphere, for example, is responsible for the motor control of the *right* hand. In addition to this, only one of these hemispheres will become dominant and the other non-dominant, largely because, Damasio hypothesizes, it is better to have "one final controller rather than two," which would make for significantly many more instances of self-defeating behavior and greater difficulty in coordinating movement with multiple limbs (Damasio 1994, 66). Imagine trying to walk while each leg attempts to do its own thing; it would not work out so well.

The other important thing to remember is that whenever a hemisphere becomes the dominant one it will also house the area in the brain responsible for language processing (Koch 2004, 290). Since most people tend to be right-handed, which is an expression of left hemisphere dominance, and language is housed in the dominant hemisphere, a pervasive myth has arisen within popular culture that the left side of the brain is for language and logic while the right side is for creativity.

---

[63] There is a smaller bundle of fibers that also connects the two hemispheres known as the *anterior commissure* (Koch 2004, 288).

It was in the 1940s when a *corpus colostomy* was performed, severing the fibers that connected the two hemispheres in a last-ditch effort to alleviate seizures in epileptic patients who had run out of options. What puzzled the clinicians was that they were expecting some significant changes when the patients recovered, anticipating that life would be very different and difficult, but to their surprise, the patients reported feeling pretty similar with no known deficits in perception (Koch 2004, 289). Why did nothing change? Or did it?

After the pioneering work of neuropsychologist Roger Sperry who believed that animals who had undergone a similar procedure had separate minds, neuroscientists Joe Bogen and his student Michael Gazzaniga tested Sperry's ideas in humans (Koch 2004, 289). What they discovered is that when an object was placed in the right hand of a split-brain patient whose eyes were closed, he was able to successfully identify it, but when the object was placed in the *left* hand, he suddenly found himself unable to say what it was. Because language is housed in his left hemisphere and that hemisphere also controls his right hand, this result makes neurobiological sense. What happened next, however, is a little more unexpected.

The researchers showed the patient a picture with a number of objects on it, asking him to point to the image of the one he had just been holding in his left hand. Strangely, his left hand was able to accurately point to the correct object but not his right. Stranger still, when the researchers inquired as to why his left hand was pointing at that particular object, not only was the patient at a loss for words, but they observed that he "often confabulates and invents some explanation to cover up the fact that he has no idea why his left hand did what it did" (Koch 2004, 291).

These split-brain experiments are often used to empirically demonstrate that these patients somehow have two separate minds, but what is more interesting is what is actually happening right in front of the researchers. Rather than simply plead ignorance, something related to the language processing center finds it necessary to *justify* the behavior of the body, as if the patient *needs* to explain himself to them even if he does not understand what he himself is actually doing, perhaps in order to salvage his reputation as a fellow human being. If this does not yet sound plausible, then consider these other experiments involving participants who have *no* apparent cognitive dysfunction.

Using electroencephalography (EEG), it is possible to painlessly record the electrical activity of various regions in the brain by attaching electrodes to a number of areas around the head. Depending on the wave pattern output, a neurologist or other specialist can interpret what is happening where. One such discovery using EEG is known as *readiness potential*, indicated by a particular kind of change in the pattern, and it is correlated with activity in the cerebral cortex when "an action is being prepared" (Nørretranders 1998, 215). For example, if you move your finger, this will show up as activity in the cerebral cortex on the EEG as well. When this imaging tool was put to use in experiments,[64] researchers discovered that this activity occurred almost an entire second *before* any muscles started to flex for the movement (214–5). That in itself is interesting enough, but a neurophysiologist by the name of Benjamin Libet decided to take the experiment a step further.

[64] For the earliest work on this, see "Hirnpotentialänderungen bei Willkürbewegungen und passive Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale" (Kornhuber and Deecke 1976).

If the supplementary motor area (SMA) in the cerebral cortex activates an entire second before our muscles begin to work when we perform an action, Libet wanted to know whether a similar discrepancy existed between our *conscious* decision to act and the activity in the SMA. On the one hand, if we decide to act at the moment the muscles ready, then our decision to act occurs *after* the SMA activity starts, suggesting that our decision is simply an illusion. On the other hand, if we decide to act an entire second before every action we perform, we are looking at a pretty significant delay between thinking and acting, a delay made all the more curious because we do not seem to be aware of it (Nørretranders 1998, 216). We will be hard-pressed to find any cognitively healthy person admit to experiencing these delays before every action. Imagine the difficulty in composing an e-mail if a one second delay must occur between every stroke of the key. So which of these alternatives is it?

By placing participants in a chair and asking them to look at the face of a clock that featured a revolving spot that completed a rotation every 2.56 seconds, Libet hoped to gather insight on when participants felt like they were making a conscious decision to act. In addition to this, he used an EEG to measure activity in the SMA as well as electrodes to measure muscle movement. As a control, he contrasted their verbal reports of when they made their conscious decisions to act with the measurements of the muscle movement in their hands, and this verified what we would reasonably expect, a conscious decision shortly before (0.20 seconds) any movement (Nørretranders 1998, 218). When all three data points were recorded, Libet was astonished. Our conscious decision to act occurred on average 0.35 seconds *after* activity in the SMA was recorded. In other words,

from the perspective of the researchers, the participants were making their conscious decisions after the brain had initiated the start of an action.

Many have used the results from these experiments and others like it to argue that the philosophical concept of free will does not exist. Some go so far as to say that this proves that consciousness is an epiphenomenon; it provides an organism with first row seats to the theatre of experience but does not permit the organism to do anything about it. Others (Libet included) have more modestly proposed that while free will may not exist, the point of conscious experience is to give the organism an opportunity to veto the action (Nørretranders 1998, 243). Here is a different interpretation of Libet's experiments though. Asking a participant to report when she consciously decides to act will recruit the language processing center of the brain to get involved, and if one of the primary functions associated with this part of the brain is to *justify* (rather than *cause*) the behavior of the body, then of course it will activate *after* an action has been initiated elsewhere! And if the split-brain patients are any indication, it may not even be accurate in its reports of what is happening.

By no means are these kinds of results exclusive to split-brain patients and Libet's lab. Consider what occurred during the landmark study conducted by social psychologists Richard Nisbett and Timothy Wilson.[65] Participants were shown four pairs of nylons and asked which of the four they preferred. Unknown to the participants was the fact that all four of the nylons were identical to one another. In spite of this fact, the participants consistently selected the nylons that were furthest to the right of them. Kurzban summarizes their results well:

[65] See "Telling more than we can know: Verbal reports on mental processes" (Nisbett and Wilson 1977).

> That is, the *position* of the object is what seems to be driving the choice. However, just like split-brain patients, the people making the choice weren't able to say what really caused them to make the choice that they did; instead, they referred to some feature of the panty hose, such as color or texture, even though these were the same for all four. (Kurzban 2010, 43)

As with the other experiments, the researchers used the results to underscore a different idea. In this case, Nisbett and Wilson believe that this demonstrates "there may be little or no direct introspective access to higher order cognitive processes" (43). While that is certainly an interesting insight, it does not answer the question *why* people feel a need to explain their behavior in the first place, even if they lack introspective access to the actual motivating factors behind their decisions. The reason why is that the participants were attempting to *justify* a behavior that already occurred rather than use reasoning to *make* a decision.

This same behavior was even exhibited by participants in another study, one that had rocked the world of oenophilia and wine-tasting. Researchers Gil Morrot, Frédéric Brochet, and Denis Dubourdieu invited sommeliers to taste glasses of wine (one white and one red) and describe the aromas and flavors. What the sommeliers did not know about the red wine is that the researchers had simply artificially dyed the white wine with an odorless red dye. This experiment was designed to test the hypothesis that visual perception influences what we taste, and sure enough, the sommeliers described the "red wine" with tasting notes characteristic of actual red wine (Morrot, Brochet, and Dubourdieu 2001, 315–6). While the researchers believe that these results confirm their hypothesis, it also consistent with the suggestion made throughout this section, that the sommeliers felt compelled to defend their reputation by justifying to themselves and others that they were tasting a red wine. They were the experts, after all.

3.8 | Damage Control

It is no secret that the capacity to construct shareable symbols that can discretely represent nearly anything, i.e. *language*, plays a vital role in reasoning, understood as the cognitive ability to draw inferences from premises.[66] It is these special statements known as *premises* and *conclusions* that uniquely serve this purpose. The former (premises) function as symbolic inputs for an inferential process that yields the latter (the conclusion) as an output. Though both count as statements, they are differentiated from one another precisely by their function. Thus, any conclusion that is yielded from one inference can be taken up and used as a premise for another by shifting the role that it plays in a chain of reasoning. Here is an example with two syllogisms:

Argument A		Argument B	
P_1	All candy is sweet	P_1	All candy is sugary
P_2	All sweet foods are sugary	P_2	All sugary foods are unhealthy
C	All candy is sugary	C	All candy is unhealthy

Given the two premises in argument A,[67] a reasoner can draw the conclusion that all candy is sugary, and once that statement is inferred, one then might opt to employ it in a second argument, argument B, to make additional inferences.

This raises a question though. If reasons are just linguistic objects, then what makes a reason different from any other linguistic object? In other words, what marks the difference between the sentence, "Shelly went to work," as, say, an expression of a fact and as a reason? There has already been a hint at the answer; it has everything to do with the functional role of the statement in question, that is, *how* the speaker uses the

---

[66] The use of the term *symbol* is meant to be understood broadly. Literal symbols, such as the small raised circle ° that designates a unit of measurement known as *degrees*, as well as words and sentences are to be counted as symbols in this discussion. The primary cognitive role of a symbol is *metarepresentational*.

[67] Alas, valid but not sound. Some sweeteners, notably those such as stevia or Splenda®, are sweet without being sugary.

statement. Any time it is used as a reason, the speaker is intending to *justify* something—perhaps some other belief that she has, action she is planning, or action she has already performed. Thus, to use "Shelly went to work" as a reason is to use it in the hopes of justifying something else, such as her absence, her lateness, her busy schedule, her missed appointment, etc. It supplies an answer to an implicit or explicit *why*.

But if reason is seldom used in *causing* actions and primarily used for *justification*, why did we develop an ability to reason at all? One provocative and attractive answer has been given by Mercier and Sperber: *reputation management*. They write:

> Whatever humans do is likely to contribute for better or worse to the way they are seen by others—in other words, to their reputation. […] By explaining and justifying themselves, people may defend or even improve their own reputation. By failing to do so, they may jeopardize it. (Mercier and Sperber 2017, 123)

Reasons come into the picture, they argue, in order to play the role of explaining and justifying oneself to others, to negotiate one's reputation with the rest of the group, and they liken reason itself to a lawyer that not only *defends* one's actions but also *recommends* and *advises against* courses of action that may impact one's image or standing within the community (123–4). This not only establishes an evolutionary continuity with our foraging roots in small, egalitarian communities, it also accounts for the fact that reason is so often employed *after* an action rather than before.[68] Still, even when reason is used beforehand, Mercier and Sperber suggest that it is exercised in *anticipation* of needing to justify oneself to others (123). When viewed in this way, the practice of giving reasons prior to undertaking an action was likely to accomplish two goals: (1) to seek approval from the community in advance, and (2) to try to establish an

[68] See the preceding section for several experimental examples.

idea or practice as a new norm for oneself and others within the community. Reason is thus a social tool, through and through.

3.9 | Lowered Expectations

This is not to suggest that there is no such thing as acting *for* reasons; rather, acting for reasons turns out to be something very different than has been traditionally supposed. If this view is correct, then there is little difference between recruiting reasons before and after an action apart from the practical benefit of decreasing the likelihood of punishment from the community when used in advance. There is a clear tradeoff, however, for the cost of reasoning in advance might be the inhibition of an action that would have otherwise been performed. Think of this in terms of risk mitigation. While taking on risk is usually dangerous, it is well-known that some of the riskiest endeavors are also the most rewarding. By reasoning in advance, one lessens the risk at the possible cost of forfeiting a better reward, and so it is not always better to reason *before* but rather to seek out a defense *after* an action has been performed. As the old adage goes within business circles, "It is better to beg for forgiveness than ask for permission." Because both uses of reason are primarily aimed at reputation management, as counterintuitive as it might seem, *both* ought to count as acting for reasons.

Although to act for a reason, especially in advance of an action, can be construed as acting intentionally insofar as it is another case of using a mental representation to guide an action, it is the reputational function of reasoning that sets apart acting for reasons from other forms of intentional action. This is clearest when one considers that the need for justification in the first place only arises against the background of *normativity*, the practice by a community of codifying standards for what counts as

permissible and impermissible. As noted in section 3.6, foragers enforce their norms with the threat of ostracism. How does normativity change things? Hobbes can provide some clarity here.

In *Leviathan*, Hobbes declares that all human beings are equal, but not in the way that we might expect by today's standards. Rather, he explains, "though there be found one man sometimes manifestly stronger in body or of quicker mind than another," when it comes down to it, "the weakest has strength enough to kill the strongest, either by secret machination, or by confederacy with others that are in the same danger with himself" (Hobbes I.13.1). Rather than advance the indefensible idea that people are *naturally* equal—as he himself notes, we can always find someone stronger, faster, or smarter than ourselves—or appeal to a noble metaphysical quality (like dignity), Hobbes very bluntly emphasizes that anybody can be killed. It is of course always possible that another human being can concoct and successfully execute an unforeseen plan against you, but the strongest weapon that works against the individual is the *community*, the "confederacy with others." It is the power that the community wields that ensures the enforcement of norms, and because of that community's power, there is every motive to negotiate one's reputation with them through the practice of giving and asking for reasons, the practice of justification.

The way that this is different from non-normative animals is that they do not seem to have any incentive to justify their behavior; they do not have to give an account of themselves to their fellow community. Simler and Hanson explain the importance of this:

> It's important to distinguish what humans are doing, in following norms, from what other animals are doing in their related patterns of behavior. An animal that decides not to pick a fight is, in most cases, simply worried about the risk of getting injured—not about some abstract "norm against violence." Likewise, an

> animal that shares food with non-kin is typically just angling for future reciprocity—not following some "norm of food-sharing." The incentives surrounding true norms are more complex. When we do something "wrong," we have to worry about reprisal not just from the wronged party but also from third parties. Frequently, this means the entire rest of our local group, or at least a majority of fit. […] *Collective enforcement*, then, is the essence of the norms. (Simler and Hanson 2018, 49)

To act rationally is thus to be sensitive to norms, to *care* about what *others* might think about oneself, and this is both promoted with and enforced by the means available to the community.[69]

3.10 | Conclusion

Through the study of cognitive dysfunction and ethology, it was shown that one can gain some important insights into what it means to act. In particular, two such insights were presented in this chapter, and both were surprising. The first is that there is more conceptual separation between acting, intending, and reasoning that has been traditionally supposed.

When the ideas of mental representation and intentionality were examined in more detail and considered along with empirical research, it emerged that while acting for reasons can still be characterized as a form of intentional action, the converse does not hold; not all intentional action occurs for reasons. In addition, there are even some actions (basic actions) that occur in the absence of intentions altogether. This helps to reinforce the argument throughout part one, that agency ought to be understood as a matter of degree.

[69] Most notably, as Simler and Hanson argue, *deadly weapons*. It was especially long-range weapons that guaranteed even the strongest individual was able to be taken down (Simler and Hanson 2018, 49–51).

The second insight presented in this chapter is that the function of reason is best understood when it is historically situated within the context out of which it came to be. When we turn to its origins, there is nothing glamorous or sexy about reason, at least as far as Enlightenment thinking goes. It developed out of the political dynamics that obtain between individuals and communities where one's membership is all but secure, thus creating a need to negotiate one's status *within* that community. So rather than leave a community with no recourse but to absorb the damage from a problematic individual and rather than leave an individual at risk for ostracism or death for no apparent reason, the development of argument through language helped shine a light on what was happening, creating a shared arena wherein the members of a community are able to more precisely represent their concerns and problems in a public manner. Members thus could not only talk about these issues but also present them to a jury, their *community*.

This view of reason, dubbed *interactionist* by Mercier and Sperber, paints a very different picture of rational action. It contends that reason did not evolve to be truth-oriented except perhaps as a secondary function of being socially-oriented; after all, when things work out well, the community will value your contributions more. Ultimately, however, to act for reasons is to act for reputations. But is this the *only* function of reason? Maybe not. Before turning to that, however, it is important to take a detour through the dark side of human behavior, the irrational. This will reveal a little more about human minds and human nature, loosening the grip of Enlightenment thinking once and for all.

Part II. The Mind's Mariana Trench

Recalling a period in his youth, Augustine shares with his readers of *The Confessions* a time when he committed an act of theft.[70] One evening while with a group of friends, they decided to sneak into a nearby garden and steal as much fruit from a pear tree as they could. When stories like this are usually shared, they are typically accompanied with a moral by way of some terrible consequence or stroke of bad luck, such as injury, disease, discovery by authority, etc. What happened to Augustine this evening? Did the owner of the tree catch them in the act? Would he learn that someone starved as a result of their mischievous undertaking? Maybe a friend fell victim to an illness after eating one of the pears?

To first-time readers, it may come as a surprise to learn that Augustine does not relay that there were any consequences. The theft was a success, and they reveled in it. In fact, they did not even undertake this crime *for* the fruit itself, which "was desirable neither in appearance nor in taste" (*Confessions*, II.iv.9). While some was eaten, many were thrown to pigs. Augustine even takes care to stress to the reader that his family had owned plenty of their own fruit and it was of much better quality; instead, this was done "to enjoy the actual theft and the sin of theft" (II.iv.9).

What were his motivations then? Surely he did this for some reason, or so Enlightenment thinking would lead us to believe. Here is Augustine's own account:

> Behold my heart, O Lord, behold my heart upon which you had mercy in the depths of the pit. Behold, now let my heart tell you what it looked for there, that I should be evil without purpose and that there should be no cause for my evil but evil itself. Foul was the evil, and I loved it. I loved to go down to death. I loved my fault, not that for which I did the fault, but I loved my fault itself. Base in soul was I, and I leaped down from your firm clasp even towards complete destruction,

[70] For a fascinating analysis on why this story in particular was chosen by Augustine, see: "Augustine's Confessions and the Source of Christian Character" (Thompson 2012, 505–522).

> and I sought nothing from the shameful deed but shame itself. (*Confessions*, II.iv.9)

From a consideration of practical reasoning, Augustine must have been thinking along these lines—if he was reasoning at all:

P_1 I desire evil
P_2 Stealing these pears is evil
C I desire stealing these pears

This certainly rationalizes Augustine's action. Or does it? Why does he desire evil after all? What advantage can it possibly confer? He does not want the thing he is stealing. Is it because he derives pleasure from it? But then why does evil perversely cause him to feel pleasure? He goes so far as to admit that he sought in this act *shame itself*, and so even if he feels pleasure, it seems to be bound up in these painful feelings as well.

One argument might be that there was something good, something desirable to be had by performing this heinous act. When people debase themselves in this way, they merely fixate on what is good in the act, failing to see that the bad outweighs any good to be had. Augustine reflects on this position and sees the attractiveness in it:

> A man commits murder: why did he do so? He coveted his victim's wife or property; or he wanted to rob him to get money to live on; or he feared to be deprived of some such thing by the other; or he had been injured, and burned for revenge. Would anyone commit murder without reason and out of delight in murder itself? (*Confessions*, II.v.11)

But when he submits his own deed to be examined by this theory, something does not add up. Either the theory is wrong, or else he is confused about his own motivations (and possibly both). As he surveys possible candidates for what motivated his behavior, he rules out the act of thievery itself, the appearance of the fruit, the taste of it, and even the desire to rebel (II.vi.12–3). As frustration mounts, he exasperates:

> O rottenness! O monstrous life and deepest death! Could a thing give pleasure which could not be done lawfully, and which was done for no other reason but because it was unlawful? (II.vi.14)

As a final consideration, Augustine entertains the idea that maybe it was because he enjoyed the company of his peers, and while he admits that this theft is not something that he would have done had he been alone, he insists that his "association with the others was itself nothing" (*Confessions*, II.viii.16). Rather than blame a phenomenon like peer pressure, Augustine looks at the matter very differently. What it proves is that he is not alone in this madness; others did the same thing and are just as confused and broken as he. He thus never manages to attain the self-understanding that would make sense of this event, accepting the possibility that this might be a part of what it means to be human:

> Who can untie this most twisted and intricate mass of knots? It is a filthy thing: I do not wish to think about it; I do not wish to look upon it. I desire you, O justice and innocence, beautiful and comely to all virtuous eyes, and I desire this unto a satiety that can never be satiated. (II.x.18)

By reflecting on this incident, Augustine became convinced of two things: his actions do not always make sense (i.e. he is broken in some way), and this problem is not unique to him.

What Augustine calls *evil*, we might refer to as *irrationality*, and it is just as senseless today as it was then. More problematic still is the fact that Enlightenment thinking encourages us to look away from it. "Surely people are rational," it says, "and if they appear otherwise, there is simply some underlying motive that we as observers have failed to identify, one that does not cohere with the norms of society." When all explanation fails, Enlightenment thinking sustains our hopes by promising that not only should nobody have a broken mind, but if there is one, it can be easily repaired through surgical or medical interventions.

Both of these Enlightenment explanations are tempting because each gets *something* right about irrationality. Medical specialists have observed for some time that our behavior can be adversely influenced by damage to different regions of the brain. This thus supports the mechanical thesis that broken minds can be repaired. But was it brain damage or some other form of cognitive dysfunction that inclined Augustine to act as he did? Is that the reason why we act against better judgment or violate social norms?

Similarly, there is little doubt that irrational behavior tends to be socially deviant, but are these two concepts equivalent to one another? Are they synonyms for the same phenomenon? Is, for example, traveling in excess of the speed limit or running late for an appointment irrational? Both are socially deviant insofar as they violate norms of the community. The former is a violation of a legal norm and the latter a cultural one, and yet they appear to be in conflict with one another in this case. If a behavior, such as speeding, is a violation of one norm (legal) but not the other (cultural), is it irrational or not? If the legal / illegal is not identical to the permissible / impermissible and the good / bad—each a set of binary concepts that describes types of norm-abiding and norm-violating behavior—then *which* is the irrational and why? Or do they all fall under the broad umbrella of the irrational, even when they contradict one another? There must be some other standard that an Enlightenment thinker can appeal to. And there is: *rational* norms.

Davidson is one who argued that irrationality was best understood as a deviation from rationality, a so-called "failure within the house of reason" (Davidson 1982, 169). I will refer to this thesis—that irrationality requires rationality and is defined as rationality gone astray—as the *Deviant Rationality Thesis* (DRT). While it is attractive and intuitively plausible, there are two significant issues with this view, even beyond the

issues of how a DRT theorist is supposed to define rationality and catalogue which kinds of behavior should count as rational.

One, DRT views irrationality as little more than a descriptive term for deviation from these norms, accomplishing little more than serving as a label for the *fact* that someone is failing to conform to a norm. As a result, this makes it unclear exactly why irrationality poses a *psychological* problem. Why is one failing to conform? Why is that, in itself, a problem? What difficulty does that create for *the person*?

Two, and perhaps more importantly, if reasons are taken to be those beliefs used in argumentation and chains of reasoning and, furthermore, if functioning rationally is taken to be the *modus operandi* of human mental activity, then one must find a way to explain why rational processes are failing at all. If we think rationally, why do we have irrational beliefs? From where do they come? Is reason flawed after all? What does that mean for DRT?

The focus over the next four chapters is to investigate common phenomena that tend to be classified as irrational—wishful thinking, self-deception, and akrasia—in the hopes of discovering whether or not there is a better way to understand irrationality itself.[71] What lessons can this teach about human psychology and behavior, if any? Does this confirm or disconfirm that DRT is the best way to analyze irrationality? Do these phenomena even exist or are they little more than literary fictions that help advance plots in novels and films? It is thus time to descend into the depths of human behavior in order to see if Enlightenment thinking can withstand the pressure.

---

[71] I will also be entertaining a new phenomenon in chapter seven: *negation*. This list, however, is by no means exhaustive. Other irrational phenomena might include bad reasoning, weakness of the warrant, compulsion, or conversion, to name a few. Unfortunately, due to the already substantial size of part two, I will not be analyzing other forms, focusing instead on the most popular as far as the literature is concerned.

Chapter 4: Kings and Queens of Wishful Thinking

One might spin a fantasy about an ordinary person riding a unicycle,
when suddenly the whole system expands a thousandfold.
Or one might describe a series of unicycles,
each bigger than the last.
In a sense, these are all appeals to intuition,
and an opponent who wishes to deny the possibility can in each case assert
that our intuitions have misled us,
but the very obviousness of what we are describing works in our favor,
and helps shift the burden of proof further onto the other side.

— David Chalmers, *The Conscious Mind*

4.1 | Nevertheless, Beliefs Persisted

It was 1993. A musician by the name of Vernon Howell had become increasingly convinced that the Book of Revelation contained cryptic, prophetic messages that detailed the end of days, and we were living in those times. What is more, he did not take himself to be just anyone; rather, he believed he had a mission as a modern-day prophet to bring about the second coming of Jesus Christ. This could be accomplished, so he thought, by unlocking the seven seals described in Revelation, using the book as one would a roadmap or blueprint.

On March 2nd, just a few days into the Federal Bureau of Investigation's siege of the Mount Carmel Center in Waco, Texas, Howell—better known at this point as David Koresh—had come to an agreement with an FBI negotiator to surrender on condition that an audio tape of his teaching be broadcast to a national audience.[72] The terms had been settled, and at 1:30 PM, the Christian Broadcasting Network aired the recording.

---

[72] For a breakdown of the timeline and what happened each day, see "Evaluation of the Handling of the Branch Davidian Stand-off in Waco, Texas February 28 to April 19, 1993" (Dennis Jr. 1993). This is a document released by the United States Department of Justice. URL = https://www.justice.gov/archives/publications/waco/evaluation-handling-branch-davidian-stand-waco-texas-february-28-april-19-1993.

According to one of his followers, David Thibodeau, Koresh likely consented to these terms because he "hoped that, on national radio, the churches and their congregations would have a chance to hear his word and recognize his teachings" (Thibodeau and Whiteson 2018, 191). But alas, "the results were disappointing," leading some to question not the message itself but how it could have been presented to the world *better* in such a short amount of airtime (193). Later that evening, another follower, Steve Schneider, reported to the FBI that Koresh had recently been instructed by God to wait, effectively reneging on the agreement to surrender. This not only came as a disappointment to the negotiators but as a surprise to everyone inside the center as well, as they struggled to understand what had gone wrong and whether they, not Koresh, had done anything themselves to cause this change in plan (196).

There is no question that Koresh and his followers (to put it lightly) miscalculated the influence he could command with a wider audience. In spite of the fact that he would essentially be a stranger to this new audience, Koresh found himself under the spell that Christians everywhere would immediately be enthralled after listening to a one hour recording of his lecture, freely and enthusiastically joining his cause. When the broadcasts failed to garner more attention and the realization started to settle in that reality was not quite aligned with their expectations,[73] he and his followers would find blame with neither his fantasy of instantaneous national appeal nor his supposed insight into the Armageddon; instead, they rigorously examined just about everything else in search for answers, including the possibility that this was some kind of divine punishment for celebrating in excess the night before (Thibodeau and Whiteson 2018, 196). Why

[73] His recording had aired several times that day.

were the most obvious culprits, those fantasies, not under more scrutiny? Why did everyone continue reinterpreting their situation in such a way so as to save the validity of those dreams? Why, above all, persist in holding on to beliefs that could not possibly be true?

It may be tempting to conclude that this is an exceptional scenario, that these people belonged to a cult and that their leader was suffering from a mental disorder. Regardless of whether that is true or false, however, this phenomenon is not nearly as uncommon as one might suspect. Neither is it unusual to find it embedded within other kinds of irrationality. In this case, the phenomenon I am discussing is *wishful thinking*. It unveils itself in a perverse display of courage, empowering one to press forward armed with a belief that is plainly outmatched against reality.

Before proceeding, it is worth mentioning that there is a tendency in common parlance to conflate the idea of wishful thinking with self-deception. In a discussion on self-deception, for example, Simler and Hanson write:

> But here's the puzzle: we don't just deceive *others*; we also deceive *ourselves*. Our minds habitually distort or ignore critical information in ways that seem, on the face of it, counterproductive. Our mental processes act in bad faith, perverting or degrading our picture of the world. In common speech, we might say that someone is engaged in "wishful thinking" or is "burying her head in the sand"—or, to use a more colorful phrase, that she's "drinking her own Kool-Aid." (Simler and Hanson 2018, 73–4; emphasis original)

Interestingly, amongst philosophers there is relative agreement that the two phenomena are distinct from one another; there is just plenty of disagreement as to *how*. Philosopher Ian Deweese-Boyd summarizes the tension in his article, "Self-Deception," briefly surveying a handful of different views before concluding, "While the precise relationship between wishful thinking and self-deception is clearly a matter of debate, there are

plausible ways of distinguishing the two that do not invoke the intention to deceive" (Deweese-Boyd 2017, 11–2). The traditional dividing line has thus been the *intent to deceive*, and in the case of wishful thinking, the assumption is that a person unwittingly deceives herself. Given the vigorous debate, it is clear that the two psychological phenomena are in desperate need of sorting.

So what *is* wishful thinking, and how might it differ from self-deception? This chapter will address the first of these questions, but the second will need to be postponed until more is said about self-deception itself.[74]

4.2 | Anybody Can Have Bouts of Madness

In *The Enigma of Reason*, Hugo Mercier and Dan Sperber tell a different tale, one that features a person who might be considered to occupy the opposite end of the psychological spectrum from someone like Koresh. As they preface this story, they remind their readers that this person, Linus Pauling, had won two Nobel Prizes in his lifetime before narrowly missing a third to James Watson, Francis Crick, and Maurice Wilkins for their discovery of the structure of DNA (Mercier and Sperber 2017, 205–6).

What happened? Possessing a brilliant, scientific mind, Pauling happened upon a study conducted by surgeon Ewan Cameron that seemed to indicate that vitamin C could be beneficial in fighting cancer. He became so enthusiastic about the results that he appealed to the Mayo Clinic to try to replicate this study with a larger, controlled trial. The Mayo Clinic agreed, but the results were a failure. Mercier and Sperber relay the decision that Pauling then faced:

> At this point, Pauling could have objectively reviewed the available evidence: on the one hand a fringe theory and a small, poorly controlled study; on the other

[74] See chapter five.

> hand, the medical consensus and a large, well-controlled clinical trial. On the basis of this evidence, the vast majority of researchers and doctors concluded that vitamin C had no proven effects on cancer. But Pauling did not reason objectively. (Mercier and Sperber 2017, 206)

Surprisingly, in spite of his intelligence, Pauling (like Koresh and his followers) continued reinterpreting reality in order to protect and persevere in his belief that vitamin C was nature's answer to the disease of cancer.

First, Pauling dismissed the findings from the Mayo Clinic on the grounds that the selection process was inadequate. In response to his concerns, a second study was performed, but it only confirmed the Mayo Clinic's conclusions. Again, however, he dismissed the results, this time arguing that the Vitamin C had not been administered long enough (Mercier and Sperber 2017, 206). This habit of framing the evidence to fit his narrative continued to no end. No matter what studies would show, nothing at all could shake him from his faith. Surely, he suspected, a method was flawed, a researcher unwittingly erred, the results were biased, the data was compromised, *something was wrong*; but whatever it was, it was not the truth of his belief. Amazingly, not even his own eventual diagnosis of cancer, which he claimed "would have struck earlier" had he not been taking high doses of the vitamin himself, could make him believe otherwise (207).

Although Mercier and Sperber use this as a cautionary tale about the dangers of reasoning and myside bias,[75] it also serves the purpose of illustrating that just about anyone can fall into the trappings of wishful thinking. As it turns out, both Koresh and Pauling formed unrealistic beliefs and protected them, not only in the absence of

[75] Myside bias is the psychological phenomenon that occurs when "reasoning systematically works to find reasons for our ideas and against ideas we oppose" (Mercier and Sperber 2017, 218).

supporting evidence but even *in spite of* evidence to the contrary. Fortunately for Pauling, however misguided his belief was, it did not cost him his life. But nonetheless, it does not make it any more reasonable. The problem that the phenomenon of wishful thinking creates cannot be stated any clearer than this:

> Our beliefs are supposed to inform us about the world in order to guide our actions. When the world is not how we would want it to be, we had better be aware of the discrepancy so as to be able to do something about it. Thinking that things are the way one wishes they were just because one so wishes goes against the main function of belief. (Mercier and Sperber 2017, 245)

Wishful thinking thus inclines us to look away from reality, often to our own detriment. How does this happen? Why do we do this? What is doing this inclining? And above all, why does it overpower a commonsense concern for reality?

4.3 | The Illusion of Objectivity

Cognitive psychologists will often distinguish between *motivational* and *cognitive* processes. The former refers to any non-cognitive mechanisms, states, or systems that can affect behavior, decisions, in short, the outcomes of thinking. Though interpretations can vary, Freud's theory of drives or even Plato's conception of an appetite-driven part of the soul would by today's standards be classified as belonging to this motivational category. As for cognitive processes, these encompass all of the mechanisms and systems involved in the processing of information, including perception, causal reasoning, facial recognition, etc. Throughout the 20th century, there have been advocates who have tried reducing one set of processes to the other, but more recently, interest has trended towards understanding how these two might *interact* and what this could mean for our understanding of thinking in general (Molden and Higgins 2005, 296).

When it comes to the topic of wishful thinking, the tradition has been to assume that it is a kind of *motivated* belief, making it a psychological phenomenon whose manifestation is driven primarily by a motivational process, typically a desire or wish. Davidson, for example, writes that wishful thinking occurs when "a desire or wish that a proposition be true causes a person to believe that it is true, but is not a reason for thinking it true" (Davidson 1997, 220). So if, for instance, Katerina begins to believe that she will win the lottery on account of a desire or wish that she will win, and, crucially, if she has no evidence to support this belief, then Davidson would take her to be engaged in a form of wishful thinking. While there is little doubt that an unreasonable belief or set of beliefs lies at the heart of wishful thinking, there is no clear way to answer the questions this phenomenon raises without better understanding some of the motivational processes that affect how we think in the first place. No analysis of cognition in isolation of motivational processes will explain how these beliefs manage to evade the demands of reality.

The evidence from the research into motivated thinking has shown that people tend to "access hypotheses, inference rules, and instances from past behavior that are most likely to support their desired conclusion" (Ditto et al. 1998, 54). But even more fascinating is that people also do the opposite to undermine the validity of evidence that does *not* support their conclusion, selectively seeking out any available inferences or prior examples that serve that end (54). While this already foreshadows the *kind* of thinking that occurs during wishful thinking, there is nonetheless a mystery. Social psychologist Ziva Kunda and others have argued that in spite of our tendency to bias the information we interpret, there are still constraints on these processes. The reason why

people do not believe whatever they want is that they have a desire to maintain an "illusion of objectivity", reaching conclusions "only if they can construct a justification that they believe could 'persuade a dispassionate observer'."[76] So, in addition to any practical advantages to be had when we allow reality to inform our beliefs, we also have reputations to uphold. But for some reason, in cases of wishful thinking, this concern for reality along with a desire to maintain an illusion of objectivity seems to be breaking down. Why believe the increasingly absurd? Surely Koresh's and Pauling's beliefs had become indefensible to any imagined dispassionate observers. Before trying to understand this, let us first develop a clearer picture of some of these motivational states and how they affect us.

4.4 | Directing the Cognitive Choir

Before the turn of the millennium, social psychologist Peter Ditto and colleagues set out to determine how a person's preferences might influence how she evaluates and accepts information. Do we play favorites when we review evidence and information? Do we treat "good" information, the kind that supports what we already believe, differently from "bad" information? If so, in what ways? Why might we do this? Ditto's experiments explored exactly these questions.

In one experiment in particular, participants received a medical test to check for the presence of an enzyme, *thioamine acetylase* (TAA). The test could not be any easier, requiring little more than a sample of saliva to be checked against a reactive strip that would change color in the presence of TAA. What the participants did not realize is that

[76] cf. Ditto et al. 1998, 54. See also: "The case for motivated reasoning" (Kunda 1990); and "Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model" (Pyszczynski and Greenberg 1987).

the test strip was actually just a glucose-sensitive strip, and the mouthwash they had been advised to use prior to collecting their saliva contained exactly that, glucose. Furthermore, though every single strip would test positive, TAA itself was nothing more than a fabrication designed to misdirect the attention of the participants (Ditto et al. 1998, 62–3).

Now before taking the test, the research subjects had been divided into two groups. Half of them had been informed that the body's production of TAA was a *good* thing. Its presence, they explained, significantly *decreases* one's chances of developing pancreatic disease compared to the rest of the population who lacks the enzyme. The other group had been lead to believe precisely the opposite: if TAA was detected, they were informed that they had a *much higher* chance of developing pancreatic disease (Ditto et al. 1998, 63).

When the subjects learned of their results, the researchers carefully offered alternative explanations for the positive tests, revealing that previous studies have shown that there is a possibility that blood sugar levels could affect the readings. A second twist was introduced at this point. Though this possibility was shared with everyone, some (in both groups) were told that this alternative was *reasonably likely* (i.e. statistically affecting 1 in 10 reactions) while the rest were advised that it was a *very rare* occurrence (i.e. statistically affecting 1 in 200 reactions). For the final leg of the experiment, the participants were asked to fill out a questionnaire, detailing their confidence in the accuracy of the results and how likely they believed future tests would yield the same outcome (Ditto et al. 1998, 63).

After collating the data, the researchers discovered that when participants believed that TAA was indicative of a *healthy* outcome, they tended to believe the tests were very accurate, *regardless of the probability of the alternative explanation*. When, on the other hand, participants believed that the presence of TAA was *unhealthy*, the researchers found an inverse correlation between their faith in the results and the probability of the alternative explanation. When the alternative was presented as *reasonably likely*, the participants regarded the test as inaccurate, but when it was presented as *very rare*, they instead believed the results were accurate (Ditto et al. 1998, 63–4). In other words, the people in this study tended to indiscriminately embrace positive evaluations of their health, regardless of the likelihood of any alternative explanations, but they were more reluctant to acknowledge negative evaluations except when their likelihood was overwhelmingly plausible.

This same phenomenon has been documented in other studies as well, most notably one conducted by Kunda,[77] which showed how women who consumed high amounts of caffeine were more skeptical of a scientific article that linked caffeine to adverse health outcomes than women who consumed low amounts (Molden and Higgins 2005, 298). In each of these cases, what psychologists are observing is something known as *directional outcome motivation*, a motivational state that inclines someone to reach a particular conclusion, usually in the form of a positive self-impression or a reinforcing belief about someone or something they already like (296–7). Such a state influences *how* we look at the information available to us. There is certainly no shortage of examples of this in the political sphere. If a person's favorite political candidate does something

[77] See "Motivated inference: Self-serving generalization and evaluation of causal theories" (Kunda 1987).

perceived as positive, regardless of any other explanations to the contrary, she will tend to celebrate the supposed achievement; and when the opposite occurs, she will be quick to entertain any alternatives that explain away the supposed flaw.

4.5 | An Orientation Towards the Top

For a theorist who places rationality at the center of human mental life, it might be frustrating to see reason so easily hijacked by non-cognitive processes. How can this be explained? Such behavior is plainly irrational, even on the deviant rationality thesis. If good health, for example, is key to optimizing quality of life, what reasonable person would turn away from the news that a lifestyle choice is demonstrably and unequivocally *unhealthy*? Who would turn a blind eye to a negative test result, especially if it can be corrected through medical intervention? And yet we do.

Even from an evolutionary perspective, this seems peculiar. Ditto goes so far so as to call it an "adaptive paradox,"[78] asking, "If beliefs are biased by wishes and fears, how can people deal effectively with negative feedback and environmental threat" (Ditto et al. 1998, 65)? However much of a problem this poses for the Enlightenment's faith in reason's icy objectivity, it seems to make even less sense to consider that we may have evolved with motivational mechanisms whose jobs are to explain away negative information (when possible) and uncritically embrace positive information. Why would we do such a thing?

---

[78] Ditto does have his own response. Though we put a lot of cognitive effort into explaining away negative information, he zeroes in on *how much* effort we exercise in examining negative information as opposed to positive information. His answer to the paradox, therefore, is that we are designed to put more cognitive effort into analyzing potential environmental threats (Ditto et al. 1998, 13–4). The paradox thus takes a new form: why are we cognitively lazy with positive information? To this, he speculates that "being an undiscriminating consumer of desirable feedback" may prove to be nothing more than an efficient use of cognitive resources (14).

There is one evolutionary perspective from which this makes perfect sense: reputation management. Consider for a moment the contexts in which directional outcome motivation most forcefully presents itself. Kevin Simler and Robin Hanson summarize a battery of these:

> Unfortunately, study after study shows that we often distort or ignore critical information about our own health in order to seem healthier than we really are.[79] One study, for example, gave patients a cholesterol test, then followed up to see what they remembered months later. Patients with the worst test results—who were judged the most at-risk of cholesterol-related health problems—were most likely to misremember their test results, and they remembered their results as better (i.e., healthier) than they actually were. Smokers, but not nonsmokers, choose not to hear about the dangerous effects of smoking. People systematically underestimate their risk of contracting HIV (human immunodeficiency virus), and avoid taking HIV tests. We also deceive ourselves about our driving skills, social skills, leadership skills, and athletic ability. (Simler and Hanson 2018, 74–5)

They identify these behaviors as forms of self-deception before emphasizing an important takeaway: "Self-deception is a tactic that's useful only to social creatures in social situations" (79). In each of these cases, it is clear that one's reputation within a community is on the line. Whether or not it is done intentionally, convincingly overestimating your athletic ability, intelligence, or leadership skills while persuasively discrediting, undervaluing, or even mocking the abilities, intelligence, and skills of others are behaviors that can enhance how a community perceives you. While the reputational advantages to be had in most of these contexts is clear, perhaps the least obvious might be how evaluating one's health has *anything* to do with reputation, but on closer inspection, even this proves consistent with the reputation management thesis.

---

[79] For more information, the studies that Simler and Hanson reference are: "How well do people recall risk factor test results? Accuracy and bias among cholesterol screening participants" (Croyle et al. 2006); "Behavioral receptivity to dissonant information" (Brock and Balloun 1967); "Perceiving AIDS-related risk: Accuracy as a function of differences in actual risk" (van der Velde, van der Pligt, and Hooykaas 1994); "'Don't tell me, I don't want to know': Understanding people's reluctance to obtain medical diagnostic information" (Dawson, Savitsky, and Dunning 2006); and "The better than average effect" (Alicke and Govorun 2005).

For one, human beings, like so much of the rest of the animal kingdom, are *sexual* creatures. We look for mates, seek attention, signal with subtle and not-so-subtle behavioral cues, and engage in a host of courting rituals. We will don the most outlandish of attire if there is the slightest chance it will separate us from our competitors, scent ourselves with the most aromatic of perfumes, flaunt our socioeconomic status with luxury goods, and more. In so many of these endeavors, there is a layer of artificiality we embrace. Of course, for many of us, our hair is not naturally straight, our scent naturally appealing, our skin naturally radiant, our financial situation in good standing. We have dry, stiff hair, body odor, wrinkles, and debt. As Simler and Hanson emphasize, however, "From the perspective of evolution, mating, not survival, is the name of the game," and so we do our best to exploit every artificial advantage that comes our way in this game (Simler and Hanson 2018, 31). In the mating game, it makes sense to desire that one's potential mate be intelligent rather than unintelligent, attractive rather than unattractive, wealthy rather than poor, and even *healthy rather than unhealthy*. Why not play the part for potential partners?

But even beyond considerations of mating, one's social *status* is equally important for *social* creatures. While the foraging communities of our ancestors tended to be egalitarian, they still found themselves in a constant state of war with the individual impulse to rise above the pack. There are, after all, benefits to be had for climbing the ladder and securing more respect and influence. "The higher your status," Simler and Hanson write, "the more other people will defer to you and the better they'll tend to treat you" (Simler and Hanson 2018, 32). One might wonder why a community would elevate an individual in the first place, but it is important to keep in mind that a community, too,

is composed of other *individuals* with the same impulses. They, too, want to get ahead and want to satisfy desires, and when you convince them that you have something to offer, they will be more willing to engage with you and recognize your value. Robert Kurzban explains this interaction in terms of investments and exchanges, writing, "Most human relationships involve, at least to some extent, the exchange of various goods and services over time" (Kurzban 2010, 90). If people do not believe you are *able* to return favors or exchange goods and services, they will not be willing to invest time and energy into you (93). In other words, giving anyone reason to believe that you are unable to be of potential benefit to them can quite easily jeopardize your social standing, and this includes giving any signals that you are in poor health.

During the 2016 United States presidential election campaign, for example, efforts were made from unsavory critics on both sides to undermine the health of the candidates. Rumors slowly bubbled to the surface from the bowels of the internet that Hillary Clinton was allegedly privately grappling with Parkinson's Disease while Donald Trump was hiding and fighting Alzheimer's Disease. The implicit message for each was the same: if you are not healthy, you are not fit to occupy the highest and most important status in our society (ultimately because you will not be able to give us what we want). What we have thus evolved to do is filter any incoming information with an eye to protecting, defending, and promoting our reputations.

4.6 | An Affinity for Somewhere Around Fifty Shades of Grey

But an overarching desire to bolster one's reputation cannot be the whole story when it comes to distorting the evidence and forming our beliefs. What good is it to convince the community that you are the strongest warrior only to decisively fail against a supposedly

weak enemy? Who wants to be remembered as the embarrassment that was Goliath? Or, as Kurzban puts it, "It does no good to act as though you're a great lion tamer if you're going to be thrown into the ring with a lion" (Kurzban 2010, 92). At some point, our beliefs will come face-to-face with reality, and better that we submit them sooner rather than later. Like the participants in Peter Ditto's experiment (described in section 4.4 above), the more plausible the counterevidence appears, the more likely we are to concede and forfeit those beliefs that run contrary to it. Yet, this is precisely what does *not* happen during wishful thinking. Why do these beliefs sometimes persist?

This is actually a topic explored by social psychologists David Dunning, Judith Meyerowitz, and Amy Holzberg. What especially piqued their interest is something known as the *above average effect*, which occurs whenever we evaluate ourselves as being better than our peers. The paradox of the above average effect lies in the fact that in many of these cases, *nearly everyone* evaluates themselves as better than their peers, leaving one to question what it even means to be above average (Dunning, Meyerowitz, and Holzberg 1989, 1082). One study found, for instance, that 94% of college professors believe they do above average work.[80] But this is not restricted to just performance evaluations; this has also been observed in studies that have covered everything from driving ability and ethics to personal health and managerial skills.[81] It is not even uncommon for such self-assessments to "appear to be favorable to a logically impossible

[80] Dunning, Meyerowitz, and Holzberg reference the following for this figure: "Not can but will college teaching be improved" (Cross 1977).

[81] For more information, they reference the following: "Are we all less risky and more skillful than our fellow driver?" (Svenson 1981); *An Honest Profit* (Baumhart 1968); "Swine flu: A field study of self-serving biases" (Larwood 1978); "Unrealistic optimism about future life events" (Weinstein 1980); "Unrealistic optimism about susceptibility to health problems" (Weinstein 1982); and "Managerial myopia: Self-serving biases in organizational planning" (Larwood and Whittaker 1977).

degree" (1082). What could be enabling these inflated self-evaluations that seem to ignore reality? According to their research, one of the most important driving factors is *ambiguity*.

Over a series of four experiments, what Dunning and his colleagues discovered is that when we use ideas that have no clear meaning (e.g. "leadership," "excellence," "sophisticated," etc.), we tend to idiosyncratically interpret them to cast ourselves in the best possible light (Dunning, Meyerowitz, and Holzberg 1989, 1082). When meanings are more clearly fixed, however, we tend to view ourselves in more realistic ways. Hugo Mercier and Dan Sperber explain this phenomenon well:

> For instance, people tend to think they are more intelligent than the average—that's an easy enough belief to defend: they can be good at math, or streetwise, or cultured, or socially skilled, and so on. By contrast, there aren't two ways of being, say, punctual. Since people can't think of ways to believe they are more punctual than average, they just give up on this belief, or on other beliefs similarly hard to justify. (Mercier and Sperber 2017, 245)

Under what better conditions can a belief persist in the face of reality than when a person perceives the countervailing evidence to be open to interpretation? Might it be the case that the more ambiguous the evidence appears to be, the more likely we are to distort it to our advantage? Interestingly enough, there is an analogous automated phenomenon that occurs during perception, and it bears a striking similarity to how we interpret information and form beliefs. Though a similar process affects hearing as well, it is well-studied in vision.

4.7 | That's One Way to Look at an Intellectual Blind Spot

If we are to see anything whatsoever, the visual information that is received from the eye needs to be transmitted to the brain in order to be processed. The informational highway

that makes this possible is the *optic nerve*, comprised of a bundle of neurons and other cells. But there are no photoreceptors where these cluster together and exit the back of the eye, and where there are no photoreceptors, there is no way to receive visual information from the surrounding environment. Where these neurons meet is known as *the blind spot*. While having two eyes helps filter out the missing information, the blind spot can still affect our field of vision whenever we look at the world with just one eye. And yet, under normal circumstances, when we do this we still do not *see* the blind spot; it looks as if we still have a uniform field of vision (Koch 2004, 53–4). There are no holes in our surroundings, no black spots, no error messages that obstruct the metaphorical visual screen. What is happening here? As it turns out, our sensory systems can deploy several tricks to maintain the illusion of the uniformity of sense experience by interpreting ambiguous sensory information.

When it comes to vision, discovering and studying how we perceive optical illusions have allowed neuroscientists to identify a special process known as *filling-in*, which is the brain's way of making educated guesses based on surrounding context (Koch 2004, 23, n.5). Given the visual information from what you *can* see, an automatic inference is made regarding what you *should* see, constructing the rest of the experience for you. In one study conducted by neuroscientist Vilaynur Ramachandran, a yellow annulus (a disc with no center) was placed precisely over participants' blind spots such that the void of the annulus lined up perfectly with it.[82] The result? Participants reported

[82] See: "Blind Spots" (Ramachandran 1992); and "Perceptual filling in of artificially induced scotomas in human vision" (Ramachandran and Gregory 1991).

seeing a *complete* yellow disc even though when they looked off to the side they could perceive the hole (54).

Another variation of filling-in is experienced when we look at Edward Adelson's classic checkerboard illusion. Atop a checkerboard sits a large cylinder that casts a shadow over a third of the board. We are then invited to contrast the color of two squares, one outside of the cylinder's shadow and one inside. Anyone with a healthy functioning visual system will concede that the two squares appear to be different colors, and yet the truth is that they are identical in color, a fact noticed whenever the illusion is covered up such that the squares can be viewed in isolation. Using the surrounding context and expectations about patterns, our visual system *assumes* that the squares must be different in color and fills in the information before presenting it to conscious experience.[83] No matter how much we gather evidence concerning these illusions and come to believe otherwise, our visual system simply does not care; it consistently interprets the sensory information as it sees fit.

In the case of Dunning, Meyerowitz, and Holzberg's experiments, they found that individuals interpreted ambiguous ideas in accordance with their own criteria and definitions, which in turn lead to inflated self-assessments (Dunning, Meyerowitz, and Holzberg 1989, 1088). This is just one variation of what has been demonstrated so frequently that it can be regarded as a psychological truism, namely, that we will draw upon the beliefs and attitudes available to us whenever we try to understand the world

[83] cf. *The Enigma of Reason* (Mercier and Sperber 2017, 30–2), and *Why everyone (else) is a hypocrite* (Kurzban 2010, 13–4).

around us.[84] Molden and Higgins summarize the findings from these studies, writing, "In typical circumstances, concepts or attitudes that have been recently or frequently activated will lead people to assimilate their judgments to this highly accessible information without considering any additional information" (Molden and Higgins 2005, 303). To put it another way, we will use what we know to make sense of what we do not know. Like our visual system, our cognitive system "fills in" ambiguous information with any surrounding cognitive context it already "sees" or has access to. As a result, the *way* that the information is interpreted often only serves to *strengthen* the pre-existing beliefs that were used to make sense of it; this is a phenomenon known as *attitude polarization* (Lord, Ross, and Lepper 1979, 2099). It is a circular cognitive strategy—and therein lies the problem—but is it *always* a problem?

Social psychologists Charles Lord, Lee Ross, and Mark Lepper studied attitude polarization in their own experimental research, finding that "completely inconsistent or even random data—when 'processed' in a suitably biased fashion—can maintain or even reinforce one's preconceptions" (Lord, Ross, and Lepper 1979, 2099). One way that they had tested this was by presenting participants with two fictitious studies on capital punishment, one that supported it and one that opposed it. An experimenter reviewed each study, explained the "results," and presented seemingly credible "criticisms." Lord and colleagues made sure that the studies, results, and criticisms were counterbalanced to minimize risk of influencing the participants' impressions.

[84] See: "Attitudes as object-evaluation associations: Determinants, consequences, and correlates of attitude accessibility" (Fazio 1995); and "Knowledge activation: Accessibility, applicability, and salience" (Higgins 1996).

The results showed that those who were *already* in favor of capital punishment (cp) dismissed the anti-cp study and *reinforced* their views with the pro-cp study; and those who were already *against* capital punishment did the opposite. Lord, Ross, and Lepper explain, "the same study can elicit entirely opposite evaluations from people who hold different initial beliefs about a complex social issue" (Lord, Ross, and Lepper 1979, 2100–2). Not only did participants *fail* to treat ambiguous information with uncertainty, they *reinforced* their pre-existing beliefs by idiosyncratically interpreting it in a way that fits those beliefs (2104).

Though this seems like a serious cognitive failure, Lord, Ross, and Lepper argue that this is actually a *good* thing. They explain that, in general, this "same bias leads most of us to be skeptical about reports of *miraculous virgin births* or *herbal cures for cancer*, and despite the risk that such theory-based and experience-based skepticism may render us unable to recognize a miraculous event when it occurs, overall we are surely well served by our bias" (Lord, Ross, and Lepper 1979, 2106; emphasis original). Furthermore, they stress that from an evolutionary consideration, it is not clear how else an organism can be expected to interpret its environment *without* relying on stored memories and experiences (2107). If every situation had to be mediated by a blank slate, we would be in a lot of trouble.

The issue with attitude polarization is not therefore *how* we exploit the stored information already available to us to filter, interpret, and color our experiences and situations; instead, the problem is when it accidentally *misfires* by increasing the strength of beliefs through circular justifications. This particular cognitive feature evolved to help us better understand our *surroundings*, not better understand scientific research or

religious beliefs. The latter are the metaphorical equivalents of placing yellow annuli in front of our intellectual blind spots.

4.8 | A Scaffold of Reasons

The psychological pieces are now in place to make sense of wishful thinking. Out of an overarching desire to improve our reputations, we form wishes and construct fantasies that entertain those possibilities that can make that happen. For some of us, these wishes never rise to the level of wishful thinking, for we possess a keen awareness of how unreasonable they are. We might daydream about packing our bags and taking a week at a lakeside resort without the slightest of inclinations to actually do it, whether we are financially able or not. We might imagine engaging in gladiatorial combat with our obnoxious co-worker, cheered on by thousands of by-standers who celebrate our impressive display of defensive finesse and martial skill as we dance to victory, again entertaining just a fantasy without any urge to put into action whatsoever. In general, wishes do not necessitate that we act on them. As Davidson observes, however, "If someone wishes that a certain proposition were true, it is natural to assume that he or she would enjoy believing it true more than not believing it true" (Davidson 1986, 205). After all, were it true, it would (so we dream) enhance our reputations. What, then, causes us to mobilize otherwise idle wishes into beliefs? *Attitude polarization*.

Our pre-existing beliefs empower us to recognize wishes as more viable than they are, and in turn, this conceptual framework interprets and distorts any ambiguous evidence in its favor, both fortifying and strengthening the wish until it spills over into belief. What would be little more than an idle wish to one person can become another person's very real possibility to enhance her reputation. Koresh, for example, believed

that his lecture would instantly appeal to a wider Christian audience *because* he believed he was ordained by God to fulfill a mission. While Pauling's motives are a little less clear, we can speculate that his belief in the brilliance of his own scientific mind fueled the fantasy that he had discovered a cure for cancer. Did neither have these kinds of auxiliary beliefs in advance of forming their wishes, they likely remain dutiful subjects who remain in the realm of harmless, idle fantasies. But how did neither yield to the force of unambiguous reality? How could they not see the unreasonableness of their fantasies when the counterevidence pushed back against them? The answer is that, for them, the evidence afforded by reality was *not* unambiguous.

The conceptual framework that can fuel wishful thinking is also the one that we use to understand reality, and, depending on the beliefs that are activated to make sense of the information, what might be clear and decisive for one person can appear unclear and uncertain to another. Our pre-existing beliefs can thus not only bias ambiguous evidence but also *ambiguate* otherwise clear evidence to the contrary! Every conspiracy theorist, for instance, argues that the official accounts of events are inadequate, that they can find evidence that has been missed or ignored, and that there are alternative explanations for the events that occurred. They see ambiguity where the rest of us see none. With the beliefs they have available, they build and use what might best be described as "an unyielding scaffold of reasons" (Mercier and Sperber 2017, 242).

4.9 | Large We Are; Multitudes Contain Us

In section 4.3, it was mentioned that people have a desire to maintain an "illusion of objectivity" by finding justifications for their beliefs that will appease a presumed audience of dispassionate observers. The question was raised why there seems to be a

breaking down of this illusion in cases of wishful thinking. How could someone like Koresh or Pauling sincerely believe that their defenses for their fantasies could be considered remotely plausible by the larger community? Why risk considerable damage to their reputations with obviously indefensible justifications? Even if their beliefs distorted their interpretation of the evidence, surely they were sensitive to the fact that we did not share in those beliefs. Well, as it turns out, *we* were not their audience; *we* were not part of the communities to whom they were appealing. The illusion of objectivity never broke down.

In *The Emotion Machine*, artificial intelligence researcher Marvin Minsky proposes a *multiple models* theory of self. The idea is that nobody has a single sense of self; rather, we construct different self-models that we adapt to our situations and circumstances. The decisions that we make, he argues, depend on which self-model we activate, using a fictional persona (Joan) to convey his point:

> When Joan is with a group of her friends, she regards herself to be a sociable person. But when surrounded by strangers, she sees herself as anxious, reclusive, and insecure. […] Joan's mind abounds with varied Self-models—Joans past, Joans present, and future Joans; some represent remnants of previous Joans, while others describe what she hopes to become; there are sexual Joans and social Joans, athletic and mathematical Joans, musical and political Joans, and various kinds of professional Joans. […] Joan's *Business Self* might be inclined to choose the option that seems more profitable; her *Ethical Self* might select an option that better conforms with her ideals; her *Social Self* might want to select the one that would most please her friends. (Minsky 2006, 306)

The idea of multiple self-models is attractive. It helps explain why the role we play on Saturday night is very different from the one we play Sunday morning. Who we portray ourselves to be in front of family is not the same as who we portray ourselves to be in the office, and both are different from who we are when surrounded by the most trusted of friends.

According to Minsky, the reason that we need multiple self-models instead of a single one is that being human is complicated. We ourselves are ambiguous beings, lazy in some ways and productive in others, confident in some areas and insecure in others. As such, Minsky believes that it would be extraordinarily counterproductive if we represented our entire complexity to ourselves in a single self-model. As he puts it, "you would be overwhelmed by seeing so many unwanted details" (Minsky 2006, 321). It is thus easier to divide our self-conceptions into smaller, more manageable models; and the upshot is that we can shift between models to suit the circumstances in which we find ourselves. We may want to deploy the *sexual* self-model when we enter the mating game while finding it better to suppress it around our children, among whom we instead deploy the *parent* self-model. "Each model," he writes, "must help us to focus on only those aspects that matter in some particular context; that's what makes a map more useful to us than seeing the entire landscape that it depicts" (321).

While the theory is attractive, the explanation for *why* we construct multiple self-models is left wanting. Why would we need a self-model to interact with the environment, let alone *multiple*? We would, however, need multiple self-models if we needed to manage multiple *reputations*. If we are members of many communities, each of which with its own standards for membership, there then arises a clear need to construct, protect, and project reputations for each of these communities. As Minsky himself observes, Charles might find Joan to be cooperative to the point that she undervalues herself at work while noticing that in social settings she comes across as "selfish and overrating" (Minsky 2006, 301–2). This is exactly the shift in behavior and image one would expect if qualities like cooperation and humility are useful for boosting one's

reputation in the business *community* but *not* one's particular social circle, where such qualities might encourage others to take advantage of you. The justification for our actions, then, is not aimed at some community at large, but instead a particular community with whom our reputation at the time is our priority. Explaining ourselves to our families will entail very different standards of justification from the justifications required in the boardroom.

Koresh and Pauling, therefore, were not ignoring the illusion of objectivity. If anything, they adhered to it all too well. For his devotees at Mount Carmel, with whom his reputation was on the line, it makes perfect sense to use as a defense for the failure of his broadcast, "God instructed me to wait." The idea of a god who could communicate directly with their teacher was precisely a belief that was central to Koresh's community. Pauling, who needed to navigate the tumultuous waters of a scientific community with high standards of evidence, found himself constructing elaborate criticisms of the studies that cast doubt on his fantasy that he had discovered the cure for cancer. It is even possible that both men were trying to find ways to defend *multiple* reputations at the same time. Perhaps Pauling was also appealing to a community comprised of family members, or maybe even an herbal medicine community?[85]

While membership in a community can often bring a host of benefits, it also comes with its fair share of danger, especially when *groupthink*—or, "the failure of group members to criticize each other's suggestions and to consider alternatives"—prevails (Mercier and Sperber 2017, 243). Hugo Mercier and Dan Sperber explain how this

[85] He did, after all, establish the Linus Pauling Institute in order to investigate not just the effects of Vitamin C, but also other vitamins, minerals, and plant compounds. See: http://lpi.oregonstate.edu/about/about-linus-pauling-institute.

phenomenon has been studied in experimental settings and is well-established in the psychological community:[86]

> Put a bunch of people together and ask them to talk about something they agree on, and some will come out with stronger beliefs. Racists become more racist, egalitarians more egalitarian. Hawks increase their support for the military; Doves decrease it. When you agree with someone, you don't scrutinize her arguments very carefully—after all, you already accept her conclusion, so why bother? (243–4)

The communities to whom we belong can create echo chambers and feedback loops, reinforcing the worst of our beliefs, enticing us to mobilize fantasies into action, and crippling our abilities to optimally interpret the world around us.

This is not to suggest that wishful thinking is inevitable—though it might be. There are, however, at least a couple of safeguards that can discourage its development. A community needs to become more self-aware of our propensity to distort evidence so as to help cultivate the virtue of intellectual humility. Like the participants in experiments with optical illusions, if we can become *aware* of our intellectual blind spots, we give ourselves the best chance to mistrust our flawed perceptions.[87] Likewise, membership in *multiple* communities is beneficial, providing opportunities for beliefs to be challenged by criteria that lie beyond that of a single community, which can be an essential check

---

[86] See: "Discussion effects on racial attitudes" (Myers and Bishop 1970); and "Discussion effects on militarism-pacifism: A test of the group polarization hypothesis" (Myers and Bach 1974).

[87] I am not confident that an individual can ever overcome her intellectual blind spots; holding out hope might itself be wishful thinking. In a study cited by Kurzban, undergraduates were informed of the above average effect, and when asked if they believed it affected them, they "uniformly judged themselves less susceptible than the average American," that is, they believed themselves above average at being susceptible to the above average effect (Kurzban 2010, 105). For more on this study, see: "Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others" (Pronin, Gilovich, and Ross 2004, 781–799).

Still, at a minimum, I believe that if the *community* can become aware of the patterns and pitfalls of *individual* thinking, it can help eliminate them at least at the level of the community, much like how seeing with *both* eyes eliminates the blind spot of each individual eye.

against conceptual frameworks that prove counterproductive or even harmful to oneself and others.

On the other hand, if it turns out that wishful thinking *is* inevitable, it might be to our benefit as much as our folly. As philosopher Nicolas Rescher argues, it is wishful thinking that inspires us to push beyond what is possible and aspire towards greatness, to "work toward the realization of a condition of things that do not and perhaps cannot exist," to open up the "sphere of ideals" (Rescher 2009, 6–7). It was wishful thinking, after all, that initially helped drive the "American experiment" in 1776 and the moon landing in 1959. The line between the possible and impossible, the reasonable and unreasonable, can sometimes be much thinner than we realize. Davidson, too, was aware of this, remarking, "When wishful thinking succeeds, one might say, there is no moment at which the thinker must be irrational" (Davidson 1986, 206). Perhaps, then, it is not wishful thinking *as such* that should concern us, but the mitigation of the more harmful forms it can take.

4.10 | Conclusion

Wishful thinking emerges whenever a wish that stands in opposition to reality is converted into a belief and then subsequently protected by a conceptual framework that distorts any evidence to the contrary in favor of preserving and supporting the wish. Wishes, in general, are relatively harmless, involving little more than reputation-enhancing fantasies (i.e. daydreams) but when they are taken up into wishful thinking, they can prove detrimental (though not necessarily). The line between madness and genius is thin, for if Pauling *had* stumbled upon the cure for cancer or Koresh *had*

precipitated the Second Coming, neither perhaps ought to have been counted as irrational on those scores.

The Deviant Rationality Thesis—that irrational behavior is simply a deviation from rational norms—cannot adequately describe what makes wishful thinking irrational *when* it is irrational. For one, what enables wishful thinking to arise are exactly those justification standards, the rational norms, of a particular community whose beliefs lend themselves to treating the fantasy as real in the first place. In addition, there are non-rational motivational processes that also help make wishful thinking possible, processes that undermine the Enlightenment assumption of human "rationality." Social psychologists have verified, for instance, how directional outcome motivations have profound influence on the reasoning process as well as how we tend to idiosyncratically interpret ambiguous information to our advantage. Because these non-cognitive processes affect how we tend to think, consistently creating cognitive failures in the same contexts, does DRT include the effects of these processes amongst the rest of the rational norms or not? This is, after all, how we *tend* to think—a *de facto* rational norm. Lastly, it is not clear how a DRT theorist can recognize wishful thinking as *anything but* irrational, considering how it stands in opposition to reality. And yet, it appears to be the case that wishful thinking ought not be regarded as *necessarily* irrational, especially when it proves to be an empowering cognitive illusion that can facilitate turning our dreams into reality. On the contrary, what makes it irrational is not the *act* of wishful thinking itself but when it begins to undermine the quality of life for an individual, becoming an *obstacle* to one's goals rather than an aid.

Wishful thinking itself takes root and thrives in environments filled with ambiguous information, and for human beings—social creatures who have multiple reputations to defend in multiple communities—ambiguity is everywhere. Not only do we distort ambiguous information, but we even use our pre-existing beliefs to ambiguate what might otherwise have been clear. Depending on the communities in which we find ourselves, we may even unwittingly create feedback loops that further enhance the power of these processes to a point where they no longer work in our favor, causing any constraints on our thought processes that could have been afforded by reality to instead become obscured and restructured to better fit with our fantasies. But all is not lost. There is plenty of evidence to suggest that we can and do concede when the counterevidence becomes overwhelming, and this is why intellectual humility and membership in multiple communities can help serve as checks against the more insidious forms of wishful thinking.

Chapter 5: To Deceive or Not to Deceive

In this book
I am using words like 'deceive' and 'lie' in a much more straight-forward sense than those philosophers.
They were interested in conscious intention to deceive.
I am talking simply about having an effect functionally equivalent to deception.

— Richard Dawkins, *The Selfish Gene*

5.1 | Remember Sammy Jankis

One of the most provocative depictions of self-deception in cinematic history can be found in an earlier film by director Christopher Nolan, *Memento*. The story follows a man named Leonard Shelby, who explains that he used to be a fraudulent claims investigator for an insurance company. As the audience learns, Leonard now suffers from anterograde amnesia, a condition that began after he suffered serious head trauma during a home invasion. Unable to form new memories, he is now driven by one purpose: to find his wife's murderer. During the film, he meets at a diner with a police officer named Teddy Gammell, and, drawing from his former experience as an investigator, he theorizes to Teddy that the function of memory is not to track truth:

> No, really. Memory's not perfect. It's not even that good. Ask the police, eyewitness testimony is unreliable. The cops don't catch a killer by sitting around remembering stuff. They collect facts, make notes, draw conclusions. Facts, not memories: that's how you investigate. I know, it's what I used to do. Memory can change the shape of a room or the color of a car. It's an interpretation, not a record. Memories can be changed or distorted and they're irrelevant if you have the facts.

The juxtaposition between memory and facts, the subjective and the objective, is noteworthy,[88] but more important are the insinuations made about memory itself.

[88] According to an interview with Nolan himself, the film's very structure and presentation plays with the boundaries between the objective and the subjective, between what is happening independently of Leonard's experience and what is filtered through it. For more on this, see his discussion with Eyes on Cinema: https://www.youtube.com/watch?v=qduOF_sl1IQ.

Mirroring the subjective / objective distinction, Leonard impresses upon the audience that memory, what we want to believe is objective, is inextricably entangled with imagination, which we know is subjective. Thus, not only is memory just an *interpretation* of events, it is an unreliable one that is susceptible to change and influence. This information ultimately works as an important clue for understanding the climax.

During a tense and confusing confrontation near the end of the film, Teddy suggests that Leonard killed his own wife, revealing that his condition (the short-term memory loss) is like the perfect alibi. It works as the first line of defense in a multi-layered web of lies that Leonard has constructed for himself. According to Teddy, both Leonard and his wife survived the home invasion, and because his condition left him unable to lead a functional life, she became increasingly unhappy. Whether skeptical that he suffered from memory dysfunction or depressed with what life has become, she challenged Leonard to repeatedly administer her insulin shots every few minutes, as if she had not received any. Consistent with someone who was unable to remember he had done so, he complied with her requests until she suffered from a fatal overdose. Faced with yet another trauma, Teddy argues that Leonard projected this experience on to a former client he investigated, Sammy Jankis, convincing himself that it was *Sammy* who had done that to his own wife. Teddy insinuates all of this with several exchanges, the most significant of which is:

> You tell everyone about Sammy. Everyone who'll listen. "Remember Sammy Jankis, remember Sammy Jankis." Great story. Gets better every time you tell it. So you lie to yourself to be happy. Nothing wrong with that—we all do. Who cares if there's a few little things you'd rather not remember?

Regardless of whether we take Teddy to be telling the truth during this moment—he is, after all, at gunpoint—it is clear that the *only* way Leonard has learned to have a

meaningful life is through acts of self-deception, and in case the audience fails to make the connection, the next scene shows him manipulating his personal notes, knowingly and willingly exploiting his faulty memory by planting the seeds for the next target of revenge: Teddy himself.

While the movie is of course fictional, the psychological phenomenon that is central to the plot's progression has been the subject of countless debates. Indeed, the very notion of self-deception, at least under normal circumstances, seems to suggest something contradictory. It is obviously possible to deceive others, but how can it be possible to deceive *oneself*? Do we hold two contradictory beliefs at the same time, ignoring what we believe to be true in favor of what we wish to believe, or is it more circuitous? Perhaps, like Leonard, we simply plant the seeds of deception, gambling on forgetting about doing so in the long run? In the previous chapter, we saw how wishful thinking functions as a reputation-enhancing strategy whereby we mobilize into action a fantasy that, from our perspective, appears defensible to at least one community. Is this the same as self-deception?

5.2 | This isn't Your Grandfather's Paradox

Perhaps because of the name, there is a temptation to believe that self-deception resembles cases of standard deception, that it is somehow like lying to another person even though the target is oneself. Known in the technical literature as the *interpersonal model*, this is in fact one of the most common ways to understand self-deception.[89] One

[89] For the name of this approach, see: "Self-Deception" (Deweese-Boyd 2017, 2).

of the more prominent defenders of this idea was the existentialist Jean-Paul Sartre, and reconstructing his analysis helps reveal the limitations of this model.[90]

In one unfortunate example that nonetheless illustrates his understanding of self-deception, Sartre suggests that a "frigid woman" is one who adopts the belief that she cannot be aroused in order to avoid a fact about her freedom, that she really *can* (Sartre 1943 / 1956, 210–11). This is accomplished, he thinks, by avoiding situations in which sexual arousal might occur, and what motivates her avoidance is her assumption that it cannot happen in the first place. She thus fools herself into believing that she really is the kind of person she takes herself to be rather than embracing who she *actually* is, a human being whose freedom has no boundaries, including sexual ones. This, he believes, is something she has managed to do intentionally, for she has committed herself to this form of self-deception for a reason: to conceal the truth *about* herself *from* herself.

Given that there is present in self-deception an intention to deceive, Sartre looks to straightforward acts of deception to see if they can provide any insight into how self-deception might work, making the following observations:

> The essence of the lie implies in fact that the liar actually is in complete possession of the truth which he is hiding. A man does not lie about what he is ignorant of; he does not lie when he spreads an error of which he himself is the dupe; he does not lie when he is mistaken. The ideal description of the liar would be a cynical consciousness, affirming truth within himself, denying it in his words, and denying that negation as such. (Sartre 1943 / 1956, 204)

---

[90] Though Sartre defends the interpersonal model, he is largely concerned with *mauvaise foi* (frequently translated as "bad faith"), which is a very specific kind of self-deception that occurs whenever someone believes herself to be other than who she *actually* is. His analysis, however, is primarily metaphysical because he believes that a person is *actually* a free, self-determining being. As such, any time a person tries to restrict or limit herself by taking herself to be something *other* than this—by living in the world and playing roles—she is living in bad faith. Though this is partly illustrated by the example used above, I will be bracketing out these metaphysical commitments because I think that what Sartre has to say about lying in general and self-deception in particular creates an important context for understanding this phenomenon. Thus, I will be using "self-deception" in the place of "mauvaise foi."

Sartre thus advances three criteria that help isolate and distinguish deception from other phenomena that can appear similar by virtue of their consequences but which, upon closer inspection, have very different motivations and conditions for occurrence. Whenever a person, for instance, volunteers information in a state of ignorance or shares the faulty conclusion she has drawn, she is not lying even though her listener might be persuaded by these erroneous beliefs. To count as lying, Sartre believes the liar must (1) acknowledge the truth to himself, (2) speak as though the truth is otherwise, and (3) deny that he is being insincere.[91] Taken together, these three conditions are constitutive of an *intention* to deceive, to see the lie *as* a lie, for it is this intention that guides one's actions for the sake of the deception, determining what to say, how to say it, when to say it, and what information to withhold. Sartre clarifies, "The liar intends to deceive and he does not seek to hide this intention from himself nor to disguise the translucency of consciousness; on the contrary, he has recourse to it when there is a question of deciding secondary behavior" (Sartre 1943 / 1956, 204). The intention to deceive is what influences how the liar will speak and act.

Sartre believes that this same kind of intention is also at work in cases of self-deception; the difference, however, is that standard deception requires something further: a complex existential relationship between oneself and another. He explains:

> [The lie] presupposes my existence, the existence of the *Other*, my existence *for* the Other, and the existence of the Other *for* me. Thus there is no difficulty in holding that the liar must make the project of the lie in entire clarity and that he must possess a complete comprehension of the lie and of the truth which he is

[91] The third criterion on the surface looks unnecessary, but it helps distinguish lying from other speech acts that share in the first and second criteria, such as satire or dry wit, both of which can involving grasping some truth and distorting it for comedic purposes. When challenged, the comedian profess that she is "just joking," signaling that her distortion was not meant to be taken seriously. The liar, by contrast, wishes to present herself as sincere.

> altering. It is sufficient that an over-all opacity hide his intentions from the *Other*; it is sufficient that the Other can take the lie for truth. By the lie consciousness affirms that it exists by nature as *hidden from the Other*. (Sartre 1943 / 1956, 205; emphasis original)

Deception is thus contingent on the existence of another person for whom the falsehood is intended and to whom it is directed. This is made possible only by having the ability to conceal and reveal the contents of one's mental states—by being able to *misrepresent* oneself *to* another, to think one thing and say another.

The kind of misrepresentation that Sartre has in mind exploits the dividing line between oneself and another, the line that separates subjective perception from objective reality. With the exception of what is only inferred to exist, much of objective reality can be interpersonally experienced. In noticing a painting on the wall, I can inquire as to whether you notice it too, and if you do, then it constitutes a shared experience, especially as we confirm through conversation and behavior that we perceive things to be relatively similar. To be sure, each of us has our own point-of-view, allowing for divergence with respect to some of the details, but even a unique point-of-view does not prohibit us from, broadly speaking, experiencing the same thing.[92]

The world of interpersonal experience stands in contrast to the psychological space that is consciously accessible to you alone, the world of *mental* experience, which includes our thoughts, feelings, imaginings, beliefs, desires, etc. If I communicate to you that I am outraged by something that happened earlier in the day, you cannot experience

[92] I used to believe that a person's particular location in space-time was a basic constraint on shared experience. After all, you experience the world from your spatial location at a point in time that is different from mine. However, with the achievements in the field of virtual reality, it is looking increasingly likely that it is possible to have identical virtual experiences, including the subject's location within the experience. The only point of differentiation then becomes the mental states that a subject has through which she understands and interprets her experiences.

my intense rage in the same way that you can see the painting that I am admiring. You might be able to conjure up a memory of what it was like to be angry yourself or you might even be able to feel *similarly* enraged, but our mental experiences are uniquely our own, unable to be fully shared in all of their richness and nuance.[93] Fortunately, through performances and utterances, we can convey our mental representations to others, inviting them to feel similarly, to understand, to empathize, or even to change our minds, but it is precisely this discrepancy between shared and unshared experiences that makes lying possible. By discovering where shared experience begins and ends, the liar learns how to take advantage of another person. She begins to understand that she can see and experience herself in a way that is different from how another sees and experiences her, and as a result, she discovers that she can misrepresent those self-experiences, presenting something very different to the target of her deception while also obscuring her intention to deceive.

What Sartre thus finds peculiar about self-deception is that there is no division between shared and private experience to exploit. The other *is* oneself, and so if self-

---

[93] While empathic people might be said to experience another's emotional states, it is still nothing more than a *similarity* to that emotion. Logically speaking, Leibniz's law—the Identity of Indiscernibles, which states that $a=b$ if and only if for every property of a, b has the same properties—precludes this unless emotions are regarded as nothing more than a set of physiological responses, as William James had proposed. This, however, seems unlikely.

Antonio Damasio, for instance, argues that it is important to consider the mental experiences that lead up to the development of the emotions (Damasio 1994, 129–30). If these are important for the fine-grained texture of emotional experience, the odds of mirroring the identical train of thoughts of another person must be pretty small. Furthermore, Damasio argues, *even if* emotions are reduced to physiological responses, it is highly unlikely that the same changes in one person's body will be identical to another's, for it would require one to duplicate an unceasing cascade of changes in the body, such as blood vessel dilations, muscle contractions, neural activity, chemical signals, etc. (142–145).

For these reasons, it seems highly probable that shared emotions are at best extremely similar but never identical.

deception is analogous to lying, then it would imply that one is somehow misrepresenting oneself *to* oneself, exploiting a division that does not exist. Sartre explains:

> It follows [in the case of self-deception] that the one to whom the lie is told and the one who lies are one and the same person, which means that I must know in my capacity as deceiver the truth which is hidden from me in my capacity as the one deceived. Better yet I must know the truth very exactly *in order to* conceal it more carefully—and this not at two different moments, which at a pinch would allow us to reestablish a semblance of duality—but in the unitary structure of a single project. (Sartre 1943 / 1956, 205; emphasis original)

Here Sartre affirms the paradoxical nature of self-deception. If it is analyzed on the interpersonal model, then the apparent contradiction raised in the introduction appears in full view, for this model implies that the same person plays the role of deceiver and deceived, somehow able to lie to oneself without recognizing the lie *as* a lie. How is this possible?

5.3 | A Breakdown of Interpersonal Communication

One philosopher who is skeptical of the interpersonal model is Donald Davidson, as he finds the paradox it creates to be wholly unsatisfactory (Davidson 1997, 219). Like Sartre, he believes that straightforward deception involves intention and misrepresentation, and though he does not use the same language as Sartre, he also believes that straightforward deception requires the existence of an other who does not have direct access to the liar's own mental experience.

For Davidson, lying requires an act of communication, specifically, an *assertion*. Assertions are, by their very nature, public speech acts, and in the case of lying, any such assertions would be accompanied by the intention to be seen by someone as a person who is speaking both meaningfully and sincerely. After all, how successful can a lie prove if one is neither making sense nor is taken to believe what one is saying? Yet, there is also a

sense in which a liar does *not* want to be seen, for she wants to conceal her true motive. For this reason, Davidson identifies twin intentions involved in straightforward deception:

> The liar succeeds in deceiving his audience only if his intention to misrepresent what he believes is not discerned. On the other hand, there is an intention he must intend to be recognized, namely the intention to be taken as making an assertion, for if *this* intention is not recognized, his utterance will not be taken as an assertion, and so must fail in its immediate intention. (Davidson 1997, 215; emphasis original)

By deploying these twin intentions, a liar hopes to conceal the true motive in the guise of the superficial meaning of the assertions.

While it is not uncommon to lie by attempting to persuade another to believe what you are saying, there are a variety of other ways to achieve this goal as well. What is most important in an act of straightforward deception is not so much the selling of a falsehood as it is the directing away from truth by any means necessary. If someone understands that her mark takes her to be untrustworthy, for example, she might speak with the *expectation* that her target will mistrust her and do otherwise, effectively indirectly succeeding with the lie. She may even use the truth to effect the lie. The intention to deceive is thus not the same thing as an intention to persuade. Davidson elaborates on this in an earlier article, "Deception and Division," writing:

> While the liar may intend his hearer to believe what he says, this intention is not essential to the concept of lying; a liar who believes that his hearer is perverse may say the opposite of what he intends his hearer to believe. A liar may not even intend to make his victim believe that he, the liar, believes what he says. The only intentions a liar must have, I think, are these: (1) he must intend to represent himself as believing what he does not (for example, and typically, by asserting what he does not believe), and (2) he must intend to keep this intention (though not necessarily what he actually believes) hidden from his hearer. So deceit of a very special kind is involved in lying, deceit with respect to the sincerity of the representation of one's beliefs. It does not seem possible that this precise form of deceit could be practised on oneself, since it would require doing something with

> the intention that that very intention should not be recognized by the intender. (Davidson 1986, 207–208)

If Davidson's analysis of straightforward deception is correct, then the sting of the paradox becomes only more painful. How can one intend to appear sincere while intentionally concealing the true motive, especially when the victim and perpetrator are one and the same?

Rather than try to make sense of these questions, Davidson advises that we not take the interpersonal model too seriously, and he even suggests that we take care to distinguish self-deception from the idea of lying to oneself. While reflecting on the absurdity of the paradox, for instance, he questions whether a *New York Times* journalist, William Safire, accurately described Ronald Reagan as having "lied to himself" when it came to negotiating with Iran the release of United States' hostages in exchange for weapons.[94] He writes:

> William Safire may be politically astute and a whiz at language, but in this case he would have done better to say Reagan was self-deceived. The reason it is more plausible to hold that Reagan deceived himself than that he lied to himself is simply that, though the aim of lying to oneself, if this were possible, would be self-deception, there are less improbable techniques for achieving this end. (Davidson 1997, 215)

If self-deception is not to be analyzed with the interpersonal model, then what is the alternative? What does Davidson have in mind when he says there are "less improbable techniques" for deceiving oneself?

---

[94] See William Safire's "Essay; Truth from Shultz," *The New York Times*, Thursday, Feb. 4, 1993, Section A, p. 23. https://www.nytimes.com/1993/02/04/opinion/essay-truth-from-shultz.html.

5.4 | Unintended Consequences

Davidson opts for a *motivational model* of self-deception. The broad idea behind a motivational model is that self-deception is not caused by an intent to deceive, but rather, it occurs as a side-effect of a motivational state that inclines one away from undesirable beliefs or feelings. On a classical Freudian view, for example, self-deception might be said to occur whenever there is an unconscious motivation to censor painful or traumatic thoughts from consciousness. Later psychologists, such as Otto Fenichel, would make a similar argument, contending that we have defense mechanisms that aim to preserve one's self esteem (Simler and Hanson 2018, 75).

Though Davidson does not commit to any particular theory of motivation, he does illustrate what he has in mind with a Freudian-inspired example. He explains that if believing something proves to be *painful*, then one might form an intention to *avoid pain* (Davidson 1986, 209). The deception would then occur through a biased handling of the evidence for the painful belief, such as, "an intentional directing of attention away from the evidence in favour of [the belief that] *p*, or it may involve the active search for evidence against *p*" (208). He also makes a similar claim elsewhere, explaining that someone might find a belief to be "disagreeable or painful or ego-deflating," and this would thus motivate this person to act or think "in a way that causes him to reject the unwelcome thought" (Davidson 1997, 216). On this model, then, it is ultimately the intention to avoid some uncomfortable belief that motivates self-deception rather than the intention to deceive oneself, and the upshot of this theory is that it effectively eliminates the paradox.

As intuitively attractive as his theory might be, there is potentially a very serious issue. In a post-Darwinian world, any claim about mental functioning will need to hold up to the demands of natural selection. To say that a mental mechanism has a function of doing Φ is to make a *de facto* assertion that the mechanism in question is an *adaptation*, that the mechanism proves to be beneficial to reproductive success. Another explanatory possibility is that the psychological phenomenon is actually just a side-effect or malfunction of some *other* mental mechanism that works towards the evolutionary end. Robert Kurzban summarizes this issue succinctly:

> In short, and I can't emphasize this strongly enough, a fundamental issue that any theory of psychology ultimately has to face is that brains are *useful*. They guide behavior. Any brain that didn't cause its owner to do useful—in the evolutionary sense—things, didn't cause reproduction. (Kurzban 2010, 143; emphasis original)

Now, superficially, the motivational model meets these demands: self-deception is a side-effect of a mechanism designed to avoid some belief, specifically, an uncomfortable or painful belief. In other words, self-deception occurs as a result of a *protective* mechanism. But therein lies the problem. What does protecting oneself *do*? What behavior does it create that contributes to reproductive success?

It cannot be the case that the protective mechanism shields us from just *any* uncomfortable thought. After all, the vast majority of us still can and do entertain these kinds of thoughts. You might imagine how awful it would feel to be buried alive or you might remember what it was like when a loved one passed away. The protective theory will have to explain why none of these representations are suppressed. Is the protective mechanism bad at its job? Is there a homunculus who selectively discriminates between which beliefs he wants to pass on to us and which he wants to hide?

One possibility might be that the protective mechanism is automated, like a thermostat, and it will activate the protective response whenever some condition is met. If, for instance, some mental representation could trigger an emotional state so intense that it would cause *inaction*, then the protective response serves the adaptive function of allowing us to continue engaging with the world by short-circuiting an intense emotional surge, making it the cognitive equivalent of a circuit breaker. On the surface, this seems plausible, but does it hold up?

5.5 | Emotionally Charged

The neurobiological and psychological understanding of emotions—e.g. fear, anger, happiness, sadness, etc.—has come a long way. From an evolutionary perspective, emotions are adaptations that have developed to help organisms advantageously interact with their environments, usually to aid in assessment of situations followed by avoidance or approach responses, all of which can happen in a matter of seconds (Cozolino 2002, 235). Disgust, for instance, has its origins in prohibiting us from making contact with potential sources of disease, such as rotting meat or excrement, and it accomplishes this goal by triggering a cascade of physiological responses (e.g. appetite suppression, nausea, etc.) that discourages such contact (Haidt 2006, 185–6). Fear, on the other hand, heightens our awareness of the environment so as to more quickly identify possible threats. The aim of this emotion is, ultimately, to increase the odds of surviving a dangerous situation, usually by initiating a fight-or-flight or even a freeze response (LeDoux 1996, 128).

Even though the same emotions tend to give rise to approximately the same set of physiological responses, it does not follow that the environmental stimuli that can trigger

such responses are hard-wired by evolution. To be sure, some are, but they tend to be general in scope, such as open spaces, heights, etc. (Cozolino 2002, 236). Similarly, snakes are also generally fairly reliable inducers of fear.[95] But if *all* of our emotions worked this way, it would make for poor biological programming. Imagine if *every* environmental cue that initiated a fear response was pre-programmed. If the environment were to rapidly change, effectively eliminating most of the fear inducers, we would have wasted quite a bit of our cognitive resources. Equally problematic is that this would lead to a failure to respond to novel sources of danger. Were there all of a sudden to appear a new threat, such as a flying predatory goldfish that is the size of a sperm whale, our species will likely not be around much longer unless we can learn rather quickly to develop a fear response to its presence.

Fortunately, however, this is not how our emotions function. Even though *some* of the stimuli that can induce emotions are pre-programmed, most are the result of an exceptional fine-tuning to the particular environment in which we find ourselves, and this is made possible through a number of learning mechanisms, such as conditioning (Damasio 1999, 57). You may not have a fear response when you *first* encounter a dog, but after a few close calls with strays, you may start to feel anxious when you spot one wandering around without an owner or leash. As a result of this learned emotional connection, you will be able to better respond to your surroundings in the presence of stray dogs. Emotions can thus be both extremely powerful regulators of behavior but also extraordinarily plastic in their orientation to the environment. As Antonio Damasio puts it, "The consequence of extending emotional value to objects that were not biologically

[95] See, for example: *Snakes, Sunrises, and Shakespeare: How Evolution Shapes Our Loves and Fears* (Orians 2014).

prescribed to be emotionally laden is that the range of stimuli that can potentially induce emotions is infinite" (58).

So far, much of this seems to support the motivational model of self-deception. Taking the science of emotions into consideration, a motivational model theorist could speculate that what must be happening is that some kind of negative emotion (e.g. sadness or shame) becomes associated through experience with some mental representation, usually a belief. Due to the intense emotional connection formed, consciously entertaining that belief would compromise one's ability to meaningfully engage with her surroundings, promoting *inaction* rather than action. Because inaction is adverse to survival, a protective mechanism triggers an avoidance response that censors that belief from conscious experience and biases the interpretation of any evidence in support of it. This effectively short-circuits the intense emotional response that would have occurred, subsequently freeing an individual so that she can continue interacting with her environment. Unfortunately for the motivational model, however, this is neither consistent with the neurobiological underpinning of emotions nor is it what scientists have observed when awareness of sources of stress have *actually* been repressed from conscious experience.

5.6 | Gone but not Forgotten

One of the most fascinating features of the human brain is how so many different processes occur in parallel to one another rather than serially. This is no less true in the case of how the brain processes an emotion like fear. The central region involved in this is known as the *amygdala*, named for its almond-like shape, and it functions as the organism's alarm system, constantly monitoring the information received from another

region, the *thalamus*, which is a structure shaped like a quail egg through which all sensory information passes (LeDoux 2002, 214; Koch 2004, 124). One of the jobs of the thalamus is to hand off this information in two different directions, to the amygdala *and* the prefrontal cortex, the latter home to higher-level cognitive functioning. As Christof Koch explains, the functions of the different regions in the prefrontal cortex are "to guide, control, and execute outputs, such as skeletal or eye movements, and the expression of emotions, speech, or internal mental states" (Koch 2004, 129).

Why would the thalamus send its information to two different sites? Since the processing time in the prefrontal cortex takes longer, there needs to be something that can initiate immediate responses to any potential threats, and this is the amygdala. The clinical psychologist Louis Cozolino, for example, discusses the tradeoffs between accuracy and speed. Any increase in one of these attributes entails a decrease in the other, and in general, natural selection will tend to favor those that can respond the quickest (Cozolino 2002, 239). The beauty of the system we have in place, however, is that we can carefully balance both of those demands. It is not *always* good to respond without contextualizing the details. So while the thalamus passes the sensory information to the amygdala, which might sound the alarms, it also sends it to the prefrontal cortex to take a closer look.

When the prefrontal cortex receives the information from the thalamus, it can then bounce that information it has now processed to other sites in the brain, including the amygdala. So, for instance, in response to the information from the thalamus, the prefrontal cortex can help direct eye movement to more specifically focus on a potential environmental threat in an effort to gather more information. While this is taking place,

the amygdala immediately initiates the fear response as a kind of precautionary measure, causing an increase in blood pressure, secretion of adrenaline, increase in muscle tension, etc. (LeDoux 2002 121; 214). As this new information passes through the prefrontal cortex, however, additional adjustments are then made to more properly respond to the situation, some of which might include instructions to the amygdala to down-regulate its initial responses (217–8). Cozolino illustrates these complex interactions with a great example:

> I walked into the basement one day to look for a tool when, out of the corner of my eye, I saw a small brown shape near my foot. There are plenty of little critters in my neighborhood and they often crawl, burrow, or fly into my house. My heart skipped a beat and I immediately jumped back. My heart rate increased, my eyes widened, and I became tense, ready to act. Moving backward, I oriented toward the brown shape, saw that it looked more like a small piece of wood than a rodent, and began to relax. After a few seconds, my heart rate and level of arousal were back to normal; the potential danger had passed. (Cozolino 2002, 241)

He then follows this up with a neurobiological analysis of what took place:

> Analyzing this experience on the basis of the two systems, my peripheral vision saw the shape and my amygdala appraised it in an overgeneralized fashion to be a threat. My amygdala activated a variety of sympathetic responses including startle, increased respiration, and avoidance. In the split second while my body was reacting, I reflexively oriented my head toward the shape, which brought it to the fovea of my retina, providing my hippocampus and cortex with more detailed visual information; they then appraised the shape differently (and more accurately) than did my skiddish amygdala. (241)

One of the important roles, then, of passing information to the prefrontal cortex is to make more accurate assessments of environmental surroundings and, if necessary, dial down or ramp up the reactions. The implication this has for the motivational model of self-deception is that it is precisely *backwards*. Entertaining an unpleasant belief and processing it more slowly should help *decrease* the anxiety rather than amplify it; it should help a person see that what is feared is not actually that important.

In fact, the stronger an emotional state becomes, the easier it is to consciously access the information correlated with the emotional event from memory. This is why we are able to recall more precise details from events that have more emotional importance to us. Many of us who are old enough to remember can recall exactly where we were when the tragedy of 9/11 occurred, and yet, we often struggle to remember what happened yesterday. As LeDoux succinctly says, "Emotions, in short, amplify memories" (LeDoux 2002, 222). There is, however, an exception to this, but it does not work in the motivational model's favor.

LeDoux explains that if emotional arousal becomes too intense and stressful, it can actually *impair* memory. Part of the reason for this is that another region of the brain—a region that resembles a seahorse known as the *hippocampus*—is also involved in coordination between the thalamus, amygdala, and prefrontal cortex. Of its many functions, it plays a critical role in the formation of memories. But during a stressful period, the body will secrete a steroid hormone known as *cortisol*, which can weaken and disrupt the ability of the hippocampus to help form memories. Meanwhile, at the other end of fear processing, the increased stress actually *amplifies* the amygdala's signals (LeDoux 2002, 223–5). Joseph LeDoux believes that this combination of effects—impaired memory, failure to regulate fear responses through prefrontal activities such as thinking and reasoning, and amplified responses to fear—might be responsible for the formation of pathological symptoms found in conditions like posttraumatic stress disorder (PTSD), and others have implicated it in a host of other anxiety disorders,

ranging from panic attack to agoraphobia.[96] Interfering with the flow of information to the prefrontal cortex is thus not without serious consequence.

Taking all of the above into consideration, a protective mechanism designed to keep something painful out of conscious awareness simply does not make any neurobiological or evolutionary sense. It would come at the cost of losing the ability to reevaluate the information and hence modulate the emotional response. In turn, the lack of feedback from the prefrontal cortex would cause the amygdala to *elevate* said response, leading to a severe pathological condition, a consequence for which there is plenty of evidence from animal and human studies (LeDoux 2002, 217; 222–3). Furthermore, it raises the question of *why* such a mechanism would function that way in the first place. How can it be adaptively beneficial to *ignore* a potential threat? If it is about something as simple as feeling good, then, as Robert Kurzban rightly asks, "Why postulate that people are willing to believe false things in order to feel good about themselves, rather than suggest that people will believe true things about themselves, but just not feel bad about it" (Kurzban 2010, 146)?

If self-deception occurs, it certainly does not happen under these conditions, where pathological outcomes are debilitating rather than motivating.

5.7 | Playing Games

Much of what motivates the start of Plato's *Republic* in Book I are questions about justice. Gathered at the home of Cephalus, a wealthy foreigner in Athens, we find Socrates along with others conversing about how to precisely define this traditional

[96] For more on this, see: *Synaptic Self: How Our Brains Become Who We Are* (LeDoux 2002, 224); *The Neuroscience of Psychotherapy: Building and Rebuilding the Human Brain* (Cozolino 2002, 242–5); and "Anxiety disorders and GABA neurotransmission: A disturbance of modulation" (Nuss 2015, 172).

virtue, culminating in a tense exchange between himself and a sophist named Thrasymachus. Book II then begins with two more of those present, Glaucon and Adeimantus, expressing some anxiety that Socrates' responses to Thrasymachus were unconvincing. Particularly bothersome to them was a nagging worry about whether or not there is a *reason* to be just, one that is not simply instrumental. Does anyone desire justice for its own sake, or always for some advantage external to it?

From the perspective of Glaucon and Adeimantus, if we only care about justice for some reason other than justice itself, then we will have every reason to act unjust when the opportunity presents itself (*Republic*, 358a–359c). To help further make his point, Glaucon shares a tale about a shepherd named Gyges who happens to find a ring that can make him invisible when he turns it to the side. Upon discovering this power, we are told that "he immediately arranged to become one of the messengers attending the king, and went and seduced the king's wife, and with her attacked and killed the king and took possession of his reign" (360a–b). While it is always possible that Gyges suffered from deep character flaws—i.e. he always had the desire; the ring just provided opportunity—Glaucon goes a step further and speculates that if there were two of such rings, one given to a just person and one to an unjust person, both would end up making exactly the same decisions, acting in the service of injustice. The reason for this, he says, is that he fears that "whenever anyone thinks he *can* do injustice, he *does* injustice" (360c–d; emphasis added). In other words, if we think we can get away with something, we most certainly will try because it is to our advantage to do so.

Might this also be true in the case of deception?

Clearly, were a person able to lie and get away with it, she will have received what she wanted from the other party and will have sacrificed nothing. This is most evident in competitive games, such as poker, the premise of which is to maximize your gains by any means necessary within the rules specified. One popular strategy is to bluff, to pretend that the cards in your hand are much more valuable than they actually are. By succeeding, you increase your earnings that you otherwise would have been without and you lose nothing. This is not without its risks, however, because your opponent will be on the lookout for such deception, watching your eyes or your hands, listening to the sound of your voice, keeping track of behavioral tendencies, etc. As a result, in a game of poker, we find ourselves constantly engaged in weighing the risks and rewards of every hand played, trying to figure out if it is worth bluffing while assessing how difficult or easy it would be to get away with it. But what if, like Gyges, it was possible to *effortlessly* bluff? It would be the poker equivalent of turning invisible, for your opponent would have no signs to detect, no way to know or even reasonably guess whether or not you were bluffing. This kind of scenario is precisely what has lead to the development of a third model of self-deception, the *adaptive model*.

It is no secret that deception plays a key role in fitness and reproductive success at many levels in the biological world. One of the most common deceptive strategies used by countless organisms is known as *mimicry*, where one type of organism evolves to take on useful traits that resemble another type of organism. This form of deception can be used in the service of protection, predation, or even reproduction. When such deception is successful, it issues in an evolutionary game between deceiver and deceived, each developing new traits and abilities to try to gain the advantage over the other. Were one

to decisively lose this game, it will cease to exist (unless it finds new sources of food for its sustenance). It is important, however, not to suppose that deception necessarily implies any *intention* to deceive; mimicry is nature's way of showing how deception can occur in the absence of any such conscious motivations. Ethologist Richard Dawkins goes to great lengths to emphasize this point by providing a number of examples:

> Many edible insects, like the butterflies of the previous chapter, derive protection by mimicking the external appearance of other distasteful or stinging insects. We ourselves are often fooled into thinking that yellow and black striped hover-flies are wasps. Some bee-mimicking flies are even more perfect in their deception.[97] Predators too tell lies. Angler fish wait patiently on the bottom of the sea, blending in with the background. The only conspicuous part is a wriggling worm-like piece of flesh on the end of a long 'fishing rod', projecting from the top of the head. When a small prey fish comes near, the angler will dance its worm-like bait in front of the little fish, and lure it down to the region of the angler's own concealed mouth. Suddenly it opens its jaws, and the little fish is sucked in and eaten. The angler is telling a lie, exploiting the little fish's tendency to approach wriggling worm-like objects. He is saying 'Here is a worm', and any little fish who 'believes' the lie is quickly eaten. (Dawkins 2006, 64–5)

Might self-deception be akin to a kind of mimicry?

In the case of human beings, although deception can prove to be an extremely rewarding strategy, it is also incredibly costly. Not only are the risks associated with discovery high, but for many of us, lying requires that we keep our stories straight while juggling a number of behavioral cues that can give our intent away. Unsurprisingly, this proves incredibly cognitively demanding. Even the simplest act of deception might involve consciously managing facial expressions, posture, and limb positions in such a

---

[97] Mimicry first caught my own attention during hikes through wooded areas and fields, where I could pursue my hobby of macrophotography, a form of photography that requires the use of a special lens to gain a magnified view of smaller objects. In particular, I derive a great deal of pleasure from photographing insects.

On one occasion, I happened upon a species of robber fly (*Laphria grossa*) that remarkably resembled in both appearance and behavior a bumblebee, a fact that I only noticed after I was able to closely inspect the photographs I had taken. This form of mimicry enables the robber fly to close the distance with some of its favored prey, including bees and wasps, before they can identify it for what it is and evade it.

way that it looks natural; deliberately tracking and withholding the information that would compromise the deception; carefully directing our attention to the promotion of the right false information; and closely monitoring ourselves for idiosyncratic deceptive cues, such as frequent use of a particular word, blinking tendencies, abnormal use of eye contact, etc. (von Hippel and Trivers 2011, 3).

What the adaptive model of self-deception proposes is that it is best understood as the ultimate form of deception. Indeed, self-deception eliminates the need for any of this resource management by promoting superficial motivations while keeping the truth out of view. Like mimicry, this does not require that a person consciously possess the truth or otherwise know it, which helps the adaptive model avoid the paradox of self-deception. Psychologist William Von Hippel and anthropologist Robert Trivers, who have championed the adaptive model, explain the obvious advantage of such a strategy, "To the degree that people can convince themselves that a deception is true or that their motives are beyond reproach, they are no longer in a position in which they must knowingly deceive others" (von Hippel and Trivers 2011, 4). Like Leonard Shelby in *Memento*, self-deceivers lack the relevant information available to conscious recall, preventing them from detecting their own efforts to deceive.

5.8 | Offense Wins Games

Like wishful thinking, much of self-deception occurs by exploiting ambiguity in information and thereby avoiding anything critical that risks overturning one's beliefs. In this way, a deceiver can claim innocence and say sincerely that she really was unaware of evidence to the contrary; what she is not admitting, or even realizing, is that she was not motivated to look for such counterevidence in the first place. In psychology, this is

known as an *information-processing bias*, a phenomenon whereby an individual gives priority to desirable information over undesirable information (von Hippel and Trivers 2011, 2).

By analyzing self-deception in terms of information-processing bias, von Hippel and Trivers explain that this especially avoids the paradox of self-deception because it does not imply that an individual must somehow both know and not know the truth. Instead, she occupies an intermediate state between knowing and not-knowing, worrying that there *might* exist some undesirable information—known within the literature as *potential awareness*—while ignoring where she might find it. They explain how this both avoids the paradox and resembles interpersonal deception:

> In this case, however, true knowing of unwelcome information is precluded because the individual ends the information search before anything unwelcome is ever encountered. Thus, the individual need not have two representations of reality to self-deceive. Rather, people can self-deceive in the same way that they deceive others, by avoiding critical information and thereby not telling (themselves) the whole truth. (von Hippel and Trivers 2011, 2)

As they observe, when it comes to interpersonal deception, a liar does not need to be in possession of the truth in order to *try* to deceive—for she can simply be mistaken in her judgment as to what is actually true—but she does need to be *potentially aware* of the kind of information that could undermine her goals, taking care to avoid it.

On the adaptive model, self-deception is more analogous to this cautious form of deception rather than the deliberate and intentional distortion of the truth so commonly associated with lying. In these cases, what primarily makes the deception possible is not so much an active concealment of the truth as it is a *failure to be truth-sensitive*. Any interest in what might be true is deprioritized in favor of protecting one's goals. This behavior has been observed in people who avoid looking at new products after making

purchases,[98] as well as in people who avoid medical testing if they believe a disease is untreatable.[99] Similarly, a person may actively avoid potential sources of undesirable information even when she has the opportunity to learn otherwise, as illustrated by von Hippel and Trivers with the following scenario:

> For example, if a person is at a dinner party where one conversation concerns the dangers of smoking and the other concerns the dangers of alcohol, she can choose to attend to one conversation or the other–and may do so selectively if she is a smoker or a drinker. In such a case she would likely be aware of the general tone of the information she is choosing not to gather, but by not attending to one of the conversations, she could avoid learning details that she may not want to know. (von Hippel and Trivers 2011, 9)

Thus, self-deception occurs not as a result of repressing something the individual knows to be true; she simply has no interest in or access to the truth (or at least the *whole* truth), and she acts with that incomplete information, hoping it is to her advantage to do so (8). This means that self-deception really does serve a protective function, just not in the way supposed by the motivational model. Individuals are not protecting themselves from pain; they are protecting themselves from the possibility for truth insofar as they fear it will obstruct their goals.

5.9 | When One Paradox Closes, Another One Opens

Interestingly, this is not the conclusion reached by von Hippel and Trivers, who instead insist that truth-insensitivity is only secondary to the fact that self-deception evolved in order to better deceive others (13). But psychologist David Dunning speculates that such a strategy would have been *maladaptive* in our foraging past. While deceiving another

[98] See: "A new look at selective exposure" (Olson and Zanna 1979).

[99] See: "Genetic testing: Psychological aspects and implications" (Lerman et al. 2002); and "Don't tell me, I don't want to know" (Dawson, Savitsky, and Dunning 2006).

person can be advantageous under the right circumstances, the smaller, egalitarian group of foragers would have met such deception with hostility, for such close-knit communities were structured around cooperation and trust. You may be able to deceive someone today without worrying about the repercussions—a luxury afforded by living in areas much more densely populated—but, Dunning asks, "what about a human evolutionary past in which people did not move on, but rather woke up each morning to deal with the same small group of individuals for most of their mortal lives" (Dunning 2011, 19)?

Dunning finds even more support for his skepticism from small group studies that have shown that those who are overconfident in what they can do are initially well-liked,[100] but eventually, such individuals become the least valued of all within the group (Dunning 2011, 19). Psychologist Delroy Paulhus, for example, conducted two studies that tracked how *self-enhancers*—those who overestimate their abilities and accomplishments—were perceived by their peers over a number of weeks (Paulhus 1998, 1198). After using questionnaires and peer evaluations to narrow down those who met the criteria, Paulhus organized all of the participants into small groups of four and five members each. Each group was to meet once per week to work together and discuss a number of different topics, but they were prohibited from interacting with one another outside of their weekly meetings. The results from both studies were the same: "self-enhancers made a better impression than non-self-enhancers in the first meeting. By the seventh meeting, however, the reverse was true" (1201; 1203). He concluded that while

[100] For more on how overly positive self-evaluations are viewed negatively by both trained and untrained observers, see: "Overly positive selfevaluations and personality: Negative implications for mental health" (Colvin, Block, and Funder 1995); and "The quest for self-insight: Theory and research on the accuracy of self-perceptions" (Robins and John 1997).

there might be a *temporary* advantage to self-deception of this type, especially if interactions are brief and limited to strangers, once discrepancies between what one says and what one can actually do begin to emerge, these types of people are held in the lowest regard, likely finding their lives "to be characterized by chronic interpersonal conflict" (1206; 1207). This does not bode well for any purported adaptive advantages to be had by self-deception in the arena of interpersonal deception.

Although Paulhus did not recognize it in his research at the time, it also pointed in the direction of a new paradox that this adaptive model creates. He explains that *narcissism*—a personality type defined by feelings of superiority, entitlement, and self-admiration—also involves self-deception, especially of the self-enhancement type (Paulhus 1998, 1198). Whereas typical interpersonal deception involves some level of actively manipulating another person and her expectations, narcissism stems from what appears to be a sincere belief in these feelings of superiority (1205). This raises the question though as to how this can still be considered a form of interpersonal deception. As philosopher Ellen Fridland states, once "one has successfully deceived oneself, what one communicates to others, though untrue, is not deceptive" (Fridland 2011, 22). If the self-deceived holds the deception to be true, how can it be considered self-*deception*? Any expected payoff for a successful deception is forfeited whenever the deceiver believes her own lie and acts accordingly. It is the equivalent of bluffing in a game of poker and following through the completion of the hand even after someone has called your bluff. There is nothing to be gained from such a strategy and much to lose. On the adaptive model, then, there is no longer any meaningful difference between self-

deception and sincerely but mistakenly believing something to be true; any distinction between the two is collapsed.

5.10 | Defense Wins Championships

When the adaptive model's order of explanations is reversed, the problems associated with it disappear. Self-deception is *primarily* about truth-insensitivity, and as a secondary advantage, it can help confer strategic benefits during interpersonal interactions. This proposal obviously raises a question: What adaptive function could self-deception possibly serve, if not the deception of another? The answer is: *avoidance*.

What the motivational model gets right is the recognition that self-deception arises in response to a discomforting or distressful thought, but the error of this model results from the assumption that an individual must somehow *know* that the distressing thought is true and, as a result, repress that thought. On the contrary, what occurs during self-deception is that an individual entertains a fearful representation of what *might* be true, and rather than repress this representation, she treats it as she would any other overwhelming external threat by *fleeing* from it. Thus, the fear of what *might* be true (as opposed to what *is* true) recruits a flight-or-fight response that is turned towards the *mental* rather than the *external* landscape, resulting in biased information processing and selective searching of evidence to disconfirm and, hence, avoid or escape the fear.

For instance, in an act of self-deception, a subject can have a fearful, undesirable representation by imagining that there is a shadowy figure in the corner of the room, and in an effort to avoid what is feared, she can begin to seek out evidence to the contrary,

eventually forming the belief that things are otherwise.[101] It would be easy to confirm or disconfirm her fear by simply looking in the corner—an effort that would provide decisive evidence one way or the other—but doing so would also require confronting precisely what she hopes to avoid. She thus opts for evidence-gathering in a cautious, indirect way, reminding herself that the doors were locked, that there is a coat hanger that usually stands in the corner, that she had just watched a terrifying movie only hours earlier, etc. Because the fearful experience occurs *prior to* the formation of a belief that there really is or is not a shadowy figure in the corner to be feared, it is not a *belief* that she is avoiding, and because truth-values are only relevant to beliefs, it cannot be what is *true* that the subject is avoiding. Instead, she is taking measures to escape her fear. Regardless of whether or not there really is a shadowy figure in her room, by simply avoiding her fear, she is as a matter of consequence *truth-insensitive*; she avoids the truth without ever learning what it is in the first place.

Consider another example, this one from von Hippel and Trivers:

> A student whose parents want her to be a physician but who wants to be an artist herself might follow her conscious goal to major in biology and attend medical school, but her unconscious goal might lead her not to study sufficiently. As a consequence, she could find herself unable to attend medical school and left with no choice but to fall back on her artistic talents to find gainful employment. By deceiving herself (and consequently others) about the unconscious motives underlying her failure, the student can gain her parents' sympathy rather than their disapproval. (von Hippel and Trivers 2011, 7)

---

[101] An alternative to forming a *belief* that things are otherwise is *acting as if* things are otherwise. If an account of self-deception can be given in terms of non-linguistic mental representations and behavior, however, then it would imply that animals can be capable of self-deception as well.

There does seem to be reason to accept this position. For instance, the philosopher Sebastian Gardner concedes that Andrew Pomiankowski, a zoologist, has persuaded him that animals are capable of a kind of self-deception, writing, "An organism preparing for combat with another may put itself into a state which misrepresents its own power, and so avoid registering its own fear or debilitation" (Gardner 1993, 251, n.6). Even in the case of animals, what is referred to as self-deception appears to result from a desire to avoid something fearful.

They argue that this student is deceiving herself for the sake of gaining her parents' sympathy, but on the avoidance model, securing such sympathy can be construed as a byproduct of the fact that she undertook actions to avoid something she feared (e.g. that she may not be able to do what she truly loves, that she is not suited for medical school, that her parents do not respect her autonomy, etc.). Any of these fears could motivate the same set of choices that this student made, explaining her self-deception without relying on an intent to deceive her parents. Rather than argue that interpersonal deception is central to self-deception, the avoidance model explains it as a side-effect of self-deception.

This also better explains how self-deception could have arisen in our foraging past. A forager, fearing the possibility of ostracism, might deceive herself by avoiding any discussion of motivations for or rationalizations of her actions, which in turn helps her better defend her reputation and abdicate responsibility for anything bad that occurs. A similar psychological tendency has been observed in previous studies, most notably a classic conducted by psychologists Caroline Preston and Stanley Harris.[102] When they queried 50 drivers who had been involved in car accidents, not only did the overwhelming majority insist that they were better drivers than average, police records indicated that 34 of them were responsible for their accidents, going so far as to blame others as well as external conditions (Preston and Harris 1965, 287). Motivated by a fear—of loss of driving privileges, of what others might think, of legal repercussions, etc.—it is only natural to expect some level of self-deception to arise regarding what happened and why.

[102] See: "In one word: Not from experience" (Brehmer 1980); and "Are we all less risky and more skillful than our fellow drivers?" (Svenson 1981).

Like the adaptive model, the avoidance model also resolves the paradox of self-deception because an individual never believes contradictory propositions in the first place. Consider, for example, the following scenario: Upon developing a fear that Rafael looks bald, he takes steps to suppress this possibility before it can even reach the level of belief. It is his information-processing biases that "repress" the belief by causing him to not sufficiently entertain any evidence that allows it to form in the first place. To do this, he might entertain a belief that expresses a contrary state of affairs (e.g. that his hair looks fine) while seeking out evidence to strengthen it by, say, viewing himself under favorable lighting conditions or placing more weight on the flattering words of an admirer. Taking this evidence to be strong, he can then explain away any potential information that might reinforce his fear and ultimately yield the belief that he is bald. Rather than believe contradictory propositions, Rafael only needs to believe one thing, that he is not bald, because he never allows the contradictory belief to be formed in the first place. He can even resist entertaining it by avoiding fearful situations that support it.

Now, it is possible that someone like Rafael gets close to acknowledging his baldness through moments of doubt, perhaps asking, "*Am* I bald?" But by avoiding any evidence that would decisively determine the answer to this question, he is able to avoid the truth altogether, leaving the question of his baldness undetermined and never quite rising to the level of belief. The evidence that he seeks out increases from his perspective the *probability* that he is not bald, allowing him to believe *that*, but he otherwise avoids serious consideration of counter-evidence because that would require him to confront the fear and anxiety he is trying to avoid. Even though, as a result of escaping the fear that he is bald, Rafael may deceive others into seeing him the way that he wants to be seen as

well, there is no reason to invoke an intent to deceive them. The consequence of interpersonal deception is simply a by-product of his meticulous hair stylings that aid him in avoiding his fears.

5.11 | Sometimes You Are Better Off

Like wishful thinking, self-deception *can* be advantageous, and so on account of that, it is not *necessarily* irrational when it works. Rafael's delusions regarding his baldness might motivate him to style his hair convincingly, in turn leading to social benefits, such as others viewing him as confident or attractive. Likewise, the artist who fails to qualify for medical school might discover that her parents suddenly treat her better and take her own career wishes more seriously. The drivers who have been involved in accidents might be persuasive enough to avoid lawsuits, and foragers might avoid ostracism. Even in the wilderness, where veridical perception is critical, a self-deceptive approach can prove advantageous. While walking through a dangerous environment, for instance, I might notice something out of the corner of my eye, freezing in my tracks. Rather than abruptly turn to confirm that it is, in fact, a grizzly bear and begin to run for dear life, I might be better served by inductive inference, scanning the ground without moving my head to look for tracks, listening for growls, etc. In fact, it might even be better for me to *not* know that it is a grizzly bear, for then I will be able to better control my breathing and appear less like a threat or source of food. In this scenario, self-deception complements a *freezing* response, which itself is an avoidance strategy.[103]

[103] For more on how human beings respond to environmental threats, see: "The ecology of human fear: Survival optimization and the nervous system" (Mobbs et al. 2015). Interestingly, the authors speculate that the flexible cognitive systems that enable human beings to predict and avoid danger "may render humans vulnerable to psychopathology" (4).

Regardless of the advantages that self-deception can afford, however, it is also easy to see how it can become a problem and, consequently, irrational. This occurs whenever a self-deceptive state undermines, interferes with, or otherwise obstructs an individual's goals. Consider the following scenario:

> Near the end of the semester, Devon must take a final exam and receive an 85% to pass her course. Though she has never struggled in a course before and often believed herself to be intelligent, she experienced great difficulty this year. Feeling stress from this and other responsibilities, she looks over the course material and lacks confidence that she understands it. Realizing that it will take tremendous effort on her part to make any progress in developing mastery of the material—such as attending study sessions, visiting with her professor, scheduling tutoring appointments, and spending a lot of time reviewing the material on her own—she forms the belief that she cannot pass the exam no matter what she does. This belief allows Devon to avoid her fear that, in spite of her best efforts, she will never be good enough.

Is Devon deceiving herself into believing that she cannot pass the exam no matter what? There is no way to know with any certainty whether she *could have* passed the exam because she does not even try (though her history of academic success suggests that there was at least a chance). What is especially interesting is that it is Devon's belief that she is unable to pass the exam that makes true the fact that she does not pass the exam. In other words, she ends up believing something about herself that turns out to be true *because* she believes it is true in the first place. The choices she makes and the actions she pursues as a result of this belief is what guarantees her failure.

On the interpersonal and motivational models, it is not clear how to count this as a case of self-deception. How could Devon be avoiding the truth that she knows—which on these models would have to be the belief that she *can* pass the exam—if she takes herself to be embracing the truth that she is academically unfit to succeed in this course? For that matter, how can Devon even be construed as avoiding the truth at all if what she believes

about herself is what ends up being true? One could argue that Devon is avoiding the truth that she can pass the exam, but it is odd to refer to this belief as a truth since it remains in a state of uncertainty, neither true nor false but simply undetermined. The adaptive model, by contrast, would suggest that Devon has an unconscious motivation to believe that she cannot pass the exam in order to better deceive others. However, the only real target of this deception is herself, and so unless other assumptions are made—e.g. that she wants sympathy from her parents or that she wants an excuse to spend more time with her friends—this explanation does not adequately account for pathological forms of self-deception, especially when it exclusively affects oneself. It is only on the avoidance model that we can see the fuller picture, namely, why Devon believes what she believes, why it counts as self-deception, and why it is irrational.

5.12 | Conclusion

Two of the biggest obstacles facing any analysis of self-deception are making sense of the paradox of self-deception and figuring out how to distinguish it from wishful thinking. The former problem, I have argued, hinges on attempting to make sense of self-deception as a phenomenon analogous to interpersonal deception, a form of deception that requires two parties to take on different roles, one of deceiver and deceived. In the case of interpersonal deception, how a deceiver interacts with her target is shaped entirely by her intention to effect some kind of lie, to misrepresent the truth in some way, and so if self-deception is at all similar, it would require that one and the same person both know and not know the truth, that one can somehow fail to recognize her own lie *as* a lie.

Davidson, wholly unhappy with any analysis that requires lying to oneself and inspired by Freud, advances an alternative: some motivational states can pull us away

from undesirable beliefs or feelings. A person, for example, can intend to avoid discomfort, and as a result, incidentally end up in a self-deceptive state. The motivational theory thus no longer demands that a person must intend to lie to herself, but it comes at the cost of proving inconsistent with our understanding of evolutionary development and neurobiological functioning. It simply does not make sense, from an evolutionary perspective, to ignore threats, keeping the information away from conscious awareness.

The adaptive model, by contrast, errs on the side of evolutionary theory, considering the possibility that self-deception might even be an evolved trait, one that proves advantageous in social interactions. A standard act of deception, the adaptive theorist argues, is cognitively costly, requiring that the deceiver track the information she is attempting to sell, her facial expressions, her posture, the reactions from her mark, etc., her goal immediately compromised were she to fail in any of these regards. If, however, we could evolve to automatically and effortlessly deceive, then the advantage now rests securely in the deceiver's hands. This, the adaptive theorist insists, is self-deception, dispensing with any intentions altogether. As if equipped with a Ring of Gyges, self-deception just happens so as to accrue the rewards from a successful interpersonal deception. Yet, while this approach dissolves the traditional paradox of self-deception, it does so by introducing a new one: what is the difference between self-deception and mistakenly believing something is true? Self-deception, as a potentially interesting irrational phenomenon, thus collapses into the rather ordinary phenomenon of being sincerely mistaken.

In a world populated with half-truths, partial truths, and ambiguous information, self-deception can be accomplished by using information-processing biases, distorting

what one learns in the interest of promoting one's goals while ignoring any information to the contrary. This is what the adaptive model gets right. What it gets wrong is the assumption that self-deception evolved for the sake of enhancing personal reward from interpersonal interactions. There is little question that such self-deception *can* prove socially advantageous, but this is a side-effect of why we do this in the first place. On the contrary, self-deception, I have argued, is a consequence of a kind of fight-or-flight response, motivating us to flee from the unknown that we fear. It is a truth-insensitive mechanism designed to optimize our survival in both natural and social worlds, and like any biological or mental mechanism, sometimes it can backfire. When it does so, it becomes an obstacle to our success, and it is at this point that it becomes irrational. This is the avoidance model of self-deception, and it preserves the insights from the adaptive model without introducing a new paradox.

On the avoidance model, self-deception is distinguished from wishful thinking precisely in its aims. Whereas wishful thinking occurs as a reputation-enhancing or -preserving strategy, self-deception is not limited to the social environment and is centrally preoccupied with fear avoidance. There are, of course, instances where the two phenomena coincide with their goals, and so may be jointly cooperating. There is little doubt, for example, that Linus Pauling's faith in Vitamin C was a result of wishful thinking *and* self-deception. There is no reason why the two phenomena cannot at times coexist, especially when their aims coincidentally intersect.

Chapter 6: Siren Songs

In the first place,
we recognize that there can be good and bad thermometers.
In the second place,
we do not even demand that a good thermometer provide a reliable reading under every condition.
We recognize that there may be special environmental conditions under which even a 'good' thermometer is unreliable.

— David Armstrong, *Belief, Truth and Knowledge*

6.1 | Decisions, Decisions

Omar is traveling home from work during rush hour. He has made this trip, or so it feels, thousands of times, and though there are multiple ways to make it home, he knows exactly which route is both the quickest and least stressful. He also has every reason to get home quickly: his son's baseball game starts at 6:30 PM, and he promised him that he would be in attendance. While on his way home, Omar faces the following two options: take Madison Avenue or Jefferson Boulevard. Over the past three months, he has consistently taken Madison Avenue, keenly aware that Jefferson Boulevard is the road that *everyone* takes, and even his navigation software is advising him to take Madison Avenue. At the last minute, however, Omar knowingly and willingly veers over into the lane for Jefferson Boulevard. It takes him an extra 45 minutes to get home, and as a result, he misses seven innings of his son's baseball game. When confronted by his son after the game, Omar apologizes, explains the traffic situation, and even confesses that he had known better. Supposing there were no ulterior motives, why would he have taken Jefferson Boulevard when Madison Avenue was clearly the better option, especially when he *knew* it was the better option?

In philosophical literature, we describe what happened to Omar as an instance of *akrasia*, a term often translated as "weakness of will" or "incontinence." Some of the earliest accounts of this phenomenon in the western tradition are attributed to Plato, but perhaps the earliest *technical* exposition of it comes from Aristotle's *Nicomachean Ethics*.[104] From an observer's point of view (and Omar's as well) it appears as though he acted against his better judgment. Why did he do this though? Are all of us susceptible to moments of weakness like this?

Such questions, of course, presume that akrasia even exists in the first place. It might be argued, for instance, that it is little more than an illusion. There is, after all, a serious obstacle for judging exactly when this occurs: we inevitably have to rely on interpreting what an individual says and checking it against what she does. This is true even if the individual is oneself. But should such utterances and beliefs be taken at face value?

Consider the case of a stroke patient, as presented by Vilaynur Ramachandran. Keeping in mind the principle of cross-communication—that any damage to the left hemisphere of the brain will affect the right side of the body and any damage to the right will affect the left side of the body—a common consequence of stroke is usually some kind of paralysis on one side of the body. In this particular case, Ramachandran met with an "intelligent and lucid" patient who had suffered from right hemisphere damage. Due to her injury, she was unable to move her left arm, but there was an interesting twist. Because one of the affected areas was the *superior parietal lobule*, which is a rectangular-shaped area of the brain associated with "body image" representation, she

[104] See, for example *Protagoras* 351b–358d and *Republic* 437d–438a.

could not recognize her own paralyzed arm. For example, when Ramachandran pointed it out to her, she claimed that it belonged to her father who was "hiding under the table," and yet at the same time, when he requested that she touch her nose with it, she picked it up with her right arm to complete the task, betraying that at some level she understood it was paralyzed. Ramachandran even pressed her further, showing her how the paralyzed arm was attached to her shoulder, and though she agreed, she still insisted that it belonged to her father (Ramachandran 2009). So not only does she have an implicit awareness of her paralysis that she does not or cannot explicitly acknowledge, she also confabulates absurd explanations for her dysfunction.

Perhaps an akratic, then, is more like this patient?

There is no denying that a case with a brain-damaged patient constitutes an exceptional example, but recall that many more cases of confabulation were discussed in section 3.7, cases involving otherwise neurologically healthy people. Given the frequency with which we try to rationalize our own behavior *ex post facto*, might an akratic simply be someone who confabulates that she knows what is best even though she does otherwise? If anything, these cases provide ample reason to be cautious when someone insists that they truly know what is best.

And yet, in spite of our propensity for confabulation, there still remains a very serious philosophical problem. Enlightenment thinking leads us to believe that reason is our savior, and so if reason cannot save us from akrasia, then what can we do about it? Are we fated to be victims to our unconscious motivations, drives, and desires? Some traditions even hold that who we are *is* the voice of reason, and if that is true, then it

appears as though we are nothing more than spectators during akratic moments. Philosopher David Pears raises these concerns well:

> This identification [of reason and self-identity] is a matter of some consequence. If we go along with it, we shall encounter a problem. When someone acts against his own better judgement the usual reaction is exasperation: 'What more could reason have done to keep him on the rails?' If his deliberation could not have been better, there was nothing more that reason could have done. So the verdict will be that reason could not have prevented the derailment. But now if we substitute *him* for his *reason* in this verdict, we shall be in trouble, because we shall have to conclude that *he* could not have prevented the derailment. (Pears 1998, 16; emphasis added)

Central to the problem of akrasia is a concern about the *power* of reason to make a difference in human action. If reason (or the self) can be overpowered and rendered impotent at any time, then what assurance is there that reason is actually doing *anything*? What can we do about it?

6.2 | Just Another Day of You and Me in Paradox

Imagine, for instance, if in deciding whether to let his ship sink or throw his cargo overboard, a captain concludes that throwing cargo overboard is the better course of action, and yet in spite of arriving at this conclusion, he refuses to do so, deciding to go down with his ship instead. Why would he do such a thing? How does this make any sense? If the captain has indeed judged that it is better to toss his cargo, why would he do otherwise, for then he fares all the worse? Somehow, his intention to do what is best has failed to cause the appropriate action.

In his essay, "How is Weakness of the Will Possible?," Donald Davidson argues that action theorists tend to be reluctant to affirm the existence of akrasia because it appears to contradict our understanding of the role of practical reasoning in producing an

intentional action. In a typical case of intentional action, it is believed that an individual goes through a process of means-end reasoning to get what she wants. Davidson explains:

> When a person acts with an intention, the following seems to be a true, if rough and incomplete, description of what goes on: he sets a positive value on some state of affairs (an end, or the performance by himself of an action satisfying certain conditions); he believes (or knows or perceives) that an action, of a kind open to him to perform, will promote or produce or realize the valued state of affairs; and so he acts (that is, he acts *because* of his value or desire and his belief). (Davidson 1969c, 31; emphasis original)

Though vague in the details, this is a fairly common account of traditional intentional action. If a person desires to practice yoga on the top of a mountain, then clearly, she will need to determine if this is possible and in what way. Once she has done so, provided that she really intends to do this, she will act accordingly to make it happen. Both the desire and the beliefs regarding how to satisfy the desire together help form an intention; they both cause the action and explain why it is done.

To help better illustrate the relation between intention and action, consider Davidson's reconstruction of an Aristotelian-inspired practical syllogism that explains why an agent performs the action of looking at his watch (Davidson 1969c, 31):

Action Description	Practical Syllogism
1. An agent desires to know what the current time is	P_1 Any act of mine that results in knowing the time is desirable
2. He believes this can be done by looking at his watch	P_2 Looking at my watch will result in my knowing the time
3. He looks at his watch	C [Action of looking at watch]

The practical syllogism on the right demonstrates how the agent from the example on the left goes through a process of reasoning that results in his action. In this particular case, what we find is that his desire to know the time, represented in the major premise, informs what he would like to do; and in the minor premise of the syllogism, we can see the belief he has regarding how he can satisfy that desire (31). Taken together, both

premises form the intention, and the conclusion of his reasoning just is the action recommended by the premises.

The idea that the conclusion of a practical syllogism can result in an action is one that Davidson borrows from Aristotle, explaining:

> Aristotle says that once a person has the desire and believes some action will satisfy it, *straightway he acts*. Since there is no distinguishing the conditions under which an agent is in a position to infer that an action he is free to perform is desirable from the conditions under which he acts, Aristotle apparently identifies drawing the inference and acting: he says, 'the conclusion is an action'. (Davidson 1969c, 32; emphasis original)

For a traditional account of agency, undertaking the action is the litmus test for desirability. If an agent is free to look at his watch, desires to know the time, has no countervailing desires, and believes that looking at his watch is the best way to know the time, then he *will* look at his watch. And so, in this case, to judge that the action is desirable just is to intentionally perform the action.

So how does this work out when there *are* countervailing desires?

Again, the traditional way of viewing the matter is to assume that an agent will act for the sake of what is *most* desirable. Why would she settle for something less desirable when there is no reason to do so? This can be summarized with the following two principles (Davidson 1969c, 23–4):

(α) If an agent wants to do *x* more than she wants to do *y* and she believes herself free to do either *x* or *y*, then she will intentionally do *x* if she does either *x* or *y* intentionally.

(β) If an agent judges that it would be better to do *x* than to do *y*, then she wants to do *x* more than she wants to do *y*.

The first principle (α) is, for a traditional action theorist, trivially true. It merely states that to act intentionally is to act for the sake of what one desires, and when there are multiple desirable actions that can be performed, then a person will intend that action that

she desires most, provided she believes it is possible to act accordingly (22). Suppose, for example, that Fran desires chocolate ice cream more than strawberry. If he is free and able to have both, and if he finds it worth acting on this desire, then he will intentionally act and have the chocolate ice cream.

The second principle (β) is a variation of the *Socratic Thesis*. Within the western intellectual tradition tracing at least as far back to Socrates, this is the long-standing thesis that associates desire with the wanting of something perceived to be good. To want something, according to this thesis, is to see something good in what is wanted, to recognize what appears to be something beneficial for oneself. In Plato's *Meno*, for example, Socrates persuades Meno to accept his conclusion that *everybody* desires good things, arguing that what is bad is harmful and nobody wants to be harmed (*Meno*, 77c–78b). Within the context of practical reasoning, this second principle more specifically entails that, when multiple courses of action are available, if a person *judges* one to be better than another, she desires most the one that she judges best. So again, if Fran judges that having chocolate ice cream is better than strawberry, then he desires having chocolate more than strawberry.

According to Davidson, when taken together, α and β imply a further principle (Davidson 1969c, 23):

(γ) A freely acting agent will intentionally do whatever she judges it best to do.

Recall that, for Davidson, to act intentionally is to judge that some course of action is desirable enough that it is worth putting into action, and so when combined with this new principle (γ), the conclusion an action theorist reaches is that an intention is not *just* a judgement that a set of beliefs and desires are worth acting on, it is a judgment that this is

what is *most* desired. But if this account of intentional action is correct, then how can akrasia be possible? According to Davidson, akrasia requires that a freely acting agent intentionally acts *contrary* to her judgment concerning what it is best to do (23). Is this not a contradiction in terms?

So what is happening in cases of akrasia? Is the agent *not* freely acting, under the spell of some powerful desire that overrides her intention? Maybe she *says* that she intends one thing but *really*, in the moment, intends the other? Or is she truly, intentionally acting for the sake of a worse course of action? If so, might something be wrong with one of the principles of practical reasoning or the traditional account of intentional action?

6.3 | In a Battle of Wills, Who Wins? The Stronger or Smarter?

Davidson believes that there have basically been two strategies in understanding how akrasia could occur: what I will call the *strength strategy* and *knowledge strategy*. Those who adopt the strength strategy will argue that, when two or more desires endorse different goals for an individual, whichever is the strongest emerges the victor. In the case of akrasia, although one may have reasoned in support of one goal, some other desire intervenes, overpowers reason, and guides the individual's action instead.

Proponents of the knowledge strategy, on the other hand, will argue that each desire makes a proposal about what is best, and after weighing considerations for each, an individual simply chooses whichever appears most convincing. This is not to suggest that all desires produce *arguments* (although that is one interpretation of a knowledge strategy), but at a minimum, all that such a strategy requires is that desires can present

something *like* an argument, something an individual can find intelligible, potentially persuasive, and actionable.

To see what each strategy entails, it is helpful to look at the dilemma that an incontinent person faces. Davidson proposes that this conflict can be characterized with the following two syllogisms (Davidson 1969c, 33):

The Side of Reason		The Side of Lust	
R_1	No fornication is lawful	L_1	Pleasure is to be pursued
R_2	This is an act of fornication	L_2	This act is pleasant
C_R	This act is not lawful	C_L	This act is to be pursued

Davidson uses the "Side of Lust" as his example, borrowing it from what he takes to be Aquinas' position.[105] Though this is part of his example, this is not meant to suggest that the side of lust is *the* side that challenges the side of reason. Nor is this to suggest that both sides *actually* reason (though again, that is one interpretation). The above is intended to be used for purposes of illustration, to show how some set of beliefs and desires can differ from another set, creating the conditions for akrasia.

According to the strength strategy, which Davidson refers to as the Aristotelian-Thomistic approach, these two sides compete with one another, and when akrasia occurs, it suggests that the "wrong" side has emerged victorious, overpowering the "right" side (Davidson 1969c, 35). Davidson, however, argues that this strategy ought to be rejected. For one, he believes it strips the agent of responsibility and thereby undermines the notion that the agent is performing an action in the first place. He writes:

> On the first story [strength], not only can we not account for incontinence; it is not clear how we can ever blame the agent for what he does: his action merely reflects the outcome of a struggle within him. What could he do about it? (35)

[105] See *Summa Theologiae*, I, q.77, a.2 ad 4.

If there is no way for the agent to influence how the drama between these practical syllogisms plays out, then she is not freely acting, a condition that Davidson's first principle states is necessary for intentional action.

As an extension of the first problem, Davidson also argues that the strength strategy precludes the possibility for any meaningful evaluation of how to act (Davidson 1969c, 36). How does an agent arrive at the conclusion that the action recommended by reason is *better* than the action recommended by lust without simply begging the question that reason is always better than lust? After all, central to the problem of akrasia is that the agent believes one course of action is better than another but decides to pursue the weaker of the two, *not* that the agent sees a course of action recommended by reason but decides to pursue the course of action recommended by lust.

The knowledge strategy, which he attributes to Plato, is able to avoid these issues, but not without introducing its own. This proposes that we introduce a third syllogism that corresponds with the decision-making of the agent (or even just the agent herself). This syllogism's role is to entertain both sides, weigh the arguments, and make a decision, taking responsibility for the outcome (Davidson 1969c, 36). Again, for illustrative purposes, Davidson calls this third part "The Will" or "Conscience," and its syllogism takes the following form (35–36):

	<u>The Will (Conscience)</u>
W_1	R_1 and L_1
W_2	R_2 and L_2
C_W	This action is wrong

Though not a logical syllogism, what the syllogism of the will captures is the ability to take into consideration *all* of the arguments and make an evaluative judgment concerning which is the better. Here, it acts very much as a judge would, as if in the courtroom of the

agent, and it hears the cases from each, the side of reason and the side of lust, before ruling in favor of one or the other.

However, as Davidson explains, even though this resolves the issue of how the agent arrives at the belief that one side is better than another, it does so by reintroducing the problem it was intended to explain, namely, it leaves unclear just how an agent can believe that one side is best and yet act otherwise. For according to the knowledge strategy, the will hears both sides, evaluates them, and executes just one. In so doing, it effectively *endorses* one side as the better of the two, as winning the case (Davidson 1969c, 36). If the will is somehow different from the agent such that what it determines is best is not identical to what the agent thinks is best, then what purpose is it serving? Similarly, if the will is not responsible for acting in accordance with its judgment, then what is the evaluation supposed to *do*? Is there a fourth part that causes the action? If so, then which argument does it favor and how does *it* evaluate them?

Though Davidson concludes that both strategies are unsatisfactory, he believes that the knowledge strategy is a step in the right direction. For him, the strength strategy ultimately begs the question by designating one side as "good" and the other side as "bad;" by contrast, the knowledge strategy allows for the possibility that an agent can compare and contrast the sides to determine if the recommendations *really are* good or bad.

So is it possible to make sense of both akrasia and agency? Davidson proposes that we can if we modify the knowledge strategy appropriately.

6.4 | A Contest of Intellect

The mistake that the knowledge strategy makes, according to Davidson, is that it assumes that the conclusion that the will reaches is an intention to act. If this were true, then we should expect the judgment the will reaches to *cause* an intentional action, but this is not what occurs. Therefore, Davidson argues, the will must not be engaging in practical reasoning after all. Instead, its function is to reach an all-things considered (*atc*) judgment (Davidson 1969c, 36).

By hearing from both sides, the will acts as a mediator who brings the argument in favor of lust into contact with the argument in favor of reason so that the agent can see both arguments together. While doing so, a comparative judgment is made as to which side looks better while considering all of the reasons presented. As explained in chapter two, an *atc* judgment is *conditional*—a judgment that such-and-such is desirable on the condition that it is evaluated in light of all the available reasons—and this is very different from an intention, which is an all-out or *unconditional* judgment that such-and-such is desirable *tout court*.

Because the will's role is to form an *atc* judgment rather than an *intention*, Davidson believes the tension between traditional agency and akrasia is relieved. On his theory, the will of the individual forms an *atc* judgment that one course of action, say, *reason*, is better than another course, say, *lust*. If an individual acts on the basis of lust *in spite of* her *atc* judgment that reason is better, then she acts for the sake of lust. There is no requirement that an individual *must* act in accordance with her *atc* judgment, and since she had a reason for acting in accordance with lust, she acts *intentionally* (Davidson 1969c, 39).

Compare how a continent individual acts: when giving a hearing to the side of reason against the side of lust, this person takes into consideration the premises that recommend both courses of action in order to arrive at an *atc* judgment as to which is best. When she acts in accordance with this *atc* judgment, she acts intentionally *not* because she is following some practical syllogism that lead to the *atc* judgment—for as Davidson argues, the *atc* judgment is not reached by a practical syllogism—but because she is following the practical syllogism that the *atc* judgment *recommends*, namely, that which is on the side of reason. Conversely, for Davidson, this means that to act akratically is to follow the practical syllogism that the *atc* judgment does *not* recommend.

But why would we do such a thing in the first place?

To act in accordance with the recommendation of the *atc* judgment is to act continently, and such persons that do so with regularity are those who recognize the value of what Davidson calls the *principle of continence*. This principle is only a *guide* for the rational person; it does not *determine* her choice. Like the fictional angel on our shoulders, such a principle "exhorts us to actions we can perform if we want; it leaves the motives to us" (Davidson 1969c, 41). But this raises the question as to why *anyone* would *not* follow this principle in the first place, much like asking a wise person for advice and doing the opposite. For Davidson, *this* is the problem of akrasia and this is what makes it irrational. He explains:

> But what, on this analysis, is the fault of incontinence? The akrates [i.e. akratic person] does not, as is now clear, hold logically contradictory beliefs, nor is his failure necessarily a moral failure. What is wrong is that the incontinent man acts, and judges, irrationally, for this is surely what we must say of a man who goes against his own best judgement. (41)

And a little further in the text, he reflects:

> Why would anyone ever perform an action when he thought that, everything considered, another action would be better? If this is a request for psychological explanation, then the answers will no doubt refer to the interesting phenomena familiar from most discussions of incontinence: self-deception, overpowering desires, lack of imagination, and the rest. But if the question is read, what is the agent's reason for doing *a* when he believes it would be better, all things considered, to do another thing, then the answer must be: for this, the agent has no reason. (42)

If Davidson is right, then the mystery of akrasia is why knowledge fails to persuade the individual as it should. A person does not act incontinently for no reason, for she has reasons for doing as she does; instead, she acts incontinently because there is no good reason for following the weaker argument.

This is, indeed, the central mystery of akrasia, but the knowledge strategy will not provide any answers. For one, as argued throughout this project thus far, Enlightenment thinking has lead traditional action theory to overestimate the role of reason in producing actions, and so the theory itself is in need of rehabilitation to better reflect what we know about how reason works and how it influences actions.

Two, Davidson's own solution, while interesting, leaves many questions unanswered. If the will is not *executing* any actions but is only forming *atc* judgments, then who or what *is* forming the intentions and causing the actions? Is this part—for the sake of convenience, I will refer to it as the *agent*, leaving open the question of whether it is an additional part or somehow the whole person—rational, non-rational, or maybe somewhat rational? Are all desires capable of forming practical syllogisms and recommending actions? Above all, is Davidson's solution even appreciably different from the classical knowledge strategy he attempts to modify? Whereas that strategy unifies the functions of evaluation and execution in the part of the will, Davidson merely separates those two functions from one another, leaving evaluation up to the will and

execution up to the agent. Recall, however, that his own criticism of the knowledge strategy is that it leaves open the question of how the will determines which side is the victor and worth acting on. While his modification clarifies how the will determines which side has the better argument, it does *not* explain how the agent determines which side is worth putting into action.

Perhaps there is virtue in strength?

6.5 | A Contest of Strength

As David Pears explains, when rendered literally, *akrasia* is better translated not as "weakness of will" or "incontinence," but rather as "lack of strength or power" (Pears 1998, 23). And in particular, he writes, "the un-negated root [*kratos*] is the ordinary word for victory or domination" (23). Understood in this way, the word was intended to make comparisons, as in, "The taller combatant lacks the strength to secure victory over the shorter, tougher combatant." To describe a phenomenon as akrasia, then, is to suggest that there is something that lacks the strength to overcome something else, and within this context, the implication is that something non-rational overpowers something rational. This idea can be traced back to Aristotle, who happens to describe akrasia in similarly vague but combative terms early in his *Nicomachean Ethics*:

> For exactly as in the case of parts of the body subject to muscular spasms, when one has chosen to move to the right they are on the contrary turned away to the left, so too is it with the soul, for the impulses of unrestrained people are contrarily directed.[106] But while in the body we see the part that swerves, with the soul we do not see it. Nevertheless one must presumably consider there to be something in the soul as well contrary to reason, which opposes it and stands in its way, though it makes no difference in what way it is distinct. (*Nicomachean Ethics*, $1102^{b}18–25$)

[106] Joe Sachs, whose translation I am following here, uses the term "unrestraint" to translate akrasia.

Was he on to something? If taken in this way, there is no need to pit practical syllogism against practical syllogism, as Davidson does, since what emerges is actually a contest of strength. In cases of akrasia, might it be the case that something nonrational influences, steers, or even overpowers reason itself?[107]

Now, if the muscle spasm analogy is pushed to the limit, it becomes difficult to see how akrasia can be anything but involuntary (much like a reflex beyond the control of the individual), and in that case, it no longer becomes a psychological problem any more than an autonomic process in the body is psychologically regulated. Is there a way to frame akrasia as a problem of strength without collapsing it into an involuntary phenomenon? This is precisely what Aristotle attempts to figure out, and his insights prove instructive. His analysis begins with a consideration of the arguments against akrasia before sorting out what might cause it.[108]

According to Aristotle, there are generally two common strategies used to deny that akrasia exists. Both inspired by the Socratic Thesis, the first position argues that the very idea of akrasia is self-contradictory because "no one acts contrary to what is best while believing that to be the case, but only from ignorance that it is" (*Nicomachean*

---

[107] Although Aristotle seems to explicitly deny this at *Nicomachean Ethics* 1147^{b}5–6, Pears entertains the possibility that akrasia could arise even in cases where reason is not involved at all. If any organism possesses *something* (even if it is not reason) that guides its actions best, something, in other words, that reliably directs it to act in its own best interests—such as a "system of primitive emotions, like fear and anger, triggering the stereotypical strategies of avoidance or aggression"—and there were some desires that overpowered *that* such that it leads the creature to act in a way contrary to its best interests, then there may be grounds to consider it a kind of akrasia. This would challenge us to widen our conception of it (Pears 1998, 17).

[108] I consider the following section to be an *Aristotelian* approach rather than *Aristotle's* approach because it draws upon ideas from interpretations of Aristotle by several scholars, notably David Pears, Amélie Rorty, and David Wiggins, each of whom focuses on a different point in Aristotle's analysis of akrasia. I will be following each of them *on* those points, combining it into a larger, comprehensive account of akrasia. The goal of adopting this strategy is simply to use Aristotelian ideas to frame and understand a problem rather than offer an authoritative, standard account of Aristotle.

Ethics, 1145^{b}26–27). Unsurprisingly, Aristotle even attributes this approach to Socrates himself. One version of the Socratic Thesis maintains that an individual always acts in accordance with what she believes is best, regardless of any insistence that she might give to the contrary. The thesis itself is non-falsifiable, for any discrepancies between what one says and does can be explained away by assuming that an individual is somehow being insincere in her self-reports or otherwise lacks reliable introspective access to her thoughts. I will refer to this as the *Behaviorist Strategy* since it effectively calls into question the value of mental states in influencing what one does; on this strategy, to believe just is to do. In other words, this strategy suggests that any actions that a person undertakes are implicit acknowledgments of what she *really* believes.

The second position, by contrast, affirms that there are cases of genuine akrasia, but it argues that they only occur when an individual fails to act with *knowledge* of what is better, instead acting only with *opinion* or *mere belief* (*Nicomachean Ethics*, 1145^{b}32–35). This position thus maintains that an individual will *necessarily* act in accordance with what she *knows* is best, and so if some nonrational part or desire overpowers her, it must be the case that it overpowered a *belief* rather than *knowledge*. What this solution does, then, is short-circuit the fear that reason or knowledge is powerless in the face of the nonrational. I will refer to this as the *Meno Strategy* because it more closely resembles Socrates' argument in Plato's *Meno* that nobody willingly errs (*Meno*, 77a–78b).

Both of these strategies ultimately make the same move; they assert that knowledge cannot be overpowered, for in both cases, if one *knows* what is best, one will *do* what is best.

Against the Behaviorist Strategy, Aristotle believes that it "disputes things that are plainly apparent" (*Nicomachean Ethics*, 1145^{b}28). That akratic acts occur is clear, and so he finds it unsatisfactory to not only deny but even explain them away by entirely writing off the value of self-evaluations and introspective access. Now, it is true that there is good reason to doubt the reliability of verbal self-evaluations.[109] It has long been known within the field of psychology that there is indeed a credibility problem at the heart of introspective access, and Joseph LeDoux summarizes a handful of these studies,[110] some showcasing how individuals can be unaware of their own emotional states and others how individuals can misremember what they were thinking and feeling at the time of an experience (LeDoux 2002, 202–3). Does this weigh in favor of the Behaviorist Strategy? Not quite. LeDoux argues that the revolution of cognitive science has convinced us that mental states matter, and though they may be unreliable, we can learn through clever experiments *how* they are unreliable. This ultimately provides invaluable insight into how the mind processes information (203–4).

On this score, then, Aristotle is correct. If we want to understand why akrasia occurs, we need to try to understand the relationship between experience, thought, and action. If people *claim* to know what is best, then we must investigate whether they *really* do. What might compromise a person's ability to know? Is there anything (e.g. an external stimulus or an internal state) that can influence a person's thought processes? These are all important questions that we need to ask, and they can only be explored by

---

[109] See, for instance, section 3.7 on *confabulations*.

[110] For more on this, he cites the following articles: "Emotion in man and animal: An analysis of the intuitive process of recognition" (Hebb 1946); "Objective happiness" (Kahneman 1999); "Ten years in the life of an expert witness" (Loftus 1986); and "Misinformation and memory: The creation of new memories" (Loftus and Hoffman 1989).

recognizing that there is at least some kind of value to our conscious mental states, even if it is not as we expect.

Against the Meno Strategy, Aristotle believes that an attempt to draw a firm qualitative line between belief and knowledge is unnecessary when it comes to analyzing akrasia, for he sees no reason why strong opinion is any more susceptible to being overcome by nonrational desire than knowledge. He writes:

> Now as for its being true opinion but not knowledge contrary to which people behave without restraint, this makes no difference to the argument, since some people who hold opinions are in no doubt, but think they know with precision. So if it is on account of having only slight belief that people with opinions will act contrary to their conceptions more than will people who know, *knowledge will be no different from opinion, since some people believe in their opinions no less strongly than others believe in the things they know*, as Heracleitus shows. (*Nicomachean Ethics*, 1146^{b}25–31; emphasis added)

It might be the case that if one has any doubts whatsoever about what she believes, her opinion can be more easily overcome by a nonrational desire, but if she holds that opinion firmly and resolutely, it is just as mysterious how it fails in the face of the nonrational as it is how knowledge fails. Thus, Aristotle believes that the problem of akrasia remains regardless of whether we consider it a failure of knowledge or a failure of belief.

Although he rejects both of these strategies, as is characteristically Aristotle, he does believe that both contain a kernel of truth that contributes to our understanding of what is happening during akrasia.

6.6 | What You Can Learn from Bad Behavior

As LeDoux pointed out, verbal self-evaluations are notoriously unreliable, but this also does not entail that we embrace the Behaviorist Strategy's conclusion that mental states

have no effect on our actions. One of the clearest proofs for this comes from the pioneering work of social psychologist Baruch Fischhoff who researches a phenomenon known as *hindsight bias*. Hindsight bias occurs whenever knowledge of the outcome of an event influences how we interpret and understand that event. In such cases, the knowledge from hindsight often leads us to erroneously believe that the outcome had been apparent all along, even to those directly involved.

To test this hypothesis, Fischhoff conducted three experiments. In the first, he divided the subjects into a handful of groups, and each group was given a description of an historical event with which they were unfamiliar. One of the passages used was the following, which comes from historian Llewellyn Woodward's *The Age of Reform*:[111]

> (1) For some years after the arrival of Hastings as governor-general of India, the consolidation of British power involved serious war. (2) The first of these wars took place on the northern frontier of Bengal where the British were faced by the plundering raids of the Gurkas of Nepal. (3) Attempts had been made to stop the raids by an exchange of lands, but the Gurkas would not give up their claims to country under British control, (4) and Hastings decided to deal with them once and for all. (5) The campaign began in November 1814. It was not glorious. (6) The Gurkas were only some 12,000 strong; (7) but they were brave fighters, fighting in territory well-suited to their raiding tactics. (8) The older British commanders were used to war in the plains where the enemy ran away from a resolute attack. (9) In the mountains of Nepal it was not easy even to find the enemy. (10) The troops and transport animals suffered from the extremes of heat and cold, (11) and the officers learned caution only after sharp reverses. (12) Major-General Sir D Octerlony was the one commander to escape from these minor defeats. (Fischhoff 1975 / 2003, 305)

Once the groups had finished reading the passage, they were presented with four possible outcomes and were asked to assign probabilities, totaling 100%, to each outcome:

a. British victory
b. Gurka victory
c. Military stalemate with no peace agreement
d. Military stalemate with peace agreement

---

111 See *The Age of Reform* (Woodward 1938, 383–4).

The twist in this experiment was the following: one group (the *Before* Group) was *not* informed of the outcome during their decision-making, while the four other groups (the *After* Groups) *were* apprised of the historical outcome. Fischhoff was particularly interested in seeing whether and how the After Groups interpreted the information differently from the Before Group. As a second twist in the experiment, the historical outcome presented to each After Group was a *different* one; each group was lead to believe that the outcome was one of the four possibilities above.

So what happened?

Each of the After Groups assigned a significantly higher probability to the outcome that was given to them, as if it were an obvious consequence from the information presented in the passage and they could have predicted it with the same level of confidence prior to knowing the outcome. This kind of retrospective identification of a pattern of inevitability in the data sample has been coined *creeping determinism*. We often perceive such necessity only through the luxury of hindsight.

In addition to creeping determinism, another feature of hindsight bias is a distortion of the information we interpret, causing us to selectively pay attention and zero in on what appears relevant to what we already know. So in this experiment, Fischhoff had observed how knowledge of the outcome changed the way that the subjects interpreted the historical passage above, leading them to believe some details to be of greater importance than others. For example, those who were told that the British were victorious believed it was relevant that the British "learned caution" during their battles in the mountain, while those who were informed that the Gurkas were victorious viewed this information as wholly irrelevant (Fischoff 1975 / 2003, 306).

In a second experiment, similarly designed, Fischhoff asked the participants to make their choices and assign relevance to information as if they were *unaware* of the outcome.[112] Unfortunately, even when instructed, the subjects were unable to ignore their hindsight bias, betraying that they were utterly incapable of viewing the event "objectively" or independently of what they already knew. This lead Fischhoff to conclude that "subjects are either unaware of outcome knowledge having an effect on their perceptions or, if aware, they are unable to ignore or rescind that effect" (Fischoff 1975 / 2003, 309). Apparently, even when we are aware of the possibility for hindsight bias, we are unable to negate its effects on our how we interpret and distort information.

The results from Fischhoff's research into hindsight bias confirms Aristotle's suspicions. While *what* we believe and know can influence our decision-making—evidenced by the fact that knowledge of different outcomes caused subjects to assign their probabilities and deem information relevant in accordance with their "knowledge"—it is *also* true that our self-descriptions of what we know, how, and why are unreliable. The implication this has for akrasia is that we should be highly suspicious of what people verbally report, especially *ex post facto*. Our memories and knowledge of an outcome inevitably distort our beliefs about what actually occurred and what we had *really* known at the time. What the Behaviorist Strategy is right to assume, then, is that "no one acts contrary to what is best *while believing that to be the case*" (*Nicomachean Ethics*, 1145^{b}26–27; emphasis added). It is easy to say what we believed through the lens

[112] The third experiment was likewise similarly designed, except in this one, participants were asked to assign probabilities to outcomes and judge relevance of information as they would expect from *others* who were unfamiliar with the event. Even in this scenario, the results were the same: hindsight bias unavoidably influenced decision-making and interpretation of the event (Fischhoff 1975 / 2003, 309–10).

of hindsight, but in all likelihood, we were entertaining and processing very different beliefs and aspects of the situation at the time.[113]

6.7 | What We Can Learn from Socratic Wisdom

Preserving this behaviorist insight, Aristotle argues that the Meno Strategy is also correct insofar as akrasia appears to result from an *epistemic* failure. To make his case, he begins to outline some distinctions concerning how an individual can be said to know something, commenting briefly about how this factors into an experience of akrasia. He writes:

> And since we speak of knowing in two senses (for both someone who has knowledge but is not using it, and someone who is using it, are said to know), it will make a difference whether one has knowledge but is not attentively considering, or is attentively considering, the things one ought not to do; for this [i.e. acting contrary to knowledge] seems terrible, but is not if one is not attentively considering it. (*Nicomachean Ethics*, 1146^{b}32–36)

This distinction is similar to one commonly used in contemporary epistemology, between *occurrent* and *dispositional* beliefs.

An occurrent belief is one that is being *entertained*, or held in thought, by an individual *here and now*, while a dispositional belief is one that an individual is *not* entertaining at the moment but *could* if the right circumstances arise. It is important to

---

[113] A handful of studies have suggested, for example, that we struggle to take long-term perspectives into consideration, and many of our decisions are influenced by the attractiveness of a reward and its temporal distance (LeBoeuf and Shafir 2005, 255–6). Depending on the context, a smaller reward here and now might appear more desirable to many of us than a larger reward at some later date and time. While fascinating, it is worth pointing out that something like this was also defended long ago by Socrates in Plato's *Protagoras* (356b–358d).

For more on this within the field of psychology, see: *Breakdown of Will* (Ainslie 2001); "Time discounting and time preference: A critical review" (Frederick, Loewenstein, and O'Donoghue 2002); and "Anomalies: Intertemporal choice" (Loewenstein and Thaler 1989).

This is also discussed below in section 6.9.

realize that dispositional beliefs are not potential beliefs but are *already* possessed by an individual and stored in memory; what makes them dispositional is simply that they are not being entertained *here and now*.

For example, while working on her car, Sabrina might have an occurrent belief that her alternator needs replaced, and while she is not currently entertaining her belief that tarantulas are not classified as true spiders, if one were to ask her whether she believed this, she would affirm it—for entomological trivia has always been a hobby of hers. This would make the tarantula belief dispositional insofar as it is a belief that Sabrina already has and *would* entertain in the right context, say, at an arachnid exhibition, even though it is not one that she is thinking about *here and now*.

Since Aristotle conceded that the problem of akrasia remains regardless of whether it is analyzed as a failure of belief or a failure of knowledge, I see no reason why the distinction in knowledge that he proposes and the subsequent concern he raises cannot also be extended to occurrent and dispositional beliefs. If so, his concerns about akrasia can be restated in terms of occurrent and dispositional beliefs, the former making for the more challenging case. Akrasia can thus be redefined along the following lines, each in need of its own separate consideration:

> *Occurrent Akrasia*: An individual is suffering from occurrent akrasia if and only if she currently holds in her mind a belief about what course of action is best and she acts otherwise in order to satisfy some nonrational desire.
>
> *Dispositional Akrasia*: An individual is suffering from dispositional akrasia if and only if she has a belief about what course of action is best, she is not currently entertaining that belief, and she acts otherwise in order to satisfy some nonrational desire.

Inspired by the Behaviorist Strategy, Aristotle denies that occurrent akrasia exists, and, following the Meno Strategy, he believes that dispositional akrasia arises from an epistemic failure analogous to a state of drunkenness.

So what might happen during akrasia then?

Aristotle first begins with the observation that it is possible for people to suffer from a condition that inhibits them from entertaining any beliefs that they possess. Something like this is what occurs during, for example, sleep or a fit of madness (*Nicomachean Ethics*, $1147^{b}13$). Philosopher Amélie Rorty explains that such a person "may have knowledge that he is not using because his condition prevents his doing so even when the occasion is appropriate (as a drunken mathematician is unable to count money to pay for his drink)" (Rorty 1980, 270). Such a mathematician knows how to count, but were he sufficiently intoxicated, he might find that his practical knowledge of counting has been suddenly compromised, rendering him unable to access this information.

Sensitive to the unreliability of verbal reports, Aristotle thus believes that something like this must be happening during akrasia; the akratic person (*akrates*) is like a drunk or an infant learning to speak. He writes:

> It is clear, then, that one ought to speak of unrestrained people as being in a condition similar to [those in a state of intense passion]. And speaking the words that come from knowledge signifies nothing, since people who are in these states of passion recite demonstrations, or verses of Empedocles; those who are first learning something also string words together, but do not yet know anything. (*Nicomachean Ethics*, $1146^{a}18$–25)

The akrates is thus stirred into a condition that prevents her from accessing her knowledge. During this episode, she lacks the control afforded by reason, and while she might even be able to repeat to herself or aloud what she believes is best, doing so

amounts to little more than a mere theatrical performance. Because she lacks the appropriate awareness of either the content of her beliefs, her surroundings, or even how her surroundings relate to her beliefs, she is akin to the drunk who finds himself able to utter lines of Shakespeare without being aware that he is doing it, that such words have any meaning, or that anybody is even listening.

6.8 | Fool Me Once, Fool Me Twice, Fool Me Something Strength All Over Again

So how might one end up in a state of akratic intoxication? Rorty helps interpret Aristotle on this point. She believes there are three different causes for this kind of epistemic failure: impulsive action, misperception, and faulty inference, explaining:

> Sometimes the akrates acts impulsively: he can fail to think about whether the situation before him falls under his general principles about what is good ($1150^{b}19$). Or if he does think about what he is doing, he does not see the particular case properly: he misperceives or misdescribes what is before him. Or even if he gets it right, he can fail to connect it with his general principles, fail to see the import of his knowledge. He then fails to draw the right conclusion about what to do, either making the wrong decision or failing to act from the decision implicit in his beliefs. (Rorty 1980, 273)

Regarding the first possible cause, Aristotle describes impulsive action as the sort of epistemic failure that occurs whenever we neglect to "wait for reason" (*Nicomachean Ethics*, $1150^{b}28$). We may not give ourselves sufficient time to weigh everything,[114] and so we do not give ourselves an opportunity to figure out how to coordinate what we believe is best with our current circumstances, ultimately resulting in undertaking an action that is contrary to our beliefs. While it is tempting to attribute impulsivity solely to

[114] Aristotle has in mind here *deliberation*, which is a technical term for him. I spend time discussing it in section 10.1, where it plays an important role in my account of agency. For present purposes, however, I am temporarily supposing that Aristotle believes that "weighing everything," "evaluating everything," or "reasoning through things," are sufficient to prevent someone from succumbing to akrasia. The goal of this chapter is to understand akrasia rather than deliberation, and a momentary deviance from deliberation eliminates what would otherwise be a distraction from that goal.

some kind of character flaw, making a decision in haste or under intense stress can also have the same consequences for decision-making.[115] As a result, acting impulsively qualifies as dispositional akrasia provided that, had we had the time to think through our situations, we would have acted in accordance with our best interests.

A misperception, on the other hand, occurs whenever we fail to see what is relevant in our environment with respect to how we frame and think about our current situation. Suppose, for instance, that we were to formulate a practical syllogism to plan what we ought to do. In such a scenario, we might reason well but struggle to precisely see how the elements of our situation relate to our premises in an appropriate way. Aristotle explains:

> One may have the [major] premises that dry food is beneficial to every human being and that oneself is a human being, or that such-and-such a food is dry, but whether this *particular* food in front of one is of that kind is a premise one either does not have or is not actively considering. (*Nicomachean Ethics*, 1147ª6–7; emphasis added)

Recasting this passage in the form of a syllogism makes it clearer, showing how we might reason in the following way:

P_1	All dry food is beneficial to human beings
P_2	I am a human being
C_1	All dry food is beneficial to me
P_3	All dry food is beneficial to me
P_4	**This food in front of me is not dry food**
C_2	[I don't know what to do with the food in front of me]

Premise four is false, and so by failing to recognize that the food in front of us is dry, we sadly never draw the right conclusion, namely, that the food in front of us is *also* one that

[115] In "Decision making under stress: Scanning of alternatives under controllable and uncontrollable threats," for instance, psychologist Giora Keinan discovered that participants struggled to consider all of the relevant alternatives in a situation when they felt pressure from the threat of an electric shock (Keinan 1987, 642).

we should eat. Consequently, we fail to act accordingly.[116] Again, if something like this is happening in cases of akrasia, then such cases should be construed as dispositional, for had we properly identified the relevant features in our environment for what they were, we would have drawn the right conclusion.

The final cause of akrasia that Aristotle discusses is faulty inference, which occurs whenever we use the right principles to determine the best course of action but unfortunately make the wrong inference about what to do, due to influence from a nonrational desire. An example of this can be located at lines $1147^{a}29$–$1147^{b}4$, but the passage itself is notoriously unclear.[117] Aristotle writes:

> If one ought to taste everything that is sweet, and this thing here, as a certain one of the particulars, is sweet, it is necessary for someone who is able to and not prevented, to do this at that same time he recognizes it. But when a universal premise is present in someone that prevents tasting it, and another that every sweet thing is pleasant, and this thing here is sweet (and this is at work on him), and a desire happens to be present, then while the one premise says to avoid this, the desire takes the lead, since it is able to set in motion each part of the body. And so, behaving without restraint [i.e. akrasia] results in a certain way from a proposition or opinion, which, while not in itself opposed to right reason, is opposed to it incidentally, since the desire, though not the opinion, is opposed. (*Nicomachean Ethics*, $1147^{a}29$–$1147^{b}4$)

---

[116] Keinan was also aware of this problem when he designed his stress experiment, taking care to control for this. He worried:

> For example, a decision maker could conceivably weigh all alternatives carefully and still reach a disastrous decision simply because some relevant data were unavailable to him or her. Hence, propositions concerning the effects of stress on the consideration of decision alternatives should be evaluated via direct observation rather than through inference. (Keinan 1987, 640)

[117] Rorty, for example, believes that this passage shows that an inferential failure can occur, but she does not explain *how* this happens nor does she acknowledge the presence of *multiple* major premises or desire itself as playing any role in the failure (Rorty 1980, 273).

Joe Sachs, by contrast, acknowledges that there are two major premises, but he does not believe they are *necessarily* in conflict. The first is that "every sweet thing is pleasant" and the second is some opinion that advises against tasting the sweet thing. According to him, the akratic action results on account of some desire to taste sweet things masquerading as a premise (Aristotle 2002, 124, n.187). This too, however, leaves much unexplained. What is the fuller syllogism that causes the akratic action? Why does the syllogism that advises against the action remain incomplete or fail to mobilize someone against akrasia?

The first sentence is an example of a standard practical syllogism with a major premise comprising a principle from which one reasons, a minor premise that derives from sense-perception, and a conclusion that *is* the act of eating the sweet. Aristotle argues that what happens during akrasia is that a desire hijacks the reasoning process, leading one to make the wrong inference regarding what to do. But how might this happen?

Following philosopher David Wiggins' interpretation of this passage,[118] it appears that Aristotle is suggesting that there are two syllogisms in conflict, illustrated by the following:

Syllogism for Reason	
P_1	No sweet thing is healthy
P_2	This is a sweet thing
C	This is not healthy

Syllogism for Appetite	
P_1	All sweet things are pleasant
P_2	This is a sweet thing
C	This is pleasant

On this interpretation, the problem of faulty inference thus begins to look very similar to Davidson's presentation of the strength model, and if so, it suffers from the same unresolved issues. How, after all, does the appetitive syllogism overpower the rational syllogism that advises against tasting the sweet thing? And for that matter, how is one to even understand the relationship between appetite (i.e. nonrational desire) and the

---

[118] Wiggins translates and paraphrases this passage in the following way:

> If one had better eat of anything that is sweet, and the object presented in some specific situation is sweet, then the man who can act and is not physically prevented *must* at the very moment [at which he brings the premises together] act accordingly. So when there is some major premise or other [which combines with some minor premise to] constrain the man from eating of something [when for instance a major premise indicating that Φ things are bad for the health combines with a minor premise that a certain *x* is Φ]; and when there is another practical syllogism in the offing with the major premise that everything sweet is nice to eat and a minor premise that *x* is sweet (this being the premise that is active) and appetite backs this syllogism; then the former syllogism forbids the man to taste but appetite's syllogism pushes him on. So it turns out that a man behaves incontinently under the influence (in some sense) of reason and belief. For he has argued himself to his practical conclusion from true beliefs, and these beliefs are not in themselves inconsistent with reason. It is the appetite itself that opposes reason, not the premises of the appetite's syllogism. (Wiggins 1980b, 248–249; emphasis original)

practical syllogism behind it? Does appetite itself reason? If so, then what is the difference between reason itself and appetitive reason? If not, then how does appetite endorse one of the syllogisms? If appetite is blind, then it is a matter of extraordinary luck that it manages to endorse a syllogism that can satisfy what it wants, but if it is not blind, then how does it "know" or "perceive" which syllogism is which?

Unfortunately, there are no obvious or easy answers to these questions, and they are problems that haunt every theory of action that adopts the strength model and prioritizes the role of reasoning in decision-making and acting. Nonetheless, the problem is not with Aristotle's etiology of akrasia, but with the account of traditional agency it appears to require.

6.9 | Too Weak to Resist

Aristotle's insight that the mind can process information poorly and draw the wrong conclusion in the heat of the moment is both innovative and insightful. For Aristotle, the mind is *vulnerable* to nonrational desires, able to be distracted by what is present *here and now*. As a result, this can cause us to lose connection with what we know and distort what we see. How might this occur?

Suppose someone were to approach you with the following options:[119]

Vacation Location A	Vacation Location B
average weather	lots of sunshine
average beaches	gorgeous beaches and coral reefs
medium-quality hotel	ultra-modern hotel
medium-temperature weather	very cold water
average nightlife	very strong winds
	no nightlife

[119] This example comes from "Reason-based choice" (Shafir, Simonson, and Tversky 1993, 17).

Now suppose this person were to ask you: "If you had to *cancel* your vacation to one of these locations, which would you cancel?"

If choosing which vacation to cancel is a matter of reasoning alone, then the choice should be straightforward. We would expect that whichever vacation you choose to cancel logically implies that the other vacation is more desirable. In logic, this is known as a *disjunctive syllogism*. We could even explore whether this is true by posing the question differently. For example, we might instead ask: "If you had to *accept* a vacation to *one* of these locations, which would you accept?"

Yet, it may come as a surprise to learn that behavioral psychologist Eldar Shafir discovered that, depending on how the question was framed (i.e. in terms of acceptance or in terms of rejection), the responses *changed*. In *both* cases, the majority of respondents selected location B. Why?

Shafir and marketing psychologist Robyn LeBoeuf explain that a person will shift how she views information and how she reasons depending on her disposition to respond to the situation, which is a phenomenon known as *response compatibility* (LeBoeuf and Shafir 2005, 253). In other words, *wanting* something causes us to think very differently about an impending choice than *not wanting* something; we will highlight and focus on differing sets of attributes, depending on our response disposition. They elaborate, "Because positive features are weighed more heavily in choice and negative features matter relatively more during rejection, the enriched destination was most frequently chosen *and* rejected" (253; emphasis added).

Other nonrational factors that have influenced decision-making include temporal distance and the size of the reward (LeBoeuf and Shafir 2005, 256). Economist Richard Thaler, for instance, found that while some people preferred a choice of one apple today as opposed to two apples tomorrow, almost nobody preferred one apple in 365 days compared to two apples in 366 days (Thaler 1981, 202). This was especially interesting for him because, from a purely economic perspective, the two sets of choices are identical with the exception of the time frame (202). When there is a chance to have a reward *now*, it becomes far more attractive to take it.

Three other important factors that influence decision-making are self-identities, emotional states, and drives. LeBoeuf and Shafir, summarizing the work of social psychologist John Turner, describe how a woman might see herself as a mother around her children but as a professional in the workplace, and as it turns out, her identity *as* a mother inclines her to use a very different set of values from her identity *as* a CEO (LeBoeuf and Shafir 2005, 257). Not only is there a clash in values between the two senses of identity (e.g. a mother prioritizes her family while a CEO prioritizes her company), such clashes can create the conditions for psychological conflict, "as when a parent commits to a late work meeting only to regret missing her child's soccer game once back at home" (258). A different study, conducted by social psychologists Jennifer Lerner and Dacher Keltner,[120] shows how a state of anger can encourage individuals to take risks while a state of fear "promotes risk aversion" (258). And yet another study—this one from social psychologist Leaf van Boven and economist George Loewenstein—

[120] See "Fear, anger, and risk" (Lerner and Keltner 2001).

reveals how a person's drives can influence decision-making.[121] These researchers discovered that when an individual was currently experiencing a state of thirst, she was far more likely to be concerned with thirst in a hypothetical scenario as well. In fact, 92% of participants believed they would be bothered more by thirst than hunger if they were hypothetically trapped in the wilderness compared to just 61% of those who were *not* thirsty at the time of responding (258).

Many of these considerations lead LeBoeuf and Shafir to conclude:

> Inconsistency thus often arises because people do not realize that their preferences are being momentarily altered by situationally induced sentiments. Evidence suggests, however, that even when people are aware of being in the grip of a transient drive or emotion, they may not be able to "correct" adequately for that influence. (258)

Now, according to Rorty, similarly (and perhaps surprisingly) Aristotle believes that an akrates "temporarily forgets his knowledge of what is good because he has put himself in a situation and in a condition in which his perceptions of pleasure are so affected that he acts from his reactions (*pathe*) rather than from his knowledge" (Rorty 1980, 272). So what happens in the case of faulty inference is that the *presence* of something sweet causes us to react to it, inducing akratic drunkenness. This, in turn, disrupts our connection with knowledge and effectively creates a state of transient ignorance. As one might expect from dispositional akrasia, had the sweet treat *not* been present, we likely could have drawn the right conclusion.

This akratic condition also explains how misperception might occur from something other than a simple mistake or inability to coordinate one's premises with one's situation. Rorty explains:

---

[121] See "Social projection of transient drive states" (van Boven and Loewenstein 2003).

> It is the manner of his reactions to pleasures that misleads the akrates: he acts from his reactions to what is before him, perceiving—misperceiving—what he does in terms of its pleasurable effects on him rather than seeing his situation, and his actions in it, as defined by his proper intentional ends. (Rorty 1980, 277)

In other words, if we are vulnerable to pleasure and something pleasant is immediately before us, we might be inclined to look at this object *as* pleasant, much as how we might look at Vacation Location B while ignoring any other features that undermine or compromise its desirability. For example, you may want something sweet and see a high-calorie dessert before you. Even though you may ordinarily reason against partaking in dessert, this situation might cause you to fail to see it *as* a high-calorie dessert, blinded by a nonrational desire for pleasure that was stirred up from its presence.

6.10 | Conclusion

The question remains as to how akrasia counts as voluntary, for if the akrates *would have* done otherwise had the circumstances been different, was she not acting compulsively, *in*voluntarily?

In Homer's *Odyssey*, the goddess Circe warns the clever hero Odysseus that his journey will take him past the Sirens, whose voices and songs can "spellbind any man alive" (*Odyssey*, XII.46). She strongly advises that he sail past them as quickly as possible, taking care to ensure that his crew place beeswax in their ears before reaching their location. "But," she says, "if *you* are bent on hearing, have them tie your hand and foot in the swift ship, erect at the mast-block, lashed by ropes to the mast so you can hear the Sirens' song to your heart's content" (XII.55–58; emphasis original).

There is always wisdom to be had in the words of Homer, for it is as if the Sirens function as the perfect metaphor for those situations that can place us in akratic

conditions. But more than that, Homer also offers advice through Circe: if we cannot avoid the situation, it would be prudent to brace ourselves in advance. If reason is to defeat akrasia, it is not on its own terms; it has to be clever like Odysseus and rig the fight before it happens. Reason is stronger by virtue of its wit and ingenuity, not its brute strength. Just as we would not expect David to best Goliath in an arm-wrestling match without a clever plan, reason too needs cleverness in its battle against desire.

Along similar lines, Rorty recalls Aristotle's account of voluntary action in *Nicomachean Ethics* Book III, arguing that even though we might lack control *in the moment*, it counts as voluntary insofar as we are responsible for having got ourselves in the situation. She writes, "If the person is responsible for having got himself in the condition where he cannot use knowledge that he has, he is responsible for what he does in that condition because he is responsible for his ignorance ($1110^{b}9$ ff.)" (Rorty 1980, 270). Whether it is that Aristotle calls an akratic condition—or what social psychologists might call nonrational influence—it is incumbent upon us to preemptively act against it, *before* we put ourselves in circumstances where it can affect us. Like one who suffers from alcoholism, the akrates must at least try to develop a self-understanding that includes an awareness of her condition, her biases, and her psychological tendencies.

Still, although an akrates acts from ignorance *in the moment of acting akratically*, it is her voluntary choice to put herself into that position *in the first place* that is especially curious. While this may happen on occasion or accidentally (as when one is surprised with a dessert from a friend even though one may be trying to avoid them), there are cases where people *consistently* put themselves into a position where they will experience the vulnerability that accompanies the akratic condition. This in itself is a kind

of akrasia, for one is voluntarily choosing to put herself in a situation where she *will* act contrary to her best interests—a choice that itself is contrary to her best interests—much like the alcoholic who decides earlier in the day to visit the bar anyway, knowing what that entails. More problematic still, this appears to be *occurrent* akrasia, the kind whose existence even Aristotle denies.

Now, in both types of akrasia, occurrent and dispositional, the end result is that we act in a manner that contradicts our best interests. Aristotle and contemporary psychological research suggests to us that akrasia occurs as a result of nonrational desires influencing, disrupting, or even overpowering rational thought processes. It is unclear how this can be adequately accounted for on a traditional theory of action, which usually adopts the Enlightenment strategy of denying the existence of akrasia altogether or refusing to entertain the idea that reason can be overpowered by the nonrational, for such an admission would undermine the idea that we are fundamentally rational. As Shafir puts it after summarizing the research:

> By this account, it may help to think of individual decision makers not as faulty economic agents, but as fundamentally different creatures. Creatures who are, to be sure, interested in improving their lot and who have preferences, but who, nonetheless, are fundamentally different processors of information from those envisioned by classical analyses. (Shafir 2003, 23)

The question remains as to how the mind is actually working if it so seldom operates rationally, traditionally understood. Answering that would also account for how these irrational phenomena seem to arise, but before addressing that topic, there is one more irrational phenomenon that is worth considering: *negation*.

Chapter 7: A Contradiction in Action

> We might ratify these translations
> by showing that his nonlinguistic ways of handling himself and others showed
> that he actually did hold such paradoxical beliefs.
> The only way to show that this suggestion cannot work,
> would be actually to tell the whole story about this hypothetical stranger.
> It might be that a story could be told
> to show the coherence of these false beliefs with each other and with actions,
> or it might not.
>
> — Richard Rorty, "The World Well Lost"

7.1 | A Moderately Severe Case of Perverse Logic

On October 1st of 1907, a bright, young male patient came to see Sigmund Freud for *Zwangsneurose* (obsessive-compulsive neurosis),[122] a condition that Freud would describe as "moderately severe." It is a disease he believed to be characterized by repetitive, uneasy thoughts that incline an individual to undertake actions that she does not want to perform, such as harming or killing oneself (Freud 1909 / 2002, 125–26). In this case, the young man, whom I shall call Mr. R,[123] found himself suffering from the following: a persistent fear that something tragic would happen to his loved ones; obsessive thoughts that are undesirable; and, in response to these thoughts, urges to punish himself with a number of prohibitions and restrictions (128; 151). Mr. R experienced each of these recurring psychological states—fear, obsession, prohibition—

[122] Psychoanalyst Jean Laplanche and philosopher Jean-Bertrand Pontalis credit Freud with being the first to isolate this disorder, even though our understanding of it has evolved since his identification. Prior to Freud, various symptoms were recognized but often treated as belonging to other, independent disease-states, but he was able to identify that there was an underlying unity to the different symptoms (Laplanche and Pontalis 1967, 281–282).

[123] The patient is usually referred to as "Ratman," named after a particular fear of his, but I have chosen to follow the convention of Jonathan Lear who instead opts for the alternative title out of respect (Lear 2005, 12).

as if they were compulsions, anxiously worrying that he had very little control. It was as if he were a mere spectator of intrusive thoughts that were not his own.

During analysis, Freud recognized this compulsive element and how it adversely affected so many aspects of Mr. R's life, even causing him to pursue actions that were sudden, extreme, or confusing. What Freud found particularly amusing was the *repetitive* component, for even though the compulsion hardly made any sense to Mr. R as he was experiencing it, to Freud, who listened carefully, there was a perverse logic. He explains:

> It is a well-known fact that compulsive ideas appear to be either without motivation or without sense, just like the phrasing of our nightly dreams, and our first task is to establish their foothold and meaning in the inner life of the individual so that their purpose is evident, self-evident in fact. One should never be misled in the exercise of translation by an appearance of insolubility; the wildest and most peculiar compulsive ideas can be solved when one is properly absorbed in the task. One arrives at such a conclusion, however, by bringing the compulsive ideas into temporal connection with the patient's experiences, that is, by examining when an individual compulsive idea first appeared and under what external circumstances it tends to be repeated. (Freud 1909 / 2002, 149)

Rather than treat the ideas as if they were purely psychological or even simply random, Freud instead sought the *cause* of the compulsion in lived experience. It was not enough to learn *that* Mr. R had obsessive thoughts; Freud wanted to know *when* they occurred, *what* was happening at the time of the occurrence, and any other potentially relevant details that filled in the context. His ingenuity in this respect was recognizing the continuity between the psychological and the world, as the person experiences it, believing resolutely that it was not possible to make sense of one without reference to the other.

With respect to Mr. R, Freud concluded that the common cause for a number of his compulsive activities had to do with his feelings toward his beloved, for so many events seemed to involve her either directly or indirectly. For instance, Freud noticed that

during a summer spent with his beloved, the young man developed an obsession to lose weight, a compulsive desire that she should wear his hat, and a compulsion to count between lightning strikes and thunderclaps when the two were together (Freud 1909 / 2002, 151). While the phenomenon of compulsion itself is fascinating enough to warrant its own treatment within a general analysis of irrationality—as it raises questions about an agent's freedom and whether a compulsive behavior ought to count as an action—of particular interest in this study is what Mr. R did during this same summer, on what just so happened to be the last day he would see his beloved for the season.

7.2 | Rationalizing the Compulsion to Throw Stones

As it turns out, Freud had learned about a peculiar event that had taken place soon after Mr. R's beloved had departed. While Mr. R was walking along a road, he happened to catch his foot on a stone. Immediately, a thought intruded that he *had* to pick up the stone and move it out of the way, and his justification for this was that his love interest might be traveling along this same road in her carriage within a few hours. "If the stone remains in the road," he reasoned, "then it could risk damaging her carriage." And so, as one might expect who shares in similar reasoning, Mr. R decided to move the stone (Freud 1909 / 2002, 151–2).

A few minutes had passed after Mr. R had completed his task when it dawned on him that what he had done was absurd. Though Freud does not reveal in more detail exactly what crossed Mr. R's mind, it is not hard to imagine a train of thought that could stir up self-doubt:

> *Why* would a stone even damage her carriage? *Will* she even be passing through here? What if her carriage slows to stop on the side of the road, and my moving of

> the stone is what *causes* her to hit it? Wasn't it *easier* to see in the road? Won't she think I'm silly for worrying *this much*?

As a result of this increased anxiety, Mr. R felt like he *had* to walk back, find the stone, and place it back on the road where he had found it (Freud 1909 / 2002, 152).

At first glance, these seem like two separate events, and one may wonder whether they even count as actions. If, for instance, compulsion was at work in producing this behavior, does Mr. R meet the minimum requirement of doing something free from external compulsion?

Let us first consider how traditional action theory might construe this as an action.

There are certainly theoretical approaches available that could support the idea. A straightforward, though unsatisfactory, solution would maintain that while Mr. R indeed acted from a sense of compulsion, the source of it was *internal*, and as a result, it meets this minimum requirement for an action. Yet such an approach seems to dilute the significance of acting free from external compulsion if it can be proven that Mr. R acted *necessarily* in these cases, as if someone else had guided his hand. Some might make the counterargument that external compulsion is whatever is external *to the agent*,[124] but if one assumes that reason is the true agent, it would follow that this behavior fails to be caused in the right way to be counted as an action, for it did not proceed from reason but something alien to it.

Rather than try to settle these theoretical issues, there is a traditional action theorist who adopts a different approach: Donald Davidson. Using his theory of intentionality, he argues that these kinds of events can be viewed as actions as long as

[124] Agent in this sense is usually understood as a *metaphysical* agent, not simply the body or brain of the person acting.

they can be *rationalized*.[125] Recall that a rationalization, for Davidson, is a way of describing an event that makes reference to the reasons for which an agent acted by reconstructing how things might have appeared from the agent's point-of-view (Davidson 1963, 3). Such descriptions are a form of causal explanation because they explain *why* an agent would undertake a particular performance in the first place (3). An event can be described in a variety of different ways, but all that is needed for it to count as an action is for us to be able to find the description that rationalizes it.

Consider Davidson's classic example that helps illustrate his use of rationalization. Here is the event to be analyzed:

> I flip the switch, turn on the light, and illuminate the room. Unbeknownst to me I also alert a prowler to the fact that I am home. Here I need not have done four things, but only one, of which four descriptions have been given. I flipped the switch because I wanted to turn on the light and by saying I wanted to turn on the light I explain (give my reason for, rationalize) the flipping. (Davidson 1963, 4–5)

We can describe this event in the following four ways:

(1) There was a flipping of the switch.
(2) There was a turning on of the light.
(3) There was an illuminating of the room.
(4) There was an alerting of the prowler.

For any of these descriptions, we might ask, "Why did this occur?," and we can provide an answer to each. The illuminating of the room occurred, for example, because there was a turning on of the light. Likewise, the alerting of the prowler occurred because there was an illuminating of the room. Each of these seem like very reasonable causal explanations for what happened. And yet, it would not make any sense to say that an alerting of the prowler occurred because the room *wanted* to illuminate. The room, as far as we know, does not and cannot have any intentions.

[125] For more on this, see section 2.2, where this idea was first introduced.

However, a causal explanation that makes use of intentions is precisely what works were we to say that a flipping of the switch occurred because I *wanted* to illuminate the room and I *believed* that turning on the light would do just that. By attributing desires and reasons to me, that is, by highlighting my intention, the rationalization relates me *as* an agent to my performance *as* an action. The flipping of the switch is something that I *caused* to happen *because of* my reasons for doing it (Davidson 1963, 3).

Can it be said of Mr. R that he wanted something from the initial moving of the stone and, if so, can this explain why it was done? On both counts, it appears that it can. Freud says of Mr. R that when he desired that his beloved should wear his hat during a windy day, it was because he felt moved by a sense of duty that "*nothing should happen to her*" (Freud 1909 / 2002, 151; emphasis original). Freud described this as a *protective compulsion*, and he believed that this same imperative motivated Mr. R's other behaviors throughout that summer, including the stone-moving incident (151). Whether this kind of compulsion rises to a level such that it leaves an agent without a choice to do otherwise is unclear, but for Davidson's purposes, as long as this can be re-described as a strong *desire* to protect, it may be just enough to rationalize Mr. R's behavior. Thus, in order for Davidson's explanatory strategy to work, a relationship needs to be established between compulsions and desires.

7.3 | The Sound (or Feeling) of Inevitability

Is it possible that compulsion is a peculiar form of *want*?

To make the case for this, it is important to keep in mind that there are different kinds of wants and not all of them are associated with pleasure. Wanting to eat a dessert,

for instance, and wanting to keep a promise in spite of the fact that it would be far more advantageous to break it are accompanied by some distinct, dissimilar phenomenological features. In the former case, one might experience a ravenous lust, anticipating the pleasure that will surely follow from savoring the dessert; whereas in the latter case, one might experience anxiety, frustration, and disappointment, knowing that keeping the promise will only make things far more complicated for oneself or others.

Philosopher Lennart Nordenfelt, who uses a more traditional theory of action, believes that internal compulsion is more akin to the second kind of want, arguing that an agent feels compelled to do something when it appears unavoidable in the light of what she believes (Nordenfelt 2007, 143).[126] He illustrates this with the following example:

> *A* is a private in the Swedish army during wartime. *A* is ordered by his officer to shoot at an enemy platoon approaching his position. *A* knows that if he does not shoot, he will be court-martialed and in the end be sentenced to death. *A*, however, intends to survive. Hence he will shoot at the enemy. (142)

Even if the soldier does not want to kill another person, he *feels* as though he has no choice when he considers his situation, his intention to live, and the limited options available along with their consequences. For Nordenfelt, a subject's intention is what limits her range of available courses of action, and the more one of those intentions restricts that range, the more the recommended course of action will seem unavoidable (151–52).

Consider a variation on another example of his: if Nordenfelt were to intend to have lunch today, then he might consider locations A, B, or C as possible options to

---

[126] It is important to mention that Nordenfelt makes a distinction between compulsion and force. What I have referred to as "external compulsion" in my analysis of action throughout this project is what he understands as force. A subject is forced when the behavior is causally necessary and no intention directs it, such as when a person remains in prison (Nordenfelt 2007, 150).

enjoy a meal, but if his intention were to have a cup of coffee, then he will discover that his options suddenly shrink to just C.[127] In a sense, he will feel compelled to visit C for lunch, provided he was strongly committed to his intention to have coffee. Thus, Nordenfelt argues, compulsions follow from intentions, and although an agent has more freedom in the act of formulating her intentions, this process is ultimately constrained by what she believes. Likewise, what one believes also has an interesting consequence for the phenomenon of compulsion itself.

In his analysis of compulsion, Nordenfelt insists that there are varying degrees, "from total compulsion to a low degree of compulsion," and this degree is influenced by the relevant beliefs an individual has that inform her intentions (Nordenfelt 2007, 142; 150). If, for instance, I intend to stop by the bank soon and I am *certain* that a branch of my bank is around the corner and that there are no other opportune times for going, then I will feel something closer to total compulsion to turn the corner, especially if I am committed to depositing a check today in order to pay bills. On the other hand, if I find myself strongly doubting that the bank is around the corner and I believe that depositing a check early tomorrow will suffice, then while I might feel a pull to investigate—after all, I still have an intention to stop by the bank soon—it may not be strong enough to overcome my desire to return to work on time. Still, although these beliefs influence the degree of compulsion experienced, they are not synonymous with the phenomenon itself.

What then is a compulsion? Might it be some kind of desire after all?

---

[127] This is not quite Nordenfelt's own example but is inspired by it. He instead considers what were to happen if option C were closed for the day, mentioning nothing about coffee (Nordenfelt 2007, 149).

Like desires, compulsions can issue in actions but they need not necessarily, especially if they are weak. When they do terminate in an action, it was because an intention was involved. Drawing upon Nordenfelt's theory that intentions restrict our range of choices, a compulsion can be characterized as a desire to act in accordance with a goal. It is the goal itself that limits our options, and the strength of one's commitment to it is proportionate to the strength of the compulsion that one experiences.

But what happens when people have a compulsion directed towards something undesirable? Is it possible to desire something bad? If not, then how can compulsion be a form of desire?

7.4 | It's All Relative to What You Want

The Socratic Thesis, mentioned briefly in the previous chapter,[128] is the thesis that a desire is always a desire *for* something that an agent sees as good. Why then do people desire bad things? According to this thesis, we only desire bad things because, Socrates explains, we *mistake* them for something good; such things only *appeared* good to us at the time of acting (*Meno*, 77e).

For example, there is a well-known, verified anecdote that Ozzy Osbourne once removed the head of a living mammal, a bat, during one of his concerts by biting into it. Ozzy himself claims that the reason this occurred was because he had mistaken it for a rubber toy. Now, in most circumstances, biting into a rubber toy is seldom viewed as a good thing by perhaps anyone but toddlers, but within the context of performing at a rock concert where the goal is to entertain the crowd, one can see how it might be reasonable

[128] See section 6.2.

to desire to bite into a rubber bat for the sake of entertainment. In this case, however, Ozzy's failure to recognize the bat as a *living* bat leads him to desire and hence act on something that was in fact undesirable.

If misjudging and making evaluative errors were not bad enough though, an additional wrinkle to consider is that desires can be taken in two very different senses: *absolute* or *relative*.[129] An absolute desire is a desire for something that appears good *in and of itself*. If, for instance, Leo wakes up in the middle of the night and has an absolute desire for some chocolate cake, then his desire is *for* the particular sweetness and flavor of chocolate cake, nothing else. He sees the chocolate cake as something good, and he intends to have it as soon as he walks into the kitchen and opens his refrigerator.

A relative desire, on the other hand, is more complex. It is a desire for something that appears *better* relative to the current alternatives.[130] In this case, for instance, suppose that Leo is now very ill with a bacterial infection (obviously from some bad chocolate cake). The medicine that he has to take is so bitter that it induces a gag reflex, causing him to feel miserable. Regardless, Leo intentionally takes his medicine. It is this intention

---

[129] Below, I reference an article where Davidson makes a similar distinction while discussing a different topic.

[130] Davidson too seems to have made a similar distinction. In discussing decision theory, he writes:

> Desire is taken in its fundamental form as a relation between three things: an agent, and two alternatives, one of which is desired more strongly than the other by the agent. Desire thought of in this way is more fundamental than simple non-relative desire, since it is often far clearer that one course of action or state of affairs is preferable to another than that either is desirable. (Davidson 1984, 26)

While he appears to endorse a reductionist account of desire, arguing that all desire is relative in some way, it is clear that absolute desires can be construed as a special type of relative desire, for they still involve two alternatives for the agent: presence and absence, or *having* and *not-having*. Davidson might argue, for instance, that Leo's absolute desire for chocolate cake is a special kind of relative desire wherein he wants the cake relative to its *absence*.

to take the medicine that conveys the fact that, to some degree, he desires to do this, and yet, because the medicine causes him to feel miserable, it can be interpreted as bad under at least one description. What makes the medicine good? From Leo's perspective, *relative to not taking it*, the medicine is good because it will cause him to feel better more quickly (and possibly save his life), but it is not something he would desire on just any occasion.[131]

Like the soldier compelled to kill mentioned in the previous section, depending on the circumstances in which we find ourselves, it is thus very possible to have a desire for something that we would otherwise consider undesirable. For example:

> The soldier does not *absolutely* desire to kill, but in this particular set of circumstances, relative to the alternatives, killing is desirable.[132]

What, then, is the difference between a compulsion and a desire?

As Nordenfelt explains, compulsion has to do with *the feeling of unavoidability* (Nordenfelt 2007, 141). When desire is coupled with *that*, it completes its transformation into compulsion, and that feeling is a function of one's intention and the relevant available beliefs about one's circumstances. Thus, when it comes to compulsion, to properly understand it as a desire we must look to how environmental context and

---

[131] Moral dilemmas also illustrate how relative desires work. Consider, for example, a situation in which there is a high probability that a pregnant mother will lose her own life if she carries her child to term. The apparent options are that she can risk losing her own life or lose the life of her child. Clearly, neither of these options is desirable, but she will intentionally choose one of them, coming to the conclusion that, relative to the alternative, one of these is good in some sense.

[132] Nordenfelt himself seems to have something like this in mind when he discusses compulsion in terms of higher-priority desires. He writes, "A desire may be considered strong simply in the sense that it is given first priority in the agent's mind," and from this, he reasons, "Hunger is normally a stronger desire for a man *A* than his desire to take a walk, also in the sense that if he has to choose between satisfying his hunger and his desire to take a walk, he will choose the former," demonstrating how desires can be relative rather than absolute (Nordenfelt 2007, 151). He does not, however, restrict his theory of compulsion to relative desires alone.

physiological states in addition to our commitment to a goal can shape and constrain our decision-making, imparting a sense of unavoidability.

When it comes to desires and their strength, context matters.

7.5 | In a Manner of Speaking, It Makes Sense

If compulsion is some kind of desire paired with the feeling of unavoidability, then Mr. R's compulsion to protect his beloved can indeed be construed as a kind of strong desire, which is half of what Davidson needs to rationalize the behavior. What is left to explain is *why*, in terms of the particulars, Mr. R ultimately does what he does. What beliefs inform his intention?

Freud writes that the reasons for removing the stone from the road was that Mr. R believed that his beloved's carriage would be passing through shortly *and* that her carriage might be damaged by it (Freud 1909 / 2002, 151–52). These reasons, taken in conjunction with his protective compulsion, enable one to construct a Davidsonian rationalization of this behavior, such as the following:

> In light of his goal to protect his beloved and his belief that her carriage would be passing by soon, Mr. R felt an unavoidable desire to move the stone on the road because he believed that moving it was a way to protect her.

Although this is a minimal description of the event, it is enough to make the moving of the stone appear reasonable from Mr. R's perspective, and because it is a behavior that follows from his beliefs and desires in light of his intention to protect her, it indeed counts as an action for someone like Davidson.

So what about Mr. R's determination moments later to return the stone back to the road?

This behavior is a little more complicated to defend, for it does not appear to be reasonable provided that one assumes Mr. R is still guided in this performance by his intention to protect his beloved, the same intention that motivated him to move it in the first place. The details of this incident are sparser, for all that Freud says is that Mr. R concluded that his original idea was nonsense and that he therefore "*had* to go back and restore the stone to its original position in the middle of the road" (Freud 1909 / 2002, 152; emphasis original).

In what way was his original idea nonsense? Was it nonsense that she would be passing by soon or that the stone could damage her carriage? Maybe it was nonsense to try to protect her in this way, if not in general? This is unclear, but even if we were to have the answer, it still leaves us with the question why Mr. R feels a compulsion to do anything about the stone *a second time*. Could he not have realized the absurdity in his idea *without* feeling the need to turn around and replace the stone he had moved? If it was a compulsion, what was the goal that framed Mr. R's decision-making, causing him to feel that he *needed* to go back?

Fortunately, there is a little more to Mr. R's psychological profile that can help answer this.

In addition to his desire to protect his beloved, Freud also detected a desire to do violence to her, a form of rage. Although this rage had deeper roots in his psychological history, Mr. R would channel it and direct it towards his beloved as well. Motivating this, says Freud, were obsessive doubts that she actually reciprocated interest in him, doubts as to "whether he is justified in taking her words as proof of her tender affection for him" (Freud 1909 / 2002, 153).

Taking then this intention to do violence as a new goal, it begins to make sense that Mr. R would feel the need to locate the stone he had moved. It was not simply a matter of correcting a foolish decision; he *hoped* that it would harm her. To see it as an action, like the initial performance, a rationalization could take the following form:

> In light of his goal to harm his beloved and his belief that her carriage would be passing by soon, Mr. R felt an unavoidable desire to replace the stone he had moved because he believed that restoring it was a way to hurt her.

It thus appears that both of these events can indeed be described as an action performed by Mr. R, as something that it makes sense for him to do from his perspective (i.e. in light of his beliefs, desires, and intentions). Perhaps surprising then is that Freud does not believe Mr. R performed two actions at all; he believes it was *one*! How is this possible?

7.6 | Somewhere There is Unity in Discontinuity

One topic that has not received nearly enough attention within event ontology and action theory is the idea that there are *discontinuous events*, single events that are spatially or temporally fractured. Davidson makes reference to these briefly when he discusses the possibility of *recurring events*—the idea that one and the same event can happen more than once—but a discontinuous event is not a species of recurring event.

In his essay, "Events as Particulars," Davidson attempts to make sense of recurrence by analyzing a dropping-of-a-saucer on two different nights, one after the other. The philosopher Roderick Chisholm had proposed that we analyze these seemingly repeating events as distinct instantiations of a singular, universal event of saucer-dropping, that is, each time Davidson dropped his saucer, shattering it into a hundred pieces, the event of saucer-dropping occurred again and again.[133]

[133] For more on his theory of events and recurrence, see "Events and Propositions" (Chisholm 1970).

Davidson rejects this position because it undermines his own view that all events are unique particulars, happening and occurring *only* once. For him, if there really is an event of saucer-dropping, it can only happen once. How, then, could it happen across multiple nights? Rejecting Chisholm's idea that some universal event of saucer-dropping is essentially happening multiple times and expressing itself in many different ways, he proposes that we look at recurrence this way:

> Events have parts that are events, and the parts may be discontinuous temporally or spatially (think of a chess tournament, an argument, a war). Thus the *sum* of all my droppings of saucers of mud is a particular event, one of whose parts (which was a dropping of a saucer of mud by me) occurred last night; another such part occurred tonight. (Davidson 1970, 183–84)

Thus, every dropping of the saucer, no matter from whom or where, is a part, making up the whole, singular event of *saucer-dropping*.

Though this would not be Davidson's final thoughts on recurrence, it does bring one's attention to the intriguing possibility that one event, and by extension one action, can be divided over space and time. It is a thought that perhaps even crossed Davidson's own mind, for he casually writes only a few lines later, as if it were obvious:

> A meeting can reconvene in another place, a play can continue after an intermission, a floating dice game can spring up again and again, and with new members. In these cases, we can talk of the same event *continuing*, perhaps after a pause. (Davidson 1970, 184; emphasis original)

Interestingly, many of the examples that he gives of a discontinuous event *are* actions.

To appreciate this insight, consider for a moment that the playing of a game can be quite involved and lengthy, especially a game like *Monopoly*® or *Risk*® or *Dungeons & Dragons*®. It is not uncommon for these kinds of games to take multiple get-togethers to reach completion. Resuming play from a previous night need not even occur at the same location, and yet, the playing of the game only makes sense when it is viewed as a

unity. The amount of possessions that one holds, the choices that one has made, the alliances that one has forged, and the enemies that one has crossed, these are aspects of the game best understood when considered this way. Not only is it the case that the players involved do not typically view this as multiple, distinct games taking place, but analytically speaking, if there is not a single event that unites it, it is not clear how to make sense of temporally disrupted, discontinuous game-playing.

Without treating such game-playings as unified events, we would be left with something like the following analysis to describe what is taking place:

> There was an event of a playing of *Risk*® that occurred on the first night, and there was a different but similar event that occurred on the following night.

An analysis such as this would have to make coincidence part of the description to account for the similarity and continuity between the two nights; whereas analyzing it as a single event avoids this complication, elegantly accounting for what has taken place.

Supposing then that there are such discontinuous actions, what justification does Freud give for analyzing Mr. R's two performances as a single action?

Recall that when it came to Mr. R's second performance (the replacing of the stone), in order to rationalize it, we had to *infer* that the reason for it was to cause harm to his beloved. Unlike the first performance, there is no explicit reason in Mr. R's own terms for why he decided to replace the stone. Furthermore, precisely because such explicit reasons are provided for the first performance but not the second, on what grounds can the two possibly be regarded as related, let alone united?

7.7 | Untying the Not

One important consequence of being human is that we struggle with self-understanding. There are times where we can supply reasons for undertaking performances, but there are also times where we find ourselves at a loss for words and unable to account for our actions. Wishful thinking, self-deception, and akrasia demonstrate just how difficult understanding ourselves can be, and in some of those cases, it is not uncommon—indeed, may even be the norm—to resort to an attempt to rationalize our behavior *ex post facto*. This might involve trying to persuade ourselves that something we fear can in no way be true, as we interpret the evidence in a biased manner for reassurance. Other times, we might reflect on what we have done to try to reconstruct *why* we did what we did, finding reasons that were not necessarily present to our minds at the time we acted. For instance, a close friend might inquire as to why Chip seems depressed, and after hearing this, even though he did not believe that he was, Chip may begin to consider that he really is depressed, seeking out reasons for possibly feeling that way.

Freud believed that these *ex post facto* rationalizations were especially common in cases like Mr. R's moving and replacing of the stone, explaining that this kind of behavior is "of course misunderstood in the conscious thought processes of the patient and given a secondary motivation—i.e. *rationalized*" (Freud 1909 / 2002, 153; emphasis original). The reason these actions in particular are misunderstood has to do with their peculiar nature, something not easily noticed even by the person performing the action. Freud deduces this by making a fascinating and careful observation.

Regarding Mr. R's second performance, Freud astutely picks up on the fact that not only did the young man *replace* the stone, he "restored the stone *to its original position* in the middle of the road" (Freud 1909 / 2002, 152; emphasis added). Given this

effort to *precisely* place the stone back where it was found, Freud believes that at some level Mr. R was effectively trying to *erase* his first performance, as if it had never happened. After all, he just as easily could have thrown the stone further away or kept it. He could have even just placed it *anywhere* along the road, anywhere but where it was. Why, specifically, did he feel compelled to recreate the original state of affairs before his intervention?

Recall that Freud's rationalization of Mr. R's second performance was that he was motivated by a destructive compulsion, a desire to do harm to his beloved as a result of his doubt that she reciprocated his feelings for her. It would be rather peculiar to see this behavior as related to the moving of the stone in the first place since, by Freud's own admission, they proceed from two very different compulsions, protection and destruction. What unites them from Freud's clinical point-of-view is the fact that they are in conflict with one another, defined by and through each other, much like a war is a higher unity of two opposing sides, they also defined by their conflict with one another.

Freud thus believes that it is ultimately psychological *conflict*—between love and hate, protection and destruction, compassion and violence—that drives Mr. R's decision-making and behavior, producing a singular action that is fractured across two points in time. It is precisely this conflict that gives the action its unusual, paradoxical character. He explains:

> A battle is raging in the lover between love and hate, both directed towards the same person, and the battle is depicted graphically in the compulsive action, which is also symbolically significant, of removing the stone from the road she is to take and then reversing this act of love. (Freud 1909 / 2002, 153)

The second performance, like the playing of a lengthy game, is thus a *continuation* of the first. It is wholly dependent on the occurrence of removing the stone at all; it functions as

an attempt to *undo* this. Such an entanglement between intentions and actions is necessary, for there cannot be a retraction of one's love until an act of love has been extended, an act to be reversed.

Hence, Freud concludes:

> Compulsive activity of this kind with two consecutive time-signatures, where the rhythm of the first cancels out the second, is a typical feature of obsessive-compulsive neurosis […] Its true meaning lies in its depiction of the conflict between two more or less equally strong impulses, opposites which, in my experience to date, are always those of love and hate. They deserve our particular theoretical interest because they reveal a new type of symptom formation. Instead of arriving at a compromise, as regularly occurs in cases of hysteria, where a single means of depiction suffices for both elements in the opposition, thus killing two birds with one stone, here the two opposing elements each find expression singly, first the one and then the other, though of course not without an attempt being made to produce some sort of logical connection between the two hostile elements—one that often defies all logic. (Freud 1909 / 2002, 153)

But even if this is a unity through some higher conflict, the question remains as to whether it can still possibly count as an action. One might accuse Freud of arbitrarily relating two distinct actions, especially when treating it as a single action invites rendering Mr. R's behavior nonsensical. As the philosopher and psychoanalyst Jonathan Lear notes, "A problem thus arises when we consider Mr R's removing-and-replacing the stone as a whole. For there doesn't seem to be a perspective from which this behavior looks reasonable" (Lear 2005, 26).

If this is true, then any hope of rationalizing Mr. R's behavior, and hence counting it as an action, seems lost.

7.8 | Acting Out the Absurd

While writing about the curious character of Mr. R's behavior with the stone in the road, Freud references a second, similar case in one of his notes. Like Mr. R, this patient, whom Lear calls Mr. S, was also suffering from a compulsive disorder (Lear 2005, 28).

One day, so the story goes, when Mr. S was walking through a park, he stumbled over a branch that was lying across the footpath. Noticing it, he took it upon himself to lift it and heave it into a hedge that ran parallel to where he was walking. Later that same day, he started feeling anxious and worried that the branch in its new location might prove to be a greater obstacle for a casual passer-by than before he handled it. Determined to correct his error, he raced back to the park "and put the branch back in its original position" (Freud 1909 / 2002, 174, n.24). Freud, intrigued by this similarly peculiar behavior, commented that "it would have been clear to anyone except my patient that the original position was bound to be more dangerous for any passer-by than the new one in the bushes" (174, n.24). Was Mr. S also trying to erase his initial performance?

Along with Freud, both Davidson and Lear find this anecdote interesting as well. Reason appears to be involved in both instances insofar as Mr. S had a motive for moving the branch each time (i.e. to keep people safe); taken individually, Davidson and Lear agree that each act of branch-moving should be counted as actions. Unlike Lear though, Davidson misses perhaps the most significant detail of the story, adapting and altering it to instead discuss the irrationality of akrasia.

On Davidson's analysis of this event, both the placement and replacement of the branch happened for reasons, namely, the beliefs that it was a danger to passers-by and that moving it would take care of the problem (Davidson 1982, 173). Davidson, however,

goes on to imagine that after Mr. S moved the branch, he debated with himself about whether he should return to the park, concluding that, all-things-considered, it was not worth the trouble to go back. In this adaptation of the story, then, because Mr. S judges that not returning is the better of the two actions, Davidson goes on to argue that the man acted akratically by returning to the park, for he did not act in accordance with his all-things-considered judgment (174).

Unfortunately, in modifying Freud's account to advance his discussion of akrasia, Davidson overlooks a deceptively significant detail. As Freud noted, Mr. S (like Mr. R) carefully placed the branch back in its *original* position, a position, as he comments, that was far worse for passers-by. This detail thus seems to undermine the *prima facie* intention that motivated both of Mr. S's performances. How could he be trying to *help* people by placing the branch in a worse position? If he judged it to be in a precarious place when he initially moved it, how can it be safe for people to replace it exactly where it was?

Taking this into consideration, unlike Davidson, Lear argues that the restoring of these objects to their original positions is of central importance in appreciating the bizarre nature of these stories (Lear 2005, 30). If this is correct, then it appears that, Mr. S (like Mr. R) attempted to erase his first action by undoing it, by trying to restore things to how they were before he intervened, and so what we are seeing as a result is that a second performance is directly *contradicting* the first, ultimately serving to undermine his explicitly stated goals. More specifically, in both cases, what is being contradicted is the *significance* of some prior performance. This is not to suggest that such initial performances are somehow insincere and are being retracted, as if either man had

changed his mind. Rather, it is quite possible (and likely) that both men *sincerely* intended *both* performances.

How can this be? Is such inconsistency a red flag for irrationality?

7.9 | The Not-So-Uncommon Negations of Everyday Life

Obviously, Freud was not unfamiliar with irrational behavior, having made important contributions to our understanding of psychological phenomena, ranging from anxiety and dreams to destructive impulses and sexual fantasies. A key component to his psychoanalytic approach is taking into consideration that a person's behavior is a manifestation of her psychological conflicts, as if she were a dramatic actor attempting to make sense of the confusing script of her internal life. Through a form of compassion, he was able to help patients put this story into words so that they could not only begin to understand themselves, but understand themselves *as* human, as persons worthy of love, respect, and attention.

Mr. R was one such person for Freud, and though he would be killed during World War I, this was not before Freud felt confident enough to declare that his treatment was successful (Freud 1909 / 2002, 202, n.18). The most surprising aspect of Mr. R's story though is how his contradictory behavior, united in a single action, was such a radical departure from the kinds of irrational behavior with which Freud had become accustomed to analyzing.

Lear believes that one of Freud's ingenuities was to look "for contradiction inside the physical symptom," explaining that he would view a person's "bodily symptom not simply as something caused by the mind's hidden contradiction, but as directly expressing it" (Lear 2005, 61–62). And so, for instance, in the case of another patient

named Elizabeth von R, who was suffering from sensitivity and pain in her leg, after extensive meetings, Freud would conclude that her symptoms were an expression of a conflict between "erotic attachments to men and attachments and obligations to her family," clearly inconsistent motivations that might pull just about anyone in two opposing behavioral directions (62). Unable to satisfy either of these, Ms. von R's symptom thus symbolizes a perverse compromise between the two; she expresses her erotic desire in this way precisely because she suppresses it, achieving a kind of confusing homeostasis that entertains her desire without satisfying it.

Prior to his analysis of Mr. R, Freud found irrational behavior like Ms. von R's to be fairly typical. If she was indeed burdened by two opposing motivations that were in conflict with each another, it is hardly surprising that these psychological forces should issue in unusual behavioral compromises. Although Ms. Von R's particular symptom is far from ordinary, the experience of contradictory motivations is a common occurrence in human mental life, usually terminating in similar behavioral half-measures along with a conscious experience of frustration or shame. Such observations, for example, are what lead Plato in Book IV of the *Republic* to conclude that the psyche must be divided.

Invoking what has come to be known as the *Principle of Opposites*, he explains through the mouthpiece of Socrates that it is not logically possible for the same thing to "do or undergo opposite things; not, at any rate, in the same respect, in relation to the same thing, at the same time;" however, were this to appear to occur, he cautions, "we will know that we are not dealing with one and the same thing, but with many" (*Republic*, 436B). Perhaps the most unusual anecdote he introduces to make his case is that of a man named Leontius, who is motivated both to look and not look at some corpses along the

road that he passes.[134] Rather than act on either motivation, Plato tells us that he struggled for a while, keeping his hands pressed over his eyes. We can imagine that he likely frantically paced back and forth during this, trying to talk himself out of his desire. Unfortunately, unable to control himself, he eventually pulled his hands away, opened his eyes wide while marching over to the corpses, and gazed upon them while simultaneously reproaching himself for doing so (439E–40A). Although Leontius, unlike Ms. von R, ultimately acted on one of his motivations, his behavior is a testament to the psychological conflict he experienced, all the way up to his final action, which both satisfied and failed to satisfy him at the same time.

Given these considerations, is it so strange that opposing inclinations might motivate a person at *different* times rather than the same time? Could not Ms. Von R have experienced the pull of her erotic desire for men when her family was not around, perhaps even entertaining it, whilst at a different time satisfy her duties and obligations to her family? Is this not, in a way, what Leontius did? Do not opposing motivations often

---

[134] Although some might agree that the argument for psychic division turns on whether there is a desire for and aversion to the same thing in the same respect, this is somewhat of an oversimplification in my view. An underappreciated clue in this passage is that each psychic part that Socrates identifies has a different *function*. Socrates himself proposes this as one of the possibilities for understanding division: "whether we *learn* by means of one of the things in us, become *spirited* by means of another, and *feel desires* in turn by means of a third for the pleasures having to do with nourishment and procreation and as many things as are closely related to these" (*Republic*, 436A–B; emphasis added).

What justifies Leontius' psychic division is thus not simply conflicting desires to look and not look, but it is the *ways* in which he is motivated—*passionate desire* for and *violent anger* against, two qualitatively different kinds of desires inclining him in opposite directions. Had Leontius not reproached himself *while* satisfying his appetite for pleasure, there would be less reason to assume he is psychologically conflicted. By contrast, even if pleasure motivates us in two different directions—say, Leontius has a desire to look at corpses but hesitates on account of a desire to have dessert at the bakery before it closes—that conflict does not persist while the desire for pleasure as such is being gratified. Lear interprets Plato similarly on this point (Lear 2005, 165–6).

This is why I speak of Leontius as *motivated* to look and not look, rather than as desiring to look and not look. It is a subtle but important distinction that more clearly emphasizes Plato's commitment to different *kinds* of desires.

cause us to start a task only to veto it moments later, taken away by a different desire? Acting on more than one motivation by no means diminishes psychological conflict—and in fact, it often merely *amplifies* it—but that such motivations *can* incline us to undertake opposing courses of action at different moments seems so self-evident that it ought to qualify as a trivial observation.

Freud's controversial proposal is not that our actions can sometimes contradict themselves—after all, we can and do change our minds—rather, it is that at least sometimes such events ought to be classified as *single actions*. Mr. R and Mr. S were almost certainly motivated by opposing desires in their performances, but if Freud's radical insight is correct, then such actions are proceeding from contradictory desires *united in a higher conflict*, expressing themselves at two different moments in time with the primary goal of the second negating the first.

Before exploring how Freud understands such actions, a question remains as to whether it is possible to avoid arbitrary designations of unified actions. There ought to be a way to distinguish between two genuinely distinct actions and what I will call *negation*, a self-contradictory action whereby an agent undertakes two performances, attempting to undo the first with the second, sometimes even expressing itself in a discontinuous manner. Fortunately, there is such a way, and it has to do with devising a test for self-understanding.

7.10 | That With Which We Get Stupefied

One of the side-effects of psychological conflict, whether it manifests itself in word or deed, is that it can render us unintelligible to others, even creating conditions for a failure in self-reporting. This unusual relationship between communication and contradiction

was brought to our attention by Aristotle, who argues that the principle of non-contradiction—that the same thing cannot both be and not be at the same time and in the same respect—is the foundation of intelligibility. So central is non-contradiction to knowledge, he believes without it, it is impossible to make any distinctions.

Imagine, for instance, if everything in your visual field were perfectly identical in color, including the same shade. How could you pick out one thing from another? How could you make any perceptual distinctions on the basis of vision alone? Aristotle believes something analogous to this holds with respect to knowledge.

First, he argues, in order to reason, the terms that we use must have different meanings from one another. C-A-T cannot mean the same thing as T-R-A-N-S-F-O-R-M-E-R, as well as every other term. There must be something to serve as a semantic point of contrast. Were our terms to fail in this most basic sense, failing to distinguish one thing from *anything* else, then, Aristotle laments:

> Reasoning would be impossible; for not to signify one thing [as distinct from anything else] is to signify nothing. And if words signify nothing, there will be no discourse with another or even with ourselve*s*. For it is impossible to understand anything unless one understands one thing. (*Metaphysics*, $1006^{b}7–11$)

Though he focuses in this passage on the semantics of referents and terms, it is evident that this insight can be generalized in the following way: without the *ability* to distinguish one thing from another (an ability that contradiction compromises), we would find ourselves epistemically stupefied, much like one who tries to perceive objects within a visual field that is uniform in color.

Continuing his argument, Aristotle next warns that if someone *judges* that something both is and is not the case, that is, if someone commits herself to a

contradiction, it makes her utterly unintelligible to others. After all, such a person would have to both affirm and deny any question regarding the thing in question.

Suppose, for instance, that Penelope believes that she is both a woman and not a woman, in exactly the same sense and respect. If anyone were to ask her whether she believes she is a woman, she would be logically obliged to respond that it is both true and false, and the same response would follow if anyone were to ask her whether she believes she is *not* a woman. Supposing she has in mind no poetic interpretation of the matter or alternative way of conceiving things, how can anyone possibly understand this? How could Penelope herself even make sense of these responses? Aristotle does not believe understanding is possible in these cases, both noting and wondering:

> But if all men are equally right and wrong, anyone who holds this [contradiction] can neither mean nor state anything; for he will both affirm and not affirm these things at the same time. And if he makes no judgment but equally thinks and does not think, in what respect will he differ from plants? (*Metaphysics*, $1008^{b}8$–12)

Though Aristotle's treatment is more technical in its mode of delivery, a similar point can be found earlier still in one of Plato's dialogues. While he does not quite explicitly conclude that contradiction leads to a plant-like state, Plato clearly comes close to suggesting as much in a brilliant dialogue known as *Euthydemus*. To explore exactly what happens whenever an individual deliberately holds contradictory opinions while denying the principle of non-contradiction, the *Euthydemus* introduces us to two brothers who are sophists, Euthydemus and Dionsyodorus, each willingly and confidently embracing every absurdity that follows from denying the principle of non-contradiction. Both brothers are fully aware that this obliges them to both affirm and deny everything, but it does not deter them in the slightest; on the contrary, they act as if they cannot be bested in any argument. And, in a sense, this turns out to be true. One of the morals that

Plato conveys to us, through Socrates, is that it is impossible to have a meaningful conversation with anyone who does this.[135]

In both cases, Aristotle and Plato show us the dangers of contradiction, how it can cause not just a failure to be intelligible to others but sometimes even a failure to be intelligible to oneself, resulting in a plant-like state. This phenomenon, this failure to make oneself intelligible *to* oneself, is what Jonathan Lear terms a *reflexive breakdown* (Lear 2005, 26). A reflexive breakdown occurs whenever we are unable to give reasons for our actions, that is, whenever we fail to rationalize our actions to ourselves. During a breakdown, we encounter difficulties responding to straightforward questions such as, "Why did you do this?," usually answering in frustration or bewilderment, "I don't know." It is a state marked by utter confusion after coming face-to-face with a contradiction that sabotages any efforts at self-understanding.

Though philosopher Sebastian Gardner does not share Lear's term, he too recommends that we use something like a reflexive breakdown to identify irrational behavior.[136] He writes:

> The seeds of irrationality lie in a discrepancy between action and self-explanation, the recognition of which is bound up with the possibility of interrogation: if the result, actual or hypothesised, of interrogating a person—calling him to account for his actions—reveals inconsistency between how he represents himself, and

---

[135] See, for example, the arguments at 285A–287D and 299D–300D, both instances ending with Dionysodorus confidently proclaiming that Socrates is unable to cope with their arguments and Socrates unsure of what to say next. At one point in the first argument, Socrates points out that their position is self-defeating, but they are unaffected by this fact, as if it too is meaningless. The dialogue leaves the reader to decide whether philosophy is foolish in the face of sophistry or vice versa.

[136] Gardner almost uses the term though. He believes that his criterion for irrationality, discussed below, "locates a problem of *reflexivity* at the heart of irrationality: the same subject who at each instant of wakeful self-consciousness experiences himself ineluctably as rational, contradicts his own rationality" (Gardner 1993, 4; emphasis added). In other words, by attempting to access her thoughts in search of reasons for her behavior, an individual experiences herself during an episode of irrationality as if she were another whom she cannot understand, her own motives suddenly inaccessible.

> how his actions show him to be, then he is on the verge, at least, of being irrational. (Gardner 1993, 3)

For Gardner, these cross-examinations do not *cause* the self-obfuscation indicative of a reflexive breakdown; rather, they bring the contradiction to the surface, revealing it for what it is. By virtue of our social interactions, we can thus come to see our own nonsense for what it is, for it is in our attempts to justify ourselves to others that we discover a hidden contradiction engendered by our irrationality, one that undermines our efforts to make sense of ourselves. Ms. von R, for instance, could not even begin to articulate her conflict, instead acting it out through hypersensitivity in her leg; and we can only imagine how awkward poor Leontius must have felt trying to explain his antics.

While using reflexive breakdown as a litmus test can prove to be a vital tool in identifying psychological conflict, it is important to be aware that there are a number of confounding factors that need to be ruled out first. A person might be fatigued, intoxicated, uninterested, etc., and such factors might lead her to appear as though she is experiencing a breakdown when she unsatisfactorily responds during a cross-examination. It is thus imperative to be on the lookout for such potential issues that could compromise someone's responses. Similarly, an interlocutor might frame questions poorly, leading an individual to be uncertain of how to answer, which is almost certainly not a reflexive breakdown. Conversely, very quick-witted people who suffer from self-deception might never experience a breakdown, even though they may nonetheless betray their irrationality in other ways. Due to these possibilities, an apparent breakdown in isolation of any additional context should be taken as neither sufficient nor necessary for the existence of some underlying psychological conflict; rather, breakdowns must be

viewed as one piece of evidence—an important one—amongst others that strengthens the case that conflict is motivating someone's behavior.

But even if psychological conflict does not result in Aristotle's plant-like state that is characteristic of reflexive breakdown, it is important to keep in mind, as Gardner suggests, that such conflict often results in inconsistencies in speech or behavior that we can identify. For example, Mr. S stated that his intention was to look after the safety of others walking in the park, but this same intention motivated both of his contradictory performances, as if he were judging that removing and replacing the branch is each simultaneously safe and unsafe. Is it any wonder that he nonsensically returned it to its original position, a position that Freud noticed made the branch unquestionably more dangerous? Even if Mr. S can account for each performance individually, when they are juxtaposed, his inconsistency is unveiled. Though not foolproof by any means, the test of reflexive breakdown in consideration with the rest of the evidence equips us with a means for picking out genuinely irrational behavior motivated by conflict, regardless of whether it *appears* reasonable or unreasonable.

Unfortunately, cross-examining Mr. R and Mr. S is no longer a possibility, and so there is no way to apply the reflexive breakdown test to investigate whether their performances ought to count as single actions unified by internal, contradictory conflicts. What can be performed though is a hypothetical analysis to discover what results when their performances are treated as if they were singular actions, and in both cases, it is evident that the behavior of each man appears nonsensical. Mr. R removes the stone to protect his beloved and then replaces it for no apparent reason, while Mr. S removes the branch *and* replaces it for the *same* reason, namely, to ensure the safety of others walking

through the park. If we were to ask of each performance in isolation why either man acted as he did, both would likely be able to produce some reasons that justify his behavior.[137] One can almost hear Mr. S, for instance, fiercely and defensively respond during his cross-examination, "It wasn't safe!" Though not indicative of psychological conflict, this degree of introspection at least signals that none of these performances were simple, mechanical responses to stimuli; at some level, they were thoughtful and intended.

Suppose that we were to ask each why he undertook *both* performances, *as if* they were a single action. Suppose, specifically, we quizzically pursued the matter with Mr. S in the following way:

> You said that you removed the branch because it wasn't safe, but you also said that you replaced it because it wasn't safe. If it wasn't safe in either position, then why did you bother to move it at all? And if safety considerations were guiding your decisions, why, for that matter, did you remove *and* replace the branch?

How could either rationalize his behavior without treating the performances as two distinct actions? *Can* they even rationalize this? How could Mr. S possibly respond to our queries?

If it turns out that both men *do* fail the reflexive breakdown test, we must still ask how such contradictory performances can possibly count as single actions.

7.11 | Breaking It Down and Putting It Back Together

If Freud's theory of a unified psychological conflict is plausible—a conflict that is expressed by the existence of contrasting pairs of wants, such that each is defined by and

[137] Mr. S gives an explanation for his second performance, but Mr. R does not. Because Mr. R fails to do so (as far as we know from Freud's case study), Lear believes that what Mr. R did was irrational. He identifies other reasons to believe this as well, and this is discussed below.

through its opposition to the other—and we accept Davidson's insight into discontinuous events,[138] then the reflexive breakdown test gives us good reason to consider that this behavior might be some type of species of irrationality. It is possible, in other words, that Mr. R and Mr. S have undertaken a kind of irrational action that I will call *negation*. But if this is correct, it only raises a host of other pressing questions.

Is one justified in treating negation as an action? Specifically, *why* would it count as an action? After all, if it is not possible to rationalize the behavior from the agent's point-of-view, how could it possibly be rendered intelligible *enough* to count as an action? Does it not make *more* sense to treat it as two distinct actions, especially if the agent can rationalize each performance? Or alternatively, why not regard it as meaningless behavior? Underpinning each of these doubts is the concern that reflexive breakdown actually *undermines* the very possibility that a subject performed an action at all by placing the behavior in question beyond rationalization. How could it be otherwise?

Jonathan Lear actually recognizes this central problem, and while he does not acknowledge the unity in the two performances of Mr. R and Mr. S that Freud sees, he does believe that reflexive breakdown indeed works as an indicator of psychological conflict, even defending the idea that the behavior that follows from conflict still counts as an action. As it turns out, his argument for this can accommodate the case of negation as well.

On the surface, it seems as though Mr. R has reasons for doing as he does—he regrets moving the stone in the first place. Likewise, Mr. S replaces the branch to help

[138] The plausibility of Freud's idea is investigated more closely in chapter nine.

people, and specifically in the second instance, to correct a previous error, fearing that he was not *really* being helpful the first time around. Wherein lies the problem?

All of this is true, and Lear admits as much, but he points out that it is not the repositioning of the stone or branch *per se* that is in need of explanation; it is the fact that these objects were placed *exactly where they originally were*. It is only when this detail is taken up into the analysis of the action that it reveals itself for what it is—curious, peculiar, *unintelligible*. As Lear explains of Mr. S, "He does have a motive to remove the stick; he doesn't have a rational reason to place it exactly where it had been" (Lear 2005, 228n14). There is nothing that Mr. S says (or can say) that satisfactorily rationalizes this fuller description of his performance. For Lear, this means that one must either forfeit the idea that the performance is rational or that the performance is an action, and of the two options, he believes it is much easier to construe it as irrational than to not count it as an action.

How so?

Similar to the recommendations outlined for intentional agency in chapter two, Lear believes that a conscious awareness of what one is doing while in pursuit of a goal is sufficient to count a performance as an action. He writes:

> In acting, [Mr. R] *knows* what he is doing in the minimal sense that he is replacing the stone, and he *sees* something to be said in favor of it. He cannot *say* much about it—'I shouldn't have removed it in the first place'—but he does see himself as acting to correct a previous error; and that's enough to make it an action. (Lear 2005, 28; emphasis added)

Thus, while Mr. R may lack a sophisticated understanding of how the particulars of his performance fit within his constellation of beliefs and desires, it does not follow that he is acting mechanically. As Lear observes, there is a sense in which he is aware of what he is

immediately doing: he is picking up a stone that he had moved and is putting it back where he thinks it belongs. He may be unable to specify the *why* of his behavior very well when pressed, but he can certainly describe the *what*.

What Lear shows us is that because a person cannot satisfactorily rationalize her behavior to others or herself, it does not follow that she is not acting; it just raises the possibility that her action is *not rational*. If conflict can compromise one's ability to speak, as proposed above, then why should we expect otherwise from Mr. R, Mr. S, or anyone else in a similar position? Of course you will not be able to give or acknowledge reasons!

As suggested above, Lear's defense can likewise be extended to include *negations*.[139] Even if we regard their contradictory performances as unified into single actions, it is still true that both Mr. R and Mr. S are aware of what they are immediately doing and have motivations for acting in each instance of the behavior; and yet, it is precisely in viewing it *as* a unified action that what they do suddenly appears wholly unintelligible, both to themselves and to us. It was Freud who detected, through his own form of cross-examination, that their two performances were motivated in every instance by desires in opposition to one another, desires to create and destroy the same act, and it is only with the second performance that they can contradict—and hence erase—their first performance. But such expression of conflict comes at a cost, for in so acting, Mr. R and Mr. S sacrifice any hope for self-understanding; in other words, they learn to express

[139] Though Lear does not analyze the two distinct actions as one, his analysis betrays their unity in the way that Freud proposed, for, as he writes of Mr. R above, "he was acting *to correct a previous error*" (Lear 2005, 28; emphasis added).

their conflict through the drama of an action, but they lose their ability to say anything meaningful about it.

7.12 | Conclusion

In addition to being a unique and potentially interesting phenomenon in and of itself, the study of negation reaffirms a recurring theme regarding irrational behavior while inviting us to reconsider some of our Enlightenment assumptions. In particular, Enlightenment thinking places a premium on human freedom, and the causative power behind it, it assumes, is none other than reason. A consequence of this idea is that for any performance to count as an action, it must follow from reason. Indeed, it is reason, action, and freedom that are inextricably bound up with one another, forming a triad that sets out to describe how human nature is so radically different from the rest of the biological world.

Prima facie, none of this appears to be a problem for either Mr. R or Mr. S at the outset. Mr. R desired to move the stone because he wanted to protect his beloved, and Mr. S desired to move the branch because he wanted to keep people safe. Surely both men performed actions?

Yet, Freud characterized Mr. R's stone-moving behavior as *compulsive*, and for any Enlightenment thinker, compulsion is the antithesis of freedom. Compulsion would imply that, in principle, no matter how much Mr. R wanted to do otherwise, no matter how much he *reasoned*, he could not have prevented himself from doing anything differently. Even if compulsion is analyzed in terms of desire, it still shows us that there are nonrational variables beyond our control that can influence our decision-making. A physiological state such as hunger or an environment with scarce resources can both

severely constrain our decision-making and restrict our options to the point where it hijacks our reasoning and prevents us from considering otherwise. It thus seems as if Mr. R actually did *not* perform an action, and if not, what was it?

And herein lies the problem.

Unless reason can do otherwise, Enlightenment thinking advises that we end our investigations prematurely, as if there is nothing further to learn. To see that this is so, consider for a moment one Enlightenment-inspired hypothesis introduced in the introduction to part two, the Deviant Rationality Thesis (DRT). DRT specifies that the moment some behavior strays from rational norms, it must be regarded as irrational. Setting aside the problem of how to define rational norms without begging the question in the first place, this looks like a promising place to begin an investigation of irrationality; however, it becomes clear that DRT quickly becomes dismissive of irrationality.

The dilemma for DRT emerges as soon as we ask, "*Why* was there a deviation from a rational norm?" If we accept the Enlightenment trinity of reason, action, and freedom, there are only two ways to address this, neither satisfactory. On the one hand, if we can find a reason for the deviation, then the behavior in question is *not irrational*, for it is no longer violating a rational norm—unless we suppose that some rational norms are somehow more important than others.[140] On the other hand, if we *cannot* find a reason for

---

[140] Although himself a strong DRT advocate, Davidson was aware of a similar issue when attempting to explain irrational behavior. In "Paradoxes of Irrationality," he writes of the dilemma:

> The underlying paradox of irrationality, from which no theory can entirely escape, is this: if we explain it too well, we turn it into a concealed form of rationality; while if we assign incoherence too glibly, we merely compromise our ability to diagnose irrationality by withdrawing the background of rationality needed to justify any diagnosis at all. (Davidson 1982, 184)

As long as irrationality continues to be defined in contrast to rationality and justification through reasons, I see no way out of either dilemma.

this, then it is not to be regarded as an action. However, an action for an Enlightenment theorist is, by definition, free, and if it is not free, then she argues that it is determined. Is causal determination synonymous with irrationality? Of course not.

The alternative to Enlightenment thinking on this topic is to broaden our understanding of intentional action. Rather than view intention as consisting of a combination of propositional attitudes, such as a belief-desire pair that can be specified in language, we must consider the matter in terms of representations of states-of-affairs and commitments to goals. In so doing, we can begin to infer through careful observation what a person's *actual* interests are and detect whether she has conflicting desires or goals. Though neither Mr. R nor Mr. S, for instance, consciously apprehended the importance of negating a previous performance—at least not in the absence of some counseling by Freud—we were able to discern through the broader context the significance of what they were doing, what they were trying to do without initially realizing it. Their neuroses had brought them to an impasse, unable to live well, and had they not submitted to cross-examination, their hidden motives would have continued to obstruct their efforts to satisfy some of their desires and goals. In the case of Mr. S, like a split-brain patient, reason had even confabulated charitable motives, creating the *appearance* that his behavior was sensible.

We must also try to understand irrationality on its own terms rather than juxtaposed to reasons and justifications. In each of these analyses throughout part two, it is clear that inconsistency and contradiction are the hallmarks of irrationality, especially when such inconsistency becomes an obstacle to one's goals, making for a frustrated lived experience that goes nowhere, such as that experienced by Ms. von R. I will coin

this *teleological inconsistency*. This cannot be understood as a mere deviation from a rational or social norm, for some of these people can even make their frustrated existence sound *reasonable*.

The cause of teleological inconsistency is some kind of psychological conflict, a conflict that engenders a contradiction, sometimes at the level of beliefs, sometimes at the level of desires, and sometimes at the level of actions. Conflict in and of itself is not always a problem. As we have seen, there can be strategic deployments of wishful thinking and self-deception that help enhance one's reputation or avoid fear. In every case, however, the conflict becomes irrational the moment that it disrupts the ability to live well.

Finally, not only does Enlightenment thinking unnecessarily—even hopelessly—restrict the scope of what counts as irrational, it has repeatedly demonstrated that if it is to make progress in understanding phenomena such as akrasia or negation, it must undermine itself. In the case of akrasia, this requires, as Davidson had recognized, ultimately dividing the mind, for there seems to be no other way to account for the failure of reason unless one denies that akrasia exists altogether. And in the case of negation, this requires acknowledging nonrational states that exercise some influence on reason itself, likewise compromising the autonomy and power of reason.

So how might a divided mind help explain these issues? Is it even necessary to accept such a thesis to best understand human behavior and irrational action? Why does rationality seem to fail in the first place? From where do our psychological conflicts come? Why are we irrational at all?

These questions, and others like them, are precisely what motivate part three.

Part III. Surveying a New Frontier

In 1978, psychologists David Premack and Guy Woodruff published an article detailing their experiments with a chimpanzee. What was revolutionary was neither the method used nor the test subject, for numerous studies had been conducted on chimpanzees prior to their work; it was the *question* that was asked: "Does the chimpanzee have a theory of mind?"

Posed today, the question strikes us as perhaps interesting but fairly standard as far as ethology or cognitive science goes. But in 1978, this was the first time that a wider audience had been exposed to the concept of a *theory of mind*, a concept that is now routinely circulated amongst cognitive psychologists, cognitive scientists, philosophers of mind, and other specialists working with mental ideas such as *intention*, *belief*, *doubt*, *knowledge*, etc. Whereas previous experiments focused on whether or not animals are capable of problem-solving or tool-use, Premack and Woodruff wanted to understand better just *how* these animals view the environment, each other, and themselves. Do they "see" the world in a manner somewhat analogous to human experience? Do they make assumptions about what others are doing? Can they empathize? Just how rich is their mental experience? They explain, "We are less interested in the ape as a physicist than as a psychologist (every layman, of course, is both); we are interested in what he knows about the physical world only insofar as this affects what he knows about what someone else knows" (Premack and Woodruff 1978, 515).

By *theory of mind*, then, what Premack and Woodruff are referring to is the capacity of an organism to attribute mental states to itself and others (Premack and Woodruff 1978, 515). Can a chimpanzee grasp, however loosely, that it intends, believes,

denies, or hates? Does it have a sense of another's sadness, frustration, or goals? Can it deliberately or consciously make inferences, process information, or draw conclusions, especially with respect to what others might be thinking and doing? How might something like this even be tested?

The researchers decided to place a human actor in four different situations, each requiring that the actor look as though she is trying to find a way to secure bananas that are inaccessible. The actor might, for instance, find that bananas have been suspended atop the cage; or she might see that they are located outside of the cage, just beyond her reach. In each situation, she was instructed to perform different tasks, only sometimes culminating in successfully reaching the bananas. If a box impeded the actor's ability to access the bananas, for instance, the researchers would follow up with her and request that she step on the box or lie on her side, stretching outside of the cage by using a rod (Premack and Woodruff 1978, 516). The purpose of recruiting an actor was to record thirty seconds of video showing how she solved the problem as well as to photograph her in various poses that may or may not lead to a solution.

Because the chimpanzee, who was named "Sarah," was familiar with watching television, the goal was to show her a video clip all the way up to the moment before the actor completed the task. A researcher would then present two photographs to her, one with the actor in a fortuitous position to complete the task and the other with the actor in a position that would result in failure. To control for social cues, the researcher would step out of the room after leaving the two photos, waiting on Sarah to pick one, drop it in a designated location, and ring a bell when she was ready (Premack and Woodruff 1978, 516).

Was Sarah able to infer what the actor was trying to do and whether she could reach the bananas?

Premack and Woodruff discovered that Sarah was able to select the correct photograph an impressive 21 out of 24 times, and each of her 3 errors were related to the same exact problem, the need to remove cement blocks from a box to reach the bananas. They emphasize that although Sarah had been mostly raised in a laboratory setting and had previously performed tasks related to causal inference and reconstruction of disassembled objects, the particular situations in which the actor was placed were entirely novel for the chimpanzee (Premack and Woodruff 1978, 516).

How was this possible?

According to Premack and Woodruff, there were four families of interpretation: matching, associationism, theory of mind, and empathy. On the matching interpretation, the animal succeeds by merely surveying the physical environment and matching one feature with another, as if she were assembling a jigsaw puzzle (Premack and Woodruff 1978, 516). If the video stopped when an actor was holding a rod, then maybe Sarah simply selected the correct photo by finding the one that contained the same elements (actor, rod, and bananas)?

Associationism, by contrast, argues that the animal succeeds by drawing inferences from previous, similar experiences, making the cognitive requirement little more than a mechanism of elaborate pattern recognition (Premack and Woodruff 1978, 516; 518). Perhaps Sarah encountered similar situations without the researchers realizing it? Was there *ever* a time where she needed to figure out how to attain an otherwise inaccessible object?

Alternatively, theory of mind insists that the animal succeeds by attributing mental states to the actor, such as intentions, goals, beliefs, desires, and knowledge (Premack and Woodruff 1978, 518). When Sarah sees the actor struggling for the banana, she can attribute to the actor a *want* or *desire* for the banana and a *belief* about how to reach it. After viewing the photographs, Sarah then infers the correct solution based on her thoughts about the actor.

Lastly, on the empathy model, the animal succeeds by imaginatively placing itself in the shoes of the actor, making inferences based on what it itself would do in that situation rather than what the actor would necessarily do (Premack and Woodruff 1978, 518). Maybe Sarah was making inferences about *herself* rather than the actor?

The proposals for understanding Sarah's success and interpreting her behavior are not meant to be presented as mutually exclusive. Premack and Woodruff clarify that there are explanations available that can draw on features from associationism, theory of mind, and empathy; rather, the problems arise whenever theorists view these ideas reductively and as mutually exclusive (Premack and Woodruff 1978, 518). There may be associative mechanisms at work, but are they the whole picture? Can *everything* be accounted for by association?

Though Premack and Woodruff defend the importance of a theory of mental states over one of brute behaviorism, they also argue along lines of biological continuity, expressing skepticism that human beings are the only organisms capable of attributing mental states and motivations to others (Premack and Woodruff 1978, 525). If it can be proven that there is a natural tendency to impute mental states to others and that other species are capable of doing this—suppose, for instance, that some organisms have

evolved to do this quite effortlessly—then, they argue, the *unnatural* supposition is the behaviorist's recommendation to suppress any reference to mental states.

Whether Sarah really does have a theory of mind with which she makes inferences, the work of Premack and Woodruff raises even more interesting questions. If there really are minds, what is one like? Is a mind supposed to be some kind of unity or is it divisible into parts? What is a whole mind like? What about the parts? Is there just a conscious mind, or is some of it unconscious as well? Does a mind help us draw inferences? Is a mind necessarily rational?

In the introduction, I attributed many of our contemporary views of agency and the mental to the Enlightenment, especially to one of the founding fathers, Descartes. The Enlightenment persuaded so many of us that to be human is to be in possession of a mind, one very different from what any other animal might possess. Our minds are not only assumed to be non-physical and unified, but they are *rational*.

And yet throughout part one, drawing on research in the sciences as well as psychological experiments, it was argued that human agency and human mental life is just one end of a much larger spectrum, that we share much in common with the rest of the animal kingdom. Compounding the problem further, throughout part two, we looked at the different ways in which our so-called rational mental life fails us, leading us not only appear less-than-rational but even inclining us to act in ways that are absurd and self-defeating. Why would a unified, rational mind fail? How can it fail?

Some of the Enlightenment thinkers, such as Locke, are not unaware of these issues, going so far as to even provide a defense that seeks to save the rationality and unity of the human mind. Others, such as Hume, argue that we are better off abandoning

these suppositions in favor of a reduction to associative principles. While the Enlightenment has furnished a wealth of knowledge and bequeathed to us ideas that are held sacred today, such as humanism and freedom, we are at a point in time where it is becoming harder to deny that it has failed us when it comes to understanding the mind and agency.

It is time to revisit Descartes' dreams and ask once more, which path shall we take in life? How does our understanding of reason and human nature move forward? A newer theory of mind is needed, one freed from the hyper-rational shackles of the Enlightenment.

Chapter 8: From Mind to Module

This is part and parcel of the fact that,
whereas *you* can use the context to figure out what Groucho meant,
Groucho can't.

If Groucho meant something that was equivocal-but-for-the-context
then, epistemically speaking, he and his interpreter would be on all fours:
if Harpo could use contextual information to figure out what Groucho meant,
then Groucho could too.

This conjures up a situation more absurd than an elephant in pajamas.

— Jerry Fodor, *Hume Variations*

8.1 | Beams in Our Eyes

For an Enlightenment figure like John Locke (and later, Donald Davidson), irrationality can only be understood against the background of rationality, as something opposed to it. Just as people tend to error even though we are supposed to be rational truth-seekers, Locke also cannot help but notice that people do think and behave irrationally at times. For him, the most pressing forms such irrationality takes are the failure of self-knowledge and the failure to consent to reason, neither of which are uncommonly experienced by people. He writes:

> There is scarce anyone that does not observe something that seems odd to him, and is in itself really extravagant in the opinions, reasonings, and actions of other men. The least flaw of this kind, if at all different from his own, everyone is quick-sighted enough to espy in another, and will by the authority of reason forwardly condemn, though he be guilty of much greater unreasonableness in his own tenets and conduct, which he never perceives, and will hardly, if at all, be convinced of.
>
> This proceeds not wholly from self-love, though that has often a great hand in it. Men of fair minds, and not given up to the overweening of self-flattery, are frequently guilty of it; and in many cases one with amazement hears the arguings, and is astonished at the obstinancy of a worthy man, who yields not to the evidence of reason, though laid before him as clear as daylight. (Locke *EHU*, II.xxxiii.1–2)

To Locke, it seems bizarre that people are able to identify what is unreasonable in others with much greater ease than what is unreasonable for oneself. Should not reason be an impartial judge? Although pride seems to suffice as a partial explanation for some instances of this, it cannot account for all of it, especially in those cases where a reasonable-minded person is presented with clear evidence and argument and yet still refuses to accept it. In fact, this phenomenon is so strange for Locke that he likens it to *madness*, believing it to be deserving of the name for its "opposition to reason" and insisting that if someone were in this state at all times, she would be "thought fitter for Bedlam, than civil conversation." Given these observations and concerns, he proposes, "And if this be a weakness to which all men are so liable; if this be a taint which so universally infects mankind, the greater care should be taken to lay it open under its due name, thereby to excite the greater care in its prevention and cure" (II.xxxiii.4). Thus, if we are to relieve ourselves of this madness that can stricken us, then we must first find the causes.

8.2 | A Habit of Madness

If it is not pride alone, then what exactly causes madness? Two other candidates that Locke considers are education and prejudice. While he believes that the former is often sufficient as a cause while the latter is a good synonym, he ultimately rejects both of these, arguing that neither reaches "the bottom of the disease, nor shows distinctly enough whence it rises, or wherein it lies" (Locke *EHU*, II.xxxiii.3). Still, it is nonetheless striking that he assigns education as a *cause* of madness and treats prejudice as a suitable synonym for it. This is owed in large part to his theory of habit. For even though the mind is rational by nature in its orientation, he does believe it is often victim

to the limitations imposed on it by experience and language. What might the significance of habit play?

According to Locke, there are two different trains of thought that people have, making him a progenitor to what is today known as *dual process theory*. Hugo Mercier and Dan Sperber define dual process theory as "the idea that there are two quite distinct basic types of processes involved in inference and more generally in human psychology" (Mercier and Sperber 2017, 44). The categories into which these two types of thinking can be allegedly sorted are familiar. One type of thinking is associative, characterized by fast, intuitive, and unconscious thoughts, while the other is practical and reasonable, requiring language to carry out a slow, deliberate, and conscious calculation (46). Though the following is a stereotype, it helps illustrate the distinction to consider the difference between the abstract artist and the mathematician. One, the artist, accomplishes her goal by tending to her *feelings* and allowing the free-flowing, spontaneous splashes of paint to express them, while the mathematician must take her time, carefully and clearly thinking through the problem before her.

Though Locke's division is similar, it would be a mistake to assume it is identical. Whereas dual process focuses on the inferences that follow from two different types of thinking, Locke's emphasis is instead on *how* ideas are associated. He does not recognize one type of thinking as *associative* and the other as *rational*; for him, *all* ideas are associated. The question then becomes *in what manner* are they associated, reasonably or spontaneously? He thus draws the line between ideas that have been associated by reason and those by habit. He writes:

> Some of our ideas have a natural correspondence and connexion one with another; it is the office and excellency of our reason to trace these, and hold them together

> in that union and correspondence which is founded in their peculiar beings. Besides this, there is another connexion of ideas wholly owing to chance or custom; ideas that in themselves are not at all of kin, come to be so united in some men's minds, that 'tis very hard to separate them, they always keep in company, and the one no sooner at any time comes into the understanding, but its associate appears with it; and if they are more than two, which are thus united, the whole gang, always inseparable, show themselves together. (Locke *EHU*, II.xxxiii.5)

Locke attempts to account for how these spontaneous associations of ideas can occur by tentatively proposing a kind of crude neurophysiology involving animal spirits—a thin and airy fluid-like substance—repeatedly tracing pathways through the brain until such a pathway becomes well-worn and allows for the spirits to effortlessly traverse.[141] Although the explanation is mechanical, he momentarily sidesteps determinism by allowing the spirits to be set in motion by external causes, such as "inclinations, educations, interests, *etc*.," causes that presumably are to some degree within our control (II.xxxiii.6). It is for this reason that Locke believes a certain degree of madness is *preventable*.

8.3 | Spontaneous Connections

When ideas are associated by reason, Locke considers the connection *natural*, and when a spontaneous connection occurs, it is *accidental*. It is not obvious what he means by a natural connection, but because accidental connections idiosyncratically vary from person to person and because natural connections are formed by reason, it is plausible that Locke takes natural connections to be associations between ideas that are nearly universal. With

[141] Locke does not appear to be committed to the idea of animal spirits in his explanation even though he acknowledges that it is the best explanation available at the time. Reflecting on how a trained musician will recall a song and be able to play it effortlessly after hearing a few initial notes, he says, "Whether the natural cause of these ideas [in the musician], as well as of that regular dancing of his fingers, be the motion of his animal spirits, I will not determine, how probable soever, by this instance, it appears to be so: But this may help us a little to conceive of intellectual habits, and of the tying together of ideas" (Locke *EHU*, II.xxxiii.6). It is clear that what is most important for his theory is the idea that repetition of the coincidence of ideas is directly proportionate to the strength of their association.

a rudimentary understanding of arithmetic, for instance, most people would make the inference that four is the sum of adding two whole units with two more whole units. By contrast, the accidental connections are a "strong combination of ideas, not allied by nature, the mind makes in itself either voluntarily, or by chance" (Locke *EHU*, II.xxxiii.6). One example in particular that Locke presents is that of an accomplished young dancer who, while honing his skill, happened by chance to always have an old trunk in his room. As a result, he could only complete his dances satisfactorily whenever that same trunk or an equivalent was present in the room during his performances, presumably because the ideas related to dancing had been so strongly, accidentally associated with the idea of the trunk (II.xxxiii.16).

Because accidental associations are formed through repetition, Locke believes their occurrence is preventable provided that they are tended to, corrected, and avoided early enough in the process. For this reason, he stresses that "those who have children, or the charge of their education, would think it worth their while diligently to watch, and carefully to prevent the undue connexion of ideas in the minds of young people" (Locke *EHU*, II.xxxiii.8). This would appear to make his cure for madness equivalent to a kind of behavioral conditioning, which is a curious recommendation for someone who trusts in rationality, and he is even explicit about the impotence of reason in the face of such well-established connections. It is not difficult to imagine him penning with regret the following concession:[142]

[142] Interestingly enough, this chapter, which is the last of Book II, was added in the Fourth Edition in 1700. The reason for this, explains Stephen Buckle, is that Locke needed "to rectify an important oversight in his treatment of the mind: associative processes are undeniable, and evidence of serious limitations on human rationality" (Buckle 2004, 146).

> When this [accidental] combination is settled, and whilst it lasts, it is not in the power of reason to help us, and relieve us from the effects of it. Ideas in our minds, when they are there, will operate according to their natures and circumstances; and here we see the cause why time cures certain afflictions, which reason, though in the right, and allowed to be so, has not power over, nor is able against them to prevail with those who are apt to hearken to it in other cases. (II.xxxiii.13)

The only correction that he can imagine is a kind of counter-balance to the formation of such associations in the first place. It is possible that with time and no reinforcement, the associations will weaken and create the opportunity for correction; likewise, a similar outcome is possible through the frequent experience of associations to the contrary.

Thus, Locke's explanation for both the existence of error and madness is the same: through chance and human physiology, the mind can be set in motion to make associations between ideas that are unreasonable. Because these arise through a nonrational, mechanical process, they can prove difficult to correct through reason, and so the solution is prevention, particularly through education and conditioning.

This raises an interesting question though: what guarantees that so-called rational processes are in fact different from spontaneous ones? Is it not possible that *all* mental associations are formed through custom and chance? Borrowing from Locke's account and pushing it to its logical conclusions, this is indeed the line that one anti-Enlightenment thinker would take, Hume.

8.4 | The Not-So-Free Association of Ideas

It might seem strange at first to consider Hume an anti-Enlightenment thinker. Cognitive psychologist Steven Pinker, for instance, includes him amongst his pantheon of

---

I disagree with Buckle's identification of Locke's spontaneous processes as *associative* since, as noted above, even rational processes are associative insofar as they connect ideas. The difference for Locke is between associations made under the guidance of reason and those that are not.

Enlightenment intellectuals (Pinker 2018, 8; 10). To classify Hume as an anti-Enlightenment thinker is not to suggest that he did not share in some familiar themes common amongst 17th century intellectuals—he clearly advocates for a criticism of tradition, especially religion, for example—but he diverges in an important way from many of them. As Hume scholar Stephen Buckle explains, "Hume seems at odds with these other [Enlightenment] thinkers: he cannot share in the enthusiastic endorsement of Reason in terms of which the attack on tradition was so often mounted; and his scepticism seems so thoroughgoing that it must rule out any optimistic hopes for social progress (or 'improvement')" (Buckle 2004, 45).

It was mentioned in the introduction that Locke was driven by a confidence in reason and a faith in its ability to discern truth (provided that it operated within its limits),[143] and as a result, the previous sections explored how he was unable to adequately tie error and irrationality into his theory. The interesting solution he proposes is that some non-rational processes must be operating alongside reason, but it is ultimately unsatisfactory because it fails to follow through with the implications of positing that there are such non-rational processes. Hume, intrigued by Locke's analysis and solution, sets out to determine what happens when one makes associative processes and error central rather than peripheral to a philosophical project (Buckle 2004, 147). One consequence is that he dispenses with the idea that reason is allied with truth. Buckle summarizes his project well: "human rationality, whatever its capacity and extent, cannot insulate itself against the corrupting effects of associative mechanical processes. The

[143] See section 0.5.

human mind *inevitably* falls into error" (146; emphasis original). Why does Hume believe this? And what function could reason possibly serve if not the search for truth?

Hume believes that the mind relates its ideas through a different kind of logic, association, and he identifies three general categories into which any particular associative activity might be classified: resemblance, contiguity, and cause and effect (Hume *EHU*, 3.2). His reason for taking association to be central to thought is inspired by two observations. At both extremes of mental activity, ratiocination and dreaming, one can observe a connection between ideas that is not obviously logical. Hume writes:

> It is evident, that there is a principle of connexion between the different thoughts or ideas of the mind, and that, in their appearance to the memory or imagination, they introduce each other with a certain degree of method and regularity. In our more serious thinking or discourse, this is so observable, that any particular thought, which breaks in upon the regular tract or chain of ideas, is immediately remarked and rejected. And even in our wildest and most wandering reveries, nay in our very dreams, we shall find, if we reflect, that the imagination ran not altogether at adventures, but that there was still a connexion upheld among the different ideas, which succeeded each other. (3.1)

The more remarkable of the two claims that Hume makes here is that association even underlies our more rational patterns of thought. The fact that our sequences of thoughts are often interrupted—no matter how strong our attempts at organization and precision—by some passing thought that we did not intend to entertain is what leads him to believe that there is a web of connections amongst our ideas that we often fail to perceive.

The experience is certainly familiar enough. While working on a logical proof, for instance, the appearance of an existential quantifier ∃ might trigger something like the following chain of ideas:

- a thought about how it looks like a backwards letter E
- a recall of Einstein's famous equation $E = mc^2$
- intrigue with the *idea* of energy (it is in the equation and starts with the letter E)

- an observation that I feel tired (a way of lacking energy)
- worry that I have been getting insufficient sleep (a justification for feeling tired)
- a concern about changing the bed sheets (a worry loosely related to sleeping)
- fantasies about rearranging my room (from thinking about my bed and solving problems)
- a passing curiosity that the film title *Room* denotes two movies that could not be more different from one another
- and so forth…

Once I become aware of these intruding thoughts, I can always try to redirect my attention and return to the problem, but part of the argument for Hume is that if thought were *necessarily* logical, these disruptions should not occur so easily, if at all. Not only is this *not* the case, but, contrary to Locke, Hume goes so far as to contend that the connections between ideas are not logically related *at all*. Buckle remarks of this account that one of the striking features is the *absence* of reason, explaining, "Ideas are connected by the imagination, and the principles that order the imagination are principles of association, not of reason" (Buckle 2004, 142).

On what grounds does Hume believe this?

8.5 | A Classic Dish of Imagination *con* Habit with a Reason Reduction

It is evident that mental associations disrupt our acts of thinking, imposing themselves on our awareness, but Hume believes that such associations are what make inferences possible. To justify this claim, we must first distinguish between reasoning and inference, and in fact, Hugo Mercier and Dan Sperber credit Hume as an important figure who makes this key distinction, believing that he takes reasoning to be part of a broader category of inference (Mercier and Sperber 2017, 51).

Throughout *An Enquiry Concerning Human Understanding*, Hume marvels at man's ability to draw inferences, and he attempts to explain how and why people can do

so *without* invoking reason. For example, he wonders, on what grounds does a person come to believe that an egg will taste pleasing or that bread will nourish (Hume *EHU*, 4.20; 4.21)? He even contends that if God had created Adam with perfect rationality, "he could not have inferred from the fluidity and transparency of water, that it would suffocate him, or from the light and warmth of fire, that it would consume him" (4.6). It is not through reason that people come to learn these things. Instead, Hume contends:

> Nature will always maintain her rights, and prevail in the end over any abstract reasoning whatsoever. Though we should conclude, for instance, as in the foregoing section, that, in all reasonings from experience, there is a step taken by the mind, which is not supported by any argument or process of the understanding; there is no danger, that these reasonings, on which almost all knowledge depends, will ever be affected by such a discovery. If the mind be not engaged by argument to make this step, it must be induced by some other principle of equal weight and authority; and that principle will preserve its influence as long as human nature remains the same. (5.2)

If it is not through reasoning that we come to draw inferences about the world, then what is it?[144]

Hume insists that we learn about the world through *custom*, a kind of inference that people make after the imagination, through the medium of experience, comes to relate constantly conjoined objects as cause to effect (Hume *EHU*, 5.5). For example, suppose that during the first instance with a hot stove, a person observes the metallic coils transition from black to red. How is this person supposed to know that a red coil will burn her skin, especially if she has never seen anything like this before? It would not be until she *feels* the heat that she would know to keep her distance, but even then, this would

[144] Hume equivocates with his use of the term *reasoning*. In some instances, as in the first sentence of the preceding quote, he has in mind *logical* inference, denoted by the qualifying adjective *abstract*. In other instances, as in his usage of the term in the second sentence, he intends it to be synonymous with *inference* in general. The difference is often signaled by the context. Whenever he discusses reasoning by *argument*, it is of the first type.

assume that she had already formed an *additional* mental association between heat and harm, an association that would not be created until some intense form of heat caused her some degree of pain prior to this moment. Hume believes that none of these ideas in her mind are *logically* related; they are formed by *custom* in accordance with the principles of association and on the basis of her particular, idiosyncratic experiences. In fact, these kinds of inferences are drawn so effortlessly that he writes of them, "All these operations are a species of natural instincts, which no reasoning or process of the thought and understanding is able, either to produce, or to prevent" (5.8).

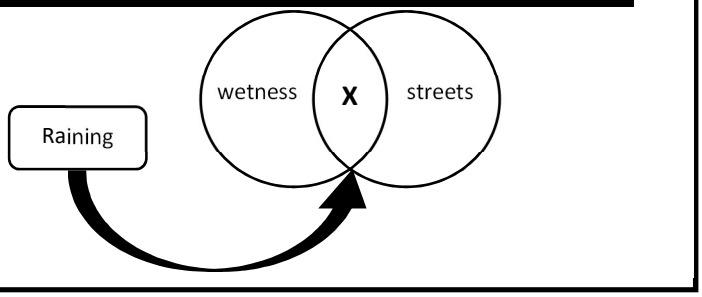
Even though reason plays no role in relating our ideas, it should be noted that Hume does not believe that man is *incapable* of reasoning. Elsewhere, he acknowledges that through acts of the will, a person can exercise freedom of thought, doing as she pleases with her ideas. Such freedom of thought, however, has its limitations, amounting "to no more than the faculty of compounding, transposing, augmenting, or diminishing the materials afforded us by the senses and experience" (Hume *EHU*, 2.5). With the exception of transposing—or moving ideas around, e.g., imagining a human head on a horse's body and vice versa—these abilities are roughly analogous to basic arithmetical operations (such as multiplying, adding, and subtracting), and when considered more closely within specific contexts, they prove incredibly conducive to how we envision reason functions. So although Hume never quite defines *reason*, it can be argued that he considers it to be a *particular way* in which the abilities of the imagination are exercised. If this is true, then logic, by extension, would be the codification of some set of rules regarding how to use these mental operations well.

For example, consider a *modus ponens* argument of the form:

P_1 If it is raining, then the streets are wet
P_2 It is raining
C The streets are wet

On the view I am attributing to Hume, using an argument form is just one way to *represent* an inference that we make between three ideas, namely [raining], [wetness], and [streets], an inference determined wholly by the imagination. In this case, with the mental operation of *compounding*, the imagination has combined [wetness] and [streets] to create [wet streets], and from the association of *cause and effect*, we make the inference [wet streets] from [it is raining]. This occurs entirely through habituation, and this process can be represented in other ways as well (see figure 8.A).

figure 8.A **(Representation of Hume on Habituation and Inference)**



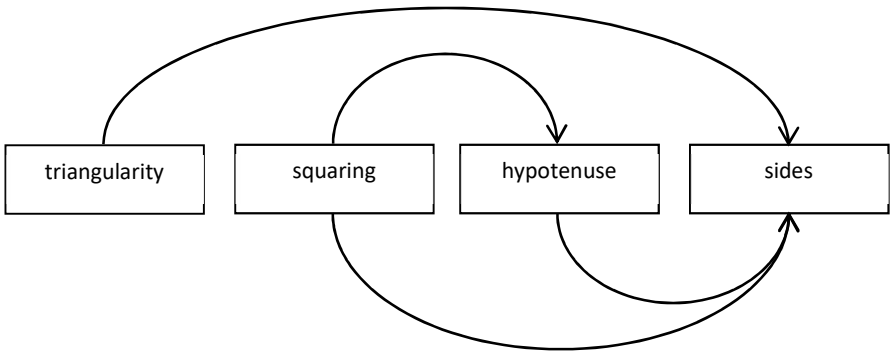
For Hume, *modus ponens* thus merely provides a *structure* on to which we can map ideas and express relations.[145] Any inferences drawn with it pertain *to that structure*, not the terms involved. This explains why when such terms are supplied, as in "real world reasoning," people sometimes make inferences that happen to resemble that pattern of argument and sometimes they make very different inferences. Such "structural reasoning" is what Hume refers to as *reasoning concerning relations of ideas*, and he

[145] See *Enquiry Concerning Human* Understanding (Hume 1999, 14). Mercier and Sperber actually make a similar argument themselves, contending that logic may be best understood as a heuristic tool that highlights and exaggerates the relationship between the reasons for a conclusion and the conclusion itself (Mercier and Sperber 2017, 158–168).

distinguishes this from the "real world reasoning" of *matters of fact* (Hume *EHU*, 4.1). Hume scholar Tom Beauchamp describes this distinction in terms of the difference between *demonstrative* and *practical* reasoning, respectively.[146]

This view of reason and logic that I have proposed as Hume's is consistent with what he has to say about demonstrative reasoning. He uses the following example: "*That the square of the hypothenuse is equal to the square of the two sides*, is a proposition, which expresses a relation between these figures" (Hume *EHU*, 4.1; emphasis original). The implication is that these ideas are *already* related by the imagination *in advance* of any reasoning (see figure 8.B).

***figure 8.B* (Ideas Related by Imagination)**



Our reasoning merely highlights these relations by directing our conscious attention to just some of the mind's associations, namely, those considered relevant to the particular inference. Reason is thus somewhat epiphenomenal, lacking the power to *create* relations and only allowing us to observe them.[147] In other words, to reason is to bring into

[146] Hume also uses the synonyms, demonstrative reasoning and *moral* reasoning, grouping not only matters of fact—or reasonings about objects of experience—under the latter but also existential claims.

[147] I say "somewhat epiphenomenal" because even if we are unable to create new associations between ideas through reasoning, *where* our attention is directed can have consequences for our behavior. Perceiving my association between [rain] and [wetness] might incline me to reach for an umbrella; whereas directing my mind to the association between [keys] and [driving] might distract me from taking one.

conscious awareness and perceive some relations that already exist between ideas in the recesses of the mind. But what of practical reasoning? Is it not the case that such means-end reasonings are under our control?

It is unclear in Hume how much he believes the mind operates mechanically and how much autonomously (i.e. under an agent's control). When discussing the operations of the imagination and freedom of thought, he makes it sound as if a person can combine and alter any ideas that she pleases—that is, that she can *voluntarily* relate ideas (Hume *EHU*, 2.5). And yet elsewhere, he argues that belief is nothing more than an involuntary feeling that something is true, a feeling that proceeds from an accumulation of frequent experiences that reinforce the mechanical operations of association made between ideas (5.12; 5.14).

This leads him to sound determinist:[148]

> I shall add, for a further confirmation of the foregoing theory, that, as the operation of the mind, by which we infer like effects from like causes, and *vice versa*, is so essential to the subsistence of all human creatures, it is not probable, that it could be trusted to the fallacious deductions of our reason, which is slow in its operations; appears not, in any degree, during the first years of infancy; and at best is, in every age and period of human life, extremely liable to error and mistake. It is more conformable to the ordinary wisdom of nature to secure so necessary an act of the mind, by some instinct or mechanical tendency, which may be infallible in its operations, may discover itself at the first appearance of life and thought, and may be independent of all the laboured deductions of the understanding. (5.22; emphasis original)

[148] Hume refers to the activities of mental association as functioning mechanically, possibly even automatically at all times ("infallibly"), and while this sounds determinist, it ultimately hinges on whether there is a will and whether it is an extension of these mental operations or independent of them. Even if the mental operations of the mind are automatic and autonomous, it might still be possible that a person can execute or veto the practical recommendations that the associations make. And Hume indeed appears to have something like this in mind. Of freedom of the will ("liberty"), he writes, "By *liberty*, then, we can only mean *a power of acting or not acting*, *according to the determinations of the will*; that is, if we choose to remain at rest, we may; if we choose to move, we also may" (Hume *EHU* 8.23; emphasis original).

Although it at first appears that he is claiming that only our inferences *from cause and effect* follow from these associative mechanisms, he says in the preceding section that "all reasonings concerning matter of fact seem to be founded on the relation of *Cause* and *Effect*," and so this would suggest that, at a minimum, practical reasoning itself springs from a mechanical, involuntary tendency (4.4). The dilemma then becomes: If practical reasoning is *involuntary*, then why refer to it as *reasoning* at all? If, however, practical reasoning is *voluntary*, then is this not ultimately a reiteration of Locke's account, viz. that there are two different kinds of mental associations that can be made between ideas?

Earlier in *An Enquiry Concerning Human Understanding*, Hume explains that to be reasonable is to act with an intention or purpose in the pursuit of happiness, defined as the "gratification of some passion or affection" (Hume *EHU*, 3.4). This suggests that the proper function of reason is *practical*; that is, rather than discover truth, reason is used to help gratify our wants and needs. However, he is also clear that such reasoning processes are automatic and mechanical.[149] When reflecting on the similarities and differences between the abilities of humans and animals to make inferences, Hume concludes:

> But our wonder [at animal instincts] will, perhaps, cease or diminish; when we consider, that the experimental reasoning itself, which we possess in common with beasts, and on which the whole conduct of life depends, is nothing but a species of instinct or mechanical power, that acts in us unknown to ourselves; and in its chief operations, is not directed by any such relations or comparisons of ideas, as are the proper objects of our intellectual faculties. Though the instinct be different, yet still it is an instinct, which teaches a man to avoid the fire; as much as that, which teaches a bird, with such exactness, the art of incubation, and the whole œconomy and order of its nursery. (9.6)

[149] In a footnote, Hume gives an extended defense of his theory that practical reasoning is reducible to custom, beginning with the rhetorical question, "Since all reasoning concerning facts or causes is derived merely from custom, it may be asked how it happens, that men so much surpass animals in reasoning, and one man so much surpasses another" (Hume *EHU*, 9.2.n20)? In short, such differences are a function of a person's span of attention, quality of memory, disposition (e.g. patient or impatient temperament), inherited biases (e.g. through upbringing or education), and naturally superior or inferior ability to make inferences.

Contra Mercier and Sperber, then, while Hume *does* distinguish between reason and inference, the distinction is ultimately superficial. Rather than follow tradition and elevate inference to reason, he reduces reason to inference. Reasoning just *is* an automatic inference that follows upon the mechanical, spontaneous associations of ideas that are created in the imagination through experience. Since there is another sense in which reasoning is the particular *expression* of ideas already related through association, and since Hume attributes the ability to make inferences to animals as well (withholding from them the ability to express such inferences), there *are* grounds on which to make a distinction between the two ideas. Part of the confusion simply stems from Hume's ambiguity with *reasoning*, which is a well-known problem.[150]

In spite of these difficulties, there are still some noteworthy takeaways from Hume. He issued a firm challenge to the Enlightenment on three important points: (1) the function of reason, (2) the relationship between reason and inference, and (3) the inherent rationality of man. All three of these follow from his radical insistence that the mind operates in accordance with associative principles rather than logical ones. For Hume, error and irrationality are inevitable because association just is how the mind works, and while it is often reliable, it does not discriminate how it makes such associations; it is entirely at the mercy of the stream of appearances in our experiences. In effect, Hume reorients us to the pre-Modern idea that we are somehow fundamentally flawed—just not in the ways that the Christian narrative had expected.

---

[150] See *Enquiry Concerning Human Understanding* (Hume 1999, 14n16). Perhaps Hume himself was aware of this and intended it to be vague? If Buckle's interpretation of him as a modern Academic Skeptic is correct, one could imagine Hume summarizing his thoughts on the matter thusly: "By my own admission, we cannot *know* whether the processes that underlie the operations of the mind are truly automatic; I have merely offered what I take to be the most probable account, based on my observations and experience" (Buckle 2004, 321–322).

8.6 | Fundamentally Flawed or Naturally Adapted?

Considering that more than one hundred years would pass before Charles Darwin would publish his *Origin of Species* in 1859, it is remarkable that Hume's psychology anticipated it in some ways. After decentralizing the role of reason and truth in human mental life, Hume inquires into an alternative, ultimately proposing a theory of mind that relies on non-rational processes that are, by and large, opaque to self-consciousness. Because mental activity is no longer viewed as truth-oriented, Hume advocates for the idea that nature has equipped man with psychological mechanisms that help him survive and succeed in his environments. This is not to suggest that his theory is correct in all of the details, but the general ideas are consistent with contemporary scientific theories. He does, however, severely underestimate just how complex the mind is in its operations.

Hugo Mercier and Dan Sperber defend a highly attractive *modularity thesis of the mind*. Borrowing the concept from biology, they explain that a module is any part of a system that is well-suited for performing a highly specialized task (Mercier and Sperber 2017, 73). Exactly what in particular counts as a module is an open-ended question, leaving the concept itself rather broad and vague. They explain, for instance, that modules can be big or small, wholes or parts, and even biological or behavioral (73–74). As long as something can be identified as having a specific function within certain contexts, it can be counted amongst the modules. A hand, for example, performs the task of grasping objects, and so in that sense, it would count as a kind of grasping module. Notice, however, that the hand also consists of other parts that have their own functions as well, such as fingers, fingernails, tendons, nerves, etc., each of which counts as their

own modules for the functions that they serve. In this case, these would be classified as sub-modules that jointly compose the hand module.

What this thesis entails for the mind is that it itself is a module suited for cognitive tasks, and it (like the hand) is comprised of submodules that perform their own functions, jointly enabling the mind to do what it does. The idea of modularity extends to the neurobiological level as well, where these task-specific submodules have as their basis locations in the brain that activate when such tasks are performed. For instance, Mercier and Sperber point out that the *fusiform face area*, located in the inferior temporal lobe, appears to play a significant role in facial recognition, evidenced by the fact that any damage to this area compromises our ability to recognize faces (Mercier and Sperber 2017, 70). Likewise, next to this location in the brain is the *visual word form area* in the left hemisphere, which activates whenever we recognize letters and words while reading (72). The comprehension of such script during an act of reading does not have its own specialized brain area but involves *both* of these (and other) modules jointly cooperating to accomplish that goal (73). The reading comprehension module, similar to the mind as a whole, is thus comprised of its own collection of several sub-modules that originate in the activity of highly localized neural substrates.

Familiar with the idea of neural modularity, Vilaynur Ramachandran and science author Sandra Blakeslee use the metaphor of a television show to help illuminate it, writing:

> As an analogy, suppose you are watching the program *Baywatch* on television. Where is *Baywatch* localized? Is it in the phosphor glowing on the TV screen or in the dancing electrons behind the cathode-ray tube? Is it in the electromagnetic waves being transmitted through air? Or is it on the celluloid film or video tape in the studio from which the show is being transmitted? Or maybe it's in the camera that's looking at the actors in the scene?

> Most people recognize right away that this is a meaningless question. You might be tempted to conclude therefore that *Baywatch* is not localized (there is no *Baywatch* "module") in any one place—that it permeates the whole universe—but that, too, is absurd. For we know it is *not* localized on the moon or in my pet cat or in the chair I'm sitting on (even though some of the electromagnetic waves may reach these locations). Clearly the phosphor, the cathode-ray tube, the electromagnetic waves and the celluloid or tape are all much more directly involved in this scenario we call *Baywatch* than is the moon, a chair or my cat. (Ramachandran and Blakeslee 1998, 11; emphasis original)

Though the metaphor is now dated, what Ramachandran's example chiefly intends to illustrate is how a collection of smaller, specialized parts in their respective locations can jointly cooperate and produce something that is far greater than the sum of their activities, taken in isolation. The pitfall of the modularity thesis, he cautions, is assuming that there is a specific neural location for every mental capacity, such as "a module for language, one for memory, one for math ability, one for face recognition and maybe even one for detecting people who cheat" (10). Instead, he believes, along with Mercier and Sperber, that "the real secret to understanding the brain lies not only in unraveling the structure and function of each module but in discovering how they interact with each other to generate the whole spectrum of abilities that we call human nature" (11).

The fact that the modularity thesis is consistent with our best current understanding of neuroanatomy is one strong reason to accept its plausibility, but there is another consideration that weighs in its favor as well. Mercier and Sperber believe that it is the most compatible explanation with evolutionary theory. Within the context of biology generally, they write:

> Individual modules are each relatively rigid, but the articulation of different modules provides complex organisms with adaptive flexibility. Could a nonmodular organism achieve a comparable kind of flexibility? It is not quite clear how. Modular systems, moreover, are more likely to overcome a local malfunction or to adjust to a changing environment. Most importantly, modular

> systems have a greater—some have argued a unique—ability to evolve. (Mercier and Sperber 2017, 74)

It is not difficult to see how modular systems are capable of evolving. Philosopher Owen Flanagan, for example, points out that "the feathers of birds are believed to have been initially selected for to serve some thermoregulatory function" (Flanagan 2000, 107). Within the context of modularity theory, feathers would be considered a biological module whose function is to help regulate the animal's body temperature. Similarly, after conducting experiments involving naturally flightless birds and birds whose wings were clipped, biologist Ken Dial proposed the theory known as Wing-Assisted Incline Running (WAIR), which conjectures that the function of flapping is to help with traction across difficult terrain and reach elevated areas, usually to escape from predators or to chase prey (Dial 2003, 403–4). On the modularity thesis, such flapping would be classified as a behavioral module. Both Flanagan and Dial contend that these traits likely played a role in the evolutionary development of proto-birds to the flighted avians known today. It is clear that these modules—as well as other relevant modules, such as hollow bones that allow for increased lung capacity—jointly contributed to the biological fitness of birds by equipping them with the ability to fly away from predators and surprise prey from new angles. What is less clear is whether and how the more complex property of flight could have emerged on its own in a non-modular system.[151]

This adaptive flexibility also carries over to the modularity thesis of mind. Mercier and Sperber clarify: "Another common misinterpretation of modularity is to assume that a modular mind must be quite rigid. It makes no sense to compare the

[151] This is not to suggest that other modules with other functions could not have sufficiently helped make flight possible; rather, this is to wonder along with Mercier and Sperber how a property like flight could have developed independently of modularity.

relative rigidity of individual modules to the flexibility of cognitive systems as a whole (or to the plasticity of the brain as a whole)" (Mercier and Sperber 2017, 75). Like the development of flight, the development of complex cognitive functions in human beings, such as reading comprehension, is a result of the flexibility that follows from the possibility of multiple modules combining their functions and cooperating, as if they were a single module.

This is not to suggest that some particular grouping of modules *always* acts in concert. If that were the case, it would be difficult to parse out the sub-modules from the aggregates that they constitute. Birds, for example, *still* help regulate their body temperature with their feathers by fluffing up or lowering down their torso so that their feathers cover their legs, and they still flap to help gain traction and escape from predators when they are on the ground and unable to fly. They just happen to *also* combine these functions to fly. Instead, it is better to think of modules as a kind of lively ecosystem, some working together with one group and then another, others disrupting that work or even overriding it. As Pinker explains, this kind of modularity allows for both, complex behaviors and complex thoughts:

> The upshot [of a modular mind] is that an urge or habit coming out of one module can be translated into behavior in different ways—or suppressed altogether—by some other module. To take a simple example, cognitive psychologists believe that a module called the "habit system" underlies our tendency to produce certain responses habitually, such as responding to a printed word by pronouncing it silently. But another module, called the "supervisory attention system," can override it and focus on the information relevant to a stated problem, such as naming the color of the ink the word is printed in, or thinking up an action that goes with the word. (Pinker 2002, 40)

If the mind is modular, then why would this create a problem for the Enlightenment attitude? How could this possibly be used to inform Hume's anti-Enlightenment insights?

8.7 | The Known Unknowns: Non-Rational Processes and Unconscious Inferences

Mental modules have the job of extracting and processing information available from the sources on which they work, such as sensory input or even information previously extracted from other modules.[152] But in order for such information to be extracted by a module in the first place, that source must be *contentful*, that is, comprised of information-rich material that can be analyzed, reconstructed, altered, abstracted, etc. Such contentful objects that meet this requirement, argue Mercier and Sperber, are *representations*,[153] information-rich symbolic patterns that are consistently related to things (Mercier and Sperber 2017, 81).

---

[152] There is some ambiguity regarding Mercier and Sperber's use of this term and others related to it. They use the term *mental module* to describe a module that implements "cognitive procedures" without explaining what they have in mind (Mercier and Sperber 2017, 83). Elsewhere, they discuss *cognitive modules*, described as "biological modules having a cognitive function," clarifying that what makes it a biological module is that its function is correlated with a physical state or region of the brain (75). They then elaborate later that "evolved cognitive modules are typically adapted to processing information belonging to a given domain and to drawing specific inferences from it," suggesting that a cognitive module is one that makes inferences (288). However, throughout their work, they discuss *inferential modules*, explaining that they aim "at providing a specific kind of cognitive benefit, and at doing so in a cost-effective way" (165). These cognitive benefits range from simple cognitive reflexes to even producing and updating beliefs about the world (84; 99). There may be some fine-grained distinctions to make between mental, cognitive, and inferential modules, but for the purposes of this chapter, I will be using the term *mental module* to denote any module that is part of a mind and tasked with making an inference from information provided to it.

[153] To be clear, Mercier and Sperber argue that there are mental modules that *exploit*—rather than *represent*— regularities in the environment by executing a procedure in response to a particular kind of stimulus, as in what occurs when a person has a flight response in the presence of a snake (Mercier and Sperber 2017, 87–88). Such exploitation is possible through a combination of sensory stimulation and stored past information (85). In chapter six of their book, however, they draw a distinction between *exploiting* and *representing* empirical regularities, suggesting that a mental module may exploit a regularity without representing anything (90). They insist that these exploitation modules make inferences, and yet earlier they contend that there is "no sensible way to talk about inference and reasoning without using [representation]," relaying an example of how a motion detector can be said to use representations (87; 81–82). The difference seems to come down to the idea that a representation makes information available while an exploitation makes an inference and triggers something, such as a reaction, an action, or even the activation of another module (87). But given their examples and explanations, it must be the case that an exploitation module still *uses* representations even if it does not *produce* them; that is, an exploitation module processes the information available to it, via a representation, in order to trigger the relevant response.

A mental module works on a representation by making an inference from it. Sometimes that inference can be the extraction or reorganization of that information (in effect, the creation of a *new* representation), while other times that inference can be the *use* of that information to initiate a reflex, fixed action pattern, or action (Mercier and Sperber 2017, 90). While there are certainly analogues to Aristotelian practical reasoning, a module need not use any propositions or require the use of any language to function.[154] Thus, Mercier and Sperber believe that Hume was right to distinguish reason from inference, arguing that modules can make all sorts of inferences without needing to reason about anything (51). To better convey this triangular relationship between inference, representation, and modules, the following examples will demonstrate how modules in both living and non-living systems can use representations to make inferences without propositions or reasoning.

First, consider (or reconsider) the well-known case of Pavlov's dogs:

> A bell sounds. Very shortly afterwards, a dog is presented with food. This sequence is repeated some time until the experimenter, Ivan Pavlov, decides to test whether the dog has successfully formed an association between the sound of the bell and the presentation of food, indicated by salivary secretions correlated with the *bell* but not the *food*. When he does so, he finds that the dog responds as expected, salivating at the sound of the bell even without food.

Nonetheless, it seems there is some ambiguity in their use of the term *representation*, for they sometimes treat it more narrowly as *mental* representation, sometimes as *nonconscious* representation, and other times even more broadly as *environmental* representation (understood as an organism's way of relating to its environment such that it treats something in it in a specifically meaningful way, as when one reacts to a snake *as* a threat). While they attempt to draw their own distinctions (e.g. mental, public, cultural, meta, abstract), these uses are not always qualified and, as such, it is at times inconsistent with discussions of representation elsewhere, cited above (Mercier and Sperber 2017, 92–93).

Keeping this in mind, I am therefore going to assume the broader use of the term throughout this section, interpreting it along the lines presented in section 1.5 of this project, summarized in the paragraph above.

154 In particular, it is generally accepted that the conclusion of a practical syllogism can be either a reason that can be used as a premise in a different chain of reasoning or an action. This is at least superficially comparable to the two functions of mental modules: represent or execute.

What is it that causes the dog to form this association?

At the time, Pavlov would argue that the association is a simple conditioned reflex, but Mercier and Sperber insist that this behaviorist explanation (which denies that there are mental states) is inadequate given today's understanding of cognition and animals (Mercier and Sperber 2017, 85). One alternative is to attribute a process of practical reasoning to the dog such that it makes something like a *modus ponens* inference from a set of beliefs it has about bells and food; however, lacking the kind of language necessary for expressing propositions, there is no evidence that dogs do any thinking in terms of propositional logic or Aristotelian practical reasoning (85). Instead, Mercier and Sperber believe this can be explained in terms of the operation of a mental module.

On the modularity thesis, the explanation becomes the following: Because of the number of times bell-ringing and food are consistently correlated in the environment, the dog comes to form an association between the two, allowing a module—whose task is to trigger the physiological processes relevant to the expectation of food (e.g. salivating)—to exploit this relationship whenever the dog has a representation of bell-ringing (Mercier and Sperber 2017, 85). In evolutionary terms, such a module has inestimable use-value in a real-world environment, for it enables animals to *expect* a food source on account of any sensory stimuli that do not *directly* indicate the nearby presence of food but merely *hint* at it. Without a module such as this, creatures would have to rely solely on the direct perception of food, which can make for poor survival odds in environments where food is scarce. While an argument might be made for dogs being able to perform some type of rudimentary reasoning based on their neuronal density, comprehension of simple

linguistic commands, and problem-solving skills, there are two additional examples of modular function that make it clear that many complex tasks can be completed without any reasoning whatsoever.

According to Mercier and Sperber, an alarm system might use different kinds of sensors to help limit the number of false positives (Mercier and Sperber 2017, 81). One sensor might be dedicated to detecting the presence of motion while another might serve the purpose of detecting heat-signatures. Neither sensor, individually, is sufficient to trigger the sound of the alarm; the two sensors must jointly pass an electric signal to a third device, whose task is to activate the alarm when this happens.

This example is meant to be more of a metaphor. While an alarm system may not *actually* make use of representations and inferences—for, as Mercier and Sperber acknowledge, "the whole process could be described in physical terms"—it can illustrate how an organism could make inferences on the basis of representations without any conscious, rational decision-making taking place (Mercier and Sperber 2017, 82). In this case, each device effectively produces representations of the environment insofar as the electric signals from each sensor and the acoustic signal of the alarm make available information about what is happening (81). Specifically, the electric signals each represent, respectively, the presence of something moving or something with a heat-signature, while the acoustic signal represents a high probability that there is an intruder in the area. Yet, the acoustic signal itself is not possible without an inference made from the information provided by both sensors to the alarm-triggering device (81). There is no central system that analyzes all of the information, sorts through it, constructs arguments, and reaches conclusions to sound or not sound the alarm; rather, a number of devices,

each performing very specific tasks, produce or use representations and make inferences. When all three perform their task successfully and in concert, it culminates in a sounding of the alarm.

Might it be the case that mental modules work similarly? Could there be a net result from their activity such that the whole exceeds the sum of the activity of the parts? If the behavior of desert ants is any indication, it seems that both questions can be answered in the affirmative.

Citing the work of biologist Rüdiger Wehner, an expert on *Cataglyphis fortis* (the desert ant),[155] Mercier and Sperber draw attention to a particularly fascinating feat that these insects manage to accomplish (Mercier and Sperber 2017, 54–55). Found primarily in the Sahara Desert, these scavengers comb the desert sands for any bits of food they can find. Although they search in whimsical, wayward fashion for more than 200 meters from their hive, once they locate food, they manage to travel in a relatively straight line back home. Mercier and Sperber explain that Wehner has discovered two different inferential mechanisms that jointly aid the ants in finding their way back to the nest; these mechanisms function in a way similar to the working of the alarm system. The first mechanism is the "navigational toolkit," which uses photosensitivity to track changes in direction from the moment they exit the hive, while the second mechanism, the "odometer," tracks the number of steps taken between changes in direction. Both function to help produce "path integration," an additional inference that determines the direction and distance of the hive (55). All of this occurs without any rational decision-making requiring the ants to draw conclusions on the basis of premises; given their

[155] See "Desert ant navigation: How miniature brains solve complex tasks," *Journal of Comparative Physiology* (Wehner 2003).

neuroanatomy, it is highly unlikely that they are even conscious of these processes at work (121). So how does this happen? It turns out that modularity accounts for their behavior rather well:

> In the brain of the desert ant, for example, the odometer and the compass feed their output to an integrative module that computes and updates a representation of the direction and the distance at which the ant's nest is located, a representation that in turn is used by a motor control module to direct the ant's movements on its way back. (83)

What one finds, then, is that the navigational module uses a representation that provides information about the sun to make inferences regarding the ant's changes in direction. Meanwhile, the odometer module uses representations from tactile feedback to make inferences regarding the number of steps taken. With the information from both of these sources now cognitively available, a third module, the path integration module, analyzes and processes both outputs to make an inference about the location of the nest. Finally, the output from *that* module is used by a fourth to steer the ant back home. Like a cognitive assembly line, the desert ant's remarkably complex behavior is ultimately the product of each module making independent contributions.

These three examples thus help demonstrate how representations and inferential modules can produce highly sophisticated behaviors without requiring any sort of conscious decision-making. Even though reasoning is not used, it does not follow that these processes are not *rational*. It is true that the behavior in two of these cases can be *rationalized*, that is, made to appear *reasonable* from the perspective of the creature performing the activity, given their environmental context and goals. But what is missing—and why we tend to withhold a judgment that they are, indeed, rational—is

evidence that they possess more than task-specific cognitive mechanisms. As Descartes famously argued:

> It is also a very remarkable fact that although many animals show more skill than we do in some of their actions, yet the same animals show none at all in many others; so what they do better does not prove that they have any intelligence, for if it did then they would have more intelligence than any of us and would excel us in everything. It proves rather that they have no intelligence at all, and that it is nature which acts in them according to the disposition of their organs. In the same way a clock, consisting only of wheels and springs, can count the hours and measure time more accurately than we can with all our wisdom. (Descartes, AT VI 58–59)

Surely, one might argue, Descartes is right about this. After all, even if animals or machines use task-specific modules to produce sophisticated behavior or other complex processes, is it not obvious that human beings, by contrast, carefully and methodically think through the problems they are trying to solve, construct arguments, make logical deductions, and, in short, *reason*? Furthermore, while Hume may be right to distinguish reasoning from inference and while the possibility for performing some cognitive tasks, such as reading, might be owed to modular processes, does this not simply weigh in favor of Locke, i.e. that there are two different *kinds* of thought processes in the human mind, spontaneous and rational (or, more specifically, inferential-modular and rational)?

8.8 | The Wizards of Ozymandias

One form of thought that a person can have is known as *intuition*. It is a kind of thought phenomenologically characterized by its sudden and spontaneous burst into conscious awareness. Colloquially, it is commonly referred to as a "gut instinct," often feeling like it is advising one to trust or distrust a new acquaintance or set of circumstances. While it is frequently contrasted with reasoning in folk psychology, the science on intuition has come up short in pinpointing any particular neurobiological underpinnings or locating

any specific cognitive mechanism of intuition. Instead, a consensus is building that it is the *method of delivery* of *other* inferential mechanisms, i.e., it is the way in which such mechanisms make their information available to conscious awareness (Mercier and Sperber 2017, 64).

Part of what defines an intuition is this character of abruptly appearing into conscious thought with no insight into how it arrived. As Mercier and Sperber explain, "When we have an intuition, we experience it as something our mind produced but without having any experience of the process of its production" (Mercier and Sperber 2017, 65). But it is also true that the experience of sense-perception shares this feature insofar as one has little to no control of how perception of the environment enters into conscious awareness. It is true that one can have small, indirect effects on what is experienced by exercising motor control of the eyes and limbs or by directing one's attention to more specific perceptible features like colors or sounds, but in general, sense-perception imposes itself on conscious experience.

To help distinguish intuitions further, then, Mercier and Sperber identify two additional characteristics: (1) a sense that they are somehow justified, that is, a feeling that reasons are *available* to support intuitions if one were to seek them; and (2) a sense of authorship, a recognition that intuitions are one's own (Mercier and Sperber 2017, 66). "While we are not the authors of our perception," they explain, "we are, or so we feel, the authors of our intuitions; we may even feel proud of them" (66). Thus, when a master of intuition like Han Solo pronounces, "I've got a bad feeling about this," he is publicly expressing his intuitive judgment that something is not right. If pressed for reasons, he might respond with facts about the environment or recall things that were said by those

involved—things that he takes to be reasons—that he believes support his judgment. It was not the case that these reasons played any conscious role in forming this judgment, for it seemed to come out of nowhere, and yet, Han recognizes it as both trustworthy and his own. So what is it that distinguishes reason from intuition?

At first glance, it seems that the defining characteristic of reason is *the use of reasons*. In such cases, a judgment is reached in the form of a conclusion that ostensibly follows from a set of reasons, known as the premises, whose role is to justify the conclusion. To relate these reasons as premise to conclusion is to construct an argument, and in doing so, one is normatively obliged to follow the rules of logic to help guarantee the truth of the conclusion reached.[156] Could this not be any more different from intuition? Unlike intuition, this process is slow, deliberate, careful, ordered, and normatively bound in its service to the truth. However, a closer look at reasoning suggests that its difference from intuition might be more superficial.

Mercier and Sperber discuss two different uses of reasons: *retrospective* and *prospective* (Mercier and Sperber 2017, 128–29). A retrospective use of reason occurs whenever one supplies reasons after the fact, much as how we might expect Han Solo to use reasons when asked why he has a bad feeling.[157] If one is reasoning in this manner then one has *already* reached a conclusion, prior to any reasoning whatsoever. This suggests that the function of retrospective reasoning is ultimately to *justify* a conclusion

---

[156] This is not to suggest that arguments are necessarily *presented* in a formal premise-conclusion format. This is obviously untrue. Idiomatic arguments are embedded within paragraphs and speeches containing a surplus of information that is not even relevant to the argument. Rhetorically, it is not uncommon for arguments to *begin* with a conclusion before supplying reasons in support of it afterward. Instead, this claim is meant to be about the *function* of these statements known as reasons: some function in support of the conclusion (premises) and others function in support of a belief or fact about the world (conclusions).

[157] See also the confabulations discussed in section 3.7.

that was reached by *intuition* (145). But even so, one might argue, there is still a difference. "*Real* intuitions," the advocate for reason might say, "lack reasons that support them; as soon as they are supported by reasons, they are no longer intuitions." If this is true, then it ought to be what is found in the paradigm case of reasoning, *prospective* reasoning, wherein one attempts to use reasons to make decisions, act, or arrive at new beliefs (128).

If there is any archetypal sage of prospective reasoning, it is the idealized conception of Descartes, sitting alone in the evening by his fireplace while he carefully and methodically composes his thoughts with an icy objectivity that leaves none of his beliefs immune to circumspection by critical reason. It is not uncommon to imagine one living up to such an ideal, sincerely believing one reaches conclusions through the application of unbiased reason. There are several problems with this image though.

First, prospective reasoning looks superior to retrospective reasoning, but it suffers from the same problem. In both cases, *there are no reasons for our reasons*! Much like any other intuitive judgment, reasons seem to come from nowhere. To be sure, they certainly feel like *our* reasons and they are often accompanied by a desire *for* them, but the *process* that produces them in the mind, like intuition, is utterly opaque to introspection. Mercier and Sperber explain:

> So, how are reasons inferred? By finding further reasons for our reasons? Sometimes, yes; most of the time, no. Assuming that the recognition of a reason must itself always be based on a higher-order reason would lead to an infinite regress: to infer a conclusion *A*, you would need some reason *B*; to infer that *B* is a reason to infer *A*, you would need a reason *C*; to infer that *C* is a reason to infer that *B* is a reason to infer *A*, you would need a reason *D*; and so on, without end. Hence, the recognition of a reason must ultimately be grounded not in further reasoning but in intuitive inference. (Mercier and Sperber 2017, 131–32)

Even if reasons play the justificatory role that supposedly marks the difference between an intuition and a reasoned conclusion, it turns out that the reasons *themselves* stand in need of justification, creating an infinite regress.

But what if one acknowledges this problem? Is it not possible to accept that reasons really are intuitions and yet still contend, along with the advocate for reason, that the difference between reasoning and intuition is not to be found in the formation or justification of any reason *individually*; the justification arises when they are taken *together* in the form of an argument?

This, however, fares no better.

It is true that arguments relate reasons as premise to conclusion with the goal of justifying the latter. If I return home to find that the door is unlocked, for instance, I might begin to form an inductive argument by reasoning:

P_1 If the door is unlocked and I didn't unlock it, then somebody else must be or has been in the house
P_2 The door is clearly unlocked, and to the best of my memory, I didn't unlock it
C Somebody either is or has been in the house

Here I begin with premises in the form of evidence and prospectively reason towards a probable conclusion, the strength of which is dependent on the strength of the premises. Mercier and Sperber contend, however, that the argument as a whole is itself intuited (Mercier and Sperber 2017, 149). How else can premises be related to a conclusion unless they are held together by something? While that something is an argument, to hold an argument in mind is precisely to have a *representation* of it, and such representations are the products of inference, specifically *intuitive* inference in this case. They continue:

> Given that an intuition, simply defined, is a conclusion accepted without attention to, or even awareness of, reasons that support it, your argument as a whole is definitely an intuitive conclusion, an intuition. This intuitive conclusion, however,

> is *about* reasons and about the support these reasons give to a second conclusion, which is embedded in the argument. (149)

Thus, every argument is a representation derived from intuitive inference. That representation itself is one whose content is linguistic in nature, binding together statements on the basis of what looks to follow from what.

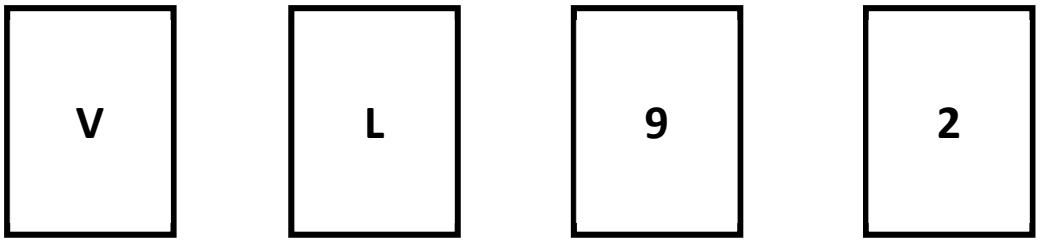
8.10 | Pairing Reasons with Choices

In the battle for an optimistic view of human nature passed down by the Enlightenment, Locke reluctantly acknowledged the tension between the rational and the irrational, concluding that there must be two processes at work within the mind of man, spontaneous and rational, both capable of associating our ideas with one another. The former, he believed, could account for the worse part of our natures, processes that are ultimately influenced through repetition by habit and education. While rational processes may not be able to correct the more entrenched spontaneous ones, reason could emerge victorious in a roundabout way. We could use reason, for example, to create an educational program and culture that is capable of influencing the development of our spontaneous processes in good ways. But recall that Locke left it unclear precisely how rational processes were in fact different from spontaneous ones. Hume, sensitive to this problem, took it upon himself to propose the radical thesis that *all* mental processes were of the spontaneous variety, resulting from habit and chance.

Given the arguments in the previous section, it is beginning to look as though favor lies in Hume's court, where reason too is the result of spontaneous mental processes. But the arguments to this point, however, have relied more on conceptual distinctions (intuition vs. reason) and phenomenological observations. Is there any

objective evidence that helps decisively tilt the scale towards Hume? As a matter of fact, yes. If reasoning really does rely on modules that make inferences behind the scenes, one should find that human beings are unreliable when it comes to introspection. How can someone possibly understand her own motivations if she does not have fuller access to what motivates her?

Mercier and Sperber recount how in 1966 cognitive psychologist Peter Wason invented what is now known as the "Wason Selection Task," a paradigm test for conditional reasoning (Mercier and Sperber 2017, 39). During this experiment, participants are presented with four cards, each with a capital letter on one side and a single-digit number on the other. As presented, the cards are divided in pairs according to which side was up, letter or number. A participant might see something like the following:



With the cards laid out in front of her, an experimenter then presents the participant with a rule, explaining that it applies to all four cards, but it may be true or false:

If a card has a V on one side, then it has a 9 on the other side.

Which cards must a participant turn over to prove that the rule is true or false?

When viewed logically, Wason argued that the correct answers were V and 2, the former representing a *modus ponens* inference that could confirm the rule true while the latter selection, a form of *modus tollens*, could disconfirm it. Yet, one early surprise was that an overwhelming majority of people (around 90%) failed to make both of the right

selections, usually selecting the V and the 9 or just the V (Evans 2005, 174). This lead Wason to speculate that people might have a *confirming* bias, as their selections were intended to prove the rule rather than disprove it (174).

In the 1970s, however, cognitive psychologist Jonathan Evans decided to make some adjustments to the experiment, altering the question so that it tested a participant's ability to *disprove* the rule. In this variation, participants are presented with a rule like the following:

If a card has a V on one side, then it does **NOT** have a 9 on the other side.

When presented in this way, the correct answers are now V and 9. As it turns out, while the modification still treats the flipping of the V card as a *modus ponens* inference, the flipping of the 9 now acts as a *modus tollens* due to the addition of the word "not" in the rule. Suddenly, the majority of people appeared to reason *correctly*. How did this happen?

Evans noticed that people were selecting the *same* cards in both versions of the task, leading him to conclude the existence of a *matching bias* that is "based not on logical reasoning but on intuitions of relevance: they turn over the cards that seem intuitively relevant" (Mercier and Sperber 2017, 43). What does this mean for introspection and self-understanding? Mercier and Sperber summarize:

> In the Wason four-card selection task, participants, before they even start reasoning, make an intuitive selection of cards. Their selection is typically correct in some versions of the task and incorrect in others, even though the problem is logically the same in all versions. Asked to explain their selection, participants have no problem providing reasons. When their choice happens to be correct, the reasons they come up with are sound. When their choice happens to be mistaken, the reasons they come up with are spurious. In both cases—sound and spurious reasons—these are demonstrably rationalizations after the fact. In neither case are participants aware of the factors that, experimental evidence shows, actually

> drove their initial selection (and which are the same factors whether their answer is correct or mistaken). (115–6)

But it does not stop here. Other experiments based on real-world scenarios also support the idea that introspective *ex post facto* rationalizations of oneself are unreliable.

After the murder of Kitty Genovese in 1964, psychologists John Darley and Bibb Latané set out to investigate why dozens of bystanders who were in a position to help ultimately did nothing.[158] Their work would lead to the discovery of *the bystander effect*, which holds that there is an inverse correlation between the number of people able to help during a situation that calls for it and the number of people who actually help (Mercier and Sperber 2017, 116). This phenomenon, like the Wason Selection Task, proved that people are oblivious to the causal factors that influence decision-making, and this phenomenon has been confirmed in subsequent psychological experiments (116).

In fact, Mercier and Sperber cite one experiment in particular, lead by Latané as well as philanthropist Judith Rodin, in which participants were greeted by an assistant

---

158 Though this is how it is commonly recounted—thanks to the reporting of Martin Gansberg and his executive editor, Abe Rosenthal, both of the *New York Times*, on March 27th, 1964—the *New York Times* has since published two follow-up articles that explain how the coverage of the original story was mired in inconsistency and hyperbole. These inconsistencies have also been the topic of a documentary created by Kitty's brother, Bill Genovese, and director James Solomon, entitled *The Witness*. Because the misreported event inspired important psychological research and because it conveys an important moral, perhaps the best way to view it is, as Genovese and Solomon suggest, as a parable.

For the original coverage, see Martin Gansberg's "37 Who Saw Murder Didn't Call the Police; Apathy at Stabbing of Queens Woman Shocks Inspector," *New York Times*, March 27, 1964. URL = https://www.nytimes.com/1964/03/27/archives/37-who-saw-murder-didnt-call-the-police-apathy-at-stabbing-of.html.

For the updated coverage, see Jim Rasenberger's "Kitty, 40 Years Later," *New York Times*, February 8, 2004. URL = https://www.nytimes.com/2004/02/08/nyregion/kitty-40-years-later.html.

See also David Dunlap's "1946 | How Many Witnessed the Murder of Kitty Genovese?," *New York Times*, April 6, 2016. URL = https://www.nytimes.com/2016/04/06/insider/1964-how-many-witnessed-the-murder-of-kitty-genovese.html; and Stephanie Merry's "Her shocking murder became the stuff of legend. But everybody got the story wrong," *Washington Post*, June 29, 2016. URL = https://www.washingtonpost.com/lifestyle/style/her-shocking-murder-became-the-stuff-of-legend-but-everyone-got-the-story-wrong/2016/06/29/544916d8-3952-11e6-9ccd-d6005beac8b3_story.html.

who handed them a questionnaire before returning to her office and closing the door behind her. From the room in which they were situated, the participants could hear the assistant moving about until eventually there was a loud crashing sound followed by cries for help. When this occurred, sometimes the participants were totally alone in the waiting room, and other times, they were accompanied by another person who happened to be a confederate in on the experiment. What Latané and Rodin observed is that whenever people were alone, they attempted to help 70% of the time, but when they were with another, that number dropped to an astonishingly low 7%. At the conclusion of the experiment, the participants were asked about their choices, and they seemed to have very reasonable explanations for their decisions to help or not. However, as Mercier and Sperber relay, "When asked whether the presence of another person in the room had had an effect on their decision not to help, most were adamant that it had had no influence at all," even though the truth is that "we know that they had been ten times more likely to intervene when they were alone in the room than when they were not" (Mercier and Sperber 2017, 116–17).

It thus appears that we really are out of touch with the many variables that influence our decision-making, swayed by hidden motives that prove opaque to introspection. Oblivious to the information that factors into a decision, we intuitively conjure up reasons that make these decisions look plausible to anyone who inquires into our behavior.

8.11 | Conclusion

Locke was not unaware of some of the issues that arise when we privilege our rationality, going so far as to make it constitutive of being human. He could very clearly see how his

Enlightenment faith needed to address the problems of error and madness. His solution, however, only resulted in paradox: reason is powerful and autonomous, except when it is not, when rational processes are disrupted by mechanical, spontaneous processes. With the luxury of Locke's efforts behind him, Hume took up his philosophical project, attempting to address the paradox by collapsing rational processes into spontaneous ones. The end result was a theory of mind that centralized the role of inference and associative processes.

Although Hume's insights into the nature of inference are still valuable today, it is the modularity theory of mind that succeeds in explaining reason's operations and its failures. Locke was right to conclude that non-rational processes exist, but, contrary to Locke and the Enlightenment attitude, reason looks more and more like just another form of intuitive inference that is as strange and mysterious as any other, operating in the background and producing effects that are consciously experienced.

What sets reason apart from other mental modules is that it appears to operate on specifically *linguistic* representations whose contents happen to be statements that play a justificatory role. Thus, when reason is employed, it is primarily aimed at *justifying* our behavior rather than *producing* it. While it subjectively feels as though reason is transparent, reliable, and trustworthy, there is plenty of evidence suggesting the contrary. It is precisely these considerations that have lead Mercier and Sperber to develop their interactionist model of reason, which when coupled with the modularity thesis of the mind, entails the possibility that reason itself might simply be just another inferential module, one that evolved in response to social and pragmatic pressures to build and manage one's reputation (Mercier and Sperber 2017, 133; 142–43). Regardless of

whether or not reason really is a module or results from the activities of several modules, it is clear that the modularity thesis of mind best accounts for "error and madness." In the next chapter, we will also see how it can correct some critical, false assumptions that Enlightenment thinking makes about the mind.

Chapter 9: From One to Many

The fact should occasion no surprise.
In the case of poisons,
the effect of which they are the cause is a gross and obvious phenomenon
and the level of causal explanation involved in simply calling a substance "a poison" is crude and simple.
But in the case of mental states,
their effects are all those complexities of behavior
that mark off men and higher animals
from the rest of the objects in the world.
Furthermore,
differences in such behavior are elaborately correlated with differences in the mental causes operating.

— David Malet Armstrong, "The Nature of Mind"

9.1 | Wholly, Somewhat, and Not-at-All Unintelligible

Even though Donald Davidson embraced the idea that we are fundamentally rational, he would eventually come to recognize the inescapable fact that we have irrational tendencies, leading him to wonder how this might best be situated against a supposed background of rationality. Recall that for Davidson, to be rational is to be the sort of creature that is sensitive to reasons, able to act *for* reasons and be open to accountability in so doing. On this view, then, to undertake any action is to be motivated by an *intention* (a belief and a desire that causes us to act), and by identifying such intentions, observers can make behavior appear reasonable from the perspective of the agent. By engaging in this practice, we *rationalize* actions, and this, Davidson believes, is what it means to make our behavior intelligible.

The idea that behavior is intelligible rests on the supposition that there is a consistent correlation between what we believe and what we do. Davidson remarks, "The reasons an agent has for acting must, if they are to explain the action, be the reasons on which he acted" (Davidson 1982, 173). The same types of actions must follow from the

same types of beliefs in an almost lawlike fashion. It is this lawlike connection, after all, that makes normativity possible, for this is what allows us to form expectations of one another, demand justifications for what we say and do, and hold each other accountable by keeping track of our behavior and reasons. These practices form the bedrock of human society, one rooted in social interaction and communication. Were our reasons to fail to make reasonable what we say and do, it is unclear how normativity can get off the ground, and any attempts to rationalize behavior under such conditions would make for exercises in futility.

And yet, the existence of irrationality, especially if it proves pervasive, threatens to abolish this nigh law of nature, thereby severing the supposed connection that holds between reasons and actions and undermining the consistency required for our normative practices to be meaningful. Against our hopes and dreams of understanding the world around us, it is irrationality that taunts us with the possibility that at least some of our behavior is *unintelligible*, perhaps even hopelessly so. Why would someone believe something to her own detriment on the basis of little more than wishing it so? What is deceiving oneself supposed to accomplish when the fear being avoided only makes one's life *more* miserable? How does the lack of self-control engendered by akrasia prove advantageous, especially when one *continues* putting herself in the same tempting situations over and over? Wherein lies the reasonable motive for creating a state of affairs only so one may attempt to undo it, as is characteristic of negation? The better reason, we want to believe, is the *stronger* one, the one that *will* persuade, and yet it so often seems to fail. How is this possible?

9.2 | A Different Kind of Hard Problem

To begin to understand this, we must first make a distinction between *internal* and *external* irrationality. What defines internal irrationality is a *psychological* conflict that can manifest itself at the level of behavior, ultimately thwarting a person's goals. Like Sisyphus of Greek mythology, we act internally irrational whenever we create impossible success conditions and engage in our own self-defeating tasks, incurring no advantages for doing so. The akrates who perpetually puts herself in a bad position, for example, is victim to this form of irrationality.

External irrationality, by contrast, is defined by *interpersonal* conflict, specifically a conflict between what one does and some particular set of norms that are external to the agent. The Davidsonian philosopher and psychoanalyst Marcia Cavell says, for example, that if a person were to believe "that the world was formed from worms and cheese," while it may not create any internal inconsistency between her beliefs, we can conclude that it is a case of *external* irrationality since it fails to be consistent with what we scientifically and socially believe about the origins of the world (Cavell 1993, 193). This form of irrationality has a social dimension, requiring the existence of a community with its own social norms against which an agent's beliefs and behavior can be compared.[159]

For someone like Davidson, external irrationality challenges us to figure out where *actual* irrationality begins and mere disagreement ends. Does every violation of a norm count as acting irrationally, or just some? He concedes, "No doubt we very often stigmatize an action, belief, attitude, or piece of reasoning as irrational simply because we disapprove, disagree, are offended, or find something not up to our own standards"

[159] This also includes speech as a form of behavior.

(Davidson 1985a, 189). Precisely when is it fair to charge someone with being irrational in this sense? What happens when we allege that someone is acting irrationally even though her actions make sense from her perspective? Davidson explains:

> Many might hold that it is irrational, given the dangers, discomforts, and meagre rewards to be expected on success, for any person to attempt to climb Mt. Everest without oxygen (or even with it). But there is no puzzle in explaining the attempt if it is undertaken by someone who has assembled all the facts he can, given full consideration to all his desires, ambitions, and attitudes, and has acted in the light of his knowledge and values. (Davidson 1982, 170)

These are of course important questions to ask, especially when it comes to deciding what it means to be a normative creature and when it is permissible to depart from norms, but Davidson's interest in these questions are secondary to what he believes are more pressing concerns. What has his attention are not problems of normativity but problems of *psychological conflict*. "It is not clear that there is a genuine case of irrationality," he writes, "unless an inconsistency in the thought of the agent can be identified, something that is inconsistent by the standards of the agent himself" (Davidson 1986, 199). To appropriate a phrase from philosopher David Chalmers, this is precisely what constitutes the "hard problem" for action theory and rationality.

The hard problem states: if we are fundamentally rational, how is (internal) irrationality possible at all? The question is difficult enough in isolation, but recall that according to traditional action theory, the reasons for which we act are the sorts of things that *cause* us to act. If this is true, then how are we to understand irrationality? On this view, it appears that if something *other* than a reason causes an action—perhaps some kind of neurobiological malfunctioning due to a brain abnormality—then it is by definition *nonrational*, and on the other hand, if it is a *reason* that is causing the action after all, then it is by definition *rational*. These concerns were raised at the end of part

two, and they invite us to ask: Where is the room for the irrational? Not only are the stronger reasons failing us, but something is causing these failures. How can we explain this?

The hard problem is one that originates in the Enlightenment's conception of the mind, and it continues to plague us to this day. It turns, in my estimation, on two Cartesian assumptions that need to be reevaluated: that rationality is the essence of the mental and that the mind is a seamless unity. In the Second Meditation, for example, one can find the inspiration for the idea that the mind *just is* rational. There, Descartes argues that he exists as a *thinking thing* (*res cogitans*) before deploying a series of terms that he treats interchangeably, "I am a mind, or intelligence, or intellect, or reason—words whose meaning I have been ignorant of until now" (Descartes, AT VII 27). Likewise, in the Sixth Meditation, he insists that the mind is indivisible, completely distinct from the body (AT VII 86). When embraced, both of these assumptions together create a paradox for irrationality, for they imply that the mind is wholly transparent, a view that Descartes himself endorses. He insists in his replies to his readers that this is plain and obvious:

> As to the fact that there can be nothing in the mind, in so far as it is a thinking thing, of which it is not aware, this seems to be me to be self-evident. For there is nothing that we can understand to be in the mind, regarded in this way, that is not a thought or dependent on a thought. If it were not a thought or dependent on a thought it would not belong to the mind *qua* thinking thing; and we cannot have any thought of which we are not aware at the very moment when it is in us. (VI 246–7)

Davidson rightly questions the Cartesian assumption of the unity of the mind by reviving the Platonic-Freudian idea that the mind is divided, but he is reluctant to abandon the assumption of our inherent rationality, fearing that it invites senselessness and behavioral anarchy by undermining the significance of norms.

9.3 | Returning to the Park

In order to contextualize the hard problem for Davidson, let us turn once more to the case of Mr. S, discussed in chapter seven. As mentioned in section 7.8, it is important to keep in mind that Davidson's use of this example diverges from Freud's own anecdote, for his interest lies in understanding the irrationality of akrasia (Davidson 1982, 172, n.3). He summarizes the case thusly:

> A man walking in a park stumbles on a branch in the path. Thinking the branch may endanger others, he picks it up and throws it in a hedge beside the path. On his way home it occurs to him that the branch may be projecting from the hedge and so still be a threat to unwary walkers. He gets off the tram he is on, returns to the park, and restores the branch to its original position. (172)

At first glance, there is nothing particularly irrational about this. At best, it could constitute an example of external irrationality if it turns out that Mr. S violated some social norms in his moving of the branch or returning to the park. Even as far as internal irrationality goes, Davidson believes little can be said of the matter, for "everything the agent does (except stumble on the branch) is done for a reason, a reason in the light of which the corresponding action was reasonable" (172-3).

To turn this into an example of internal irrationality, some psychological conflict must be introduced, and so Davidson proposes modifying it accordingly: imagine that Mr. S realizes that his desire to return to the park for the sake of moving the branch is nonsense. This is not to suggest that he has *no reason* to return; it is just that the reasons for *not* returning, "the time and trouble it costs," are judged to be better than the reasons for returning (Davidson 1982, 174). But why would Mr. S act against his better judgment? More specifically, why would he *intentionally* act against his better judgment?

As intriguing as these questions may seem, even with this additional modification, Davidson does not believe there is any difficulty in explaining this predicament quite yet, for he does not think that an individual is necessarily irrational for acting contrary to her better judgment. These better judgments—otherwise known as *all things considered* (*atc*) judgments—are evaluative in nature, and as such, they are not reached by practical syllogisms. After viewing the arguments in light of all of the available reasons, the job of an *atc* judgment, as explained in chapter six, is to issue a conclusion that recommends one argument as better relative to the others. Intentions are thus not part of the syllogisms that result in *atc* judgments, and so such judgments do not necessitate that we act in accordance with them.

Indeed, if Mr. S is to act irrational, then something further still is required. What is needed, thinks Davidson, is for Mr. S to hold the belief that *atc* judgments are worth acting on in the first place. To have such a belief is to affirm the *principle of continence*—the principle that one ought always to act in accordance with *atc* judgments—and so if a person avows this principle, then she has every reason to act accordingly. If Mr. S believes this, then it is in intentionally acting against this principle wherein the psychological conflict obtains. But why might someone act against this principle for *no* reason? How is that even possible?

To these questions, Davidson believes traditional action theory has no good answer. The usual strategy is to either deny the existence of irrationality or else reduce it to the nonrational, to argue that a person does not *really* believe that the alternative course of action was better or to argue that some kind of mechanical dysfunction caused the person to act otherwise. Davidson, however, believes that traditional action theory

can be salvaged *and* irrationality explained by forging an alliance with an unlikely hero: Freud.

9.4 | 1915: A Year of Freudian Excursions into Unknown Territory

It is not an uncommon philosophical strategy to posit a divided mind whenever serious psychological conflict can be detected. The most famous historical example of this in western thought might be Plato's tri-partite theory of the soul that he advances in Book IV of the *Republic*. This was discussed briefly in section 7.9, where I explained how the character of Socrates uses the Principle of Opposites to argue that a person cannot both have a desire for and aversion to one and the same thing, in the same respect, and at the same time. Clearly, however, people *do* have conflicting motivations, and so rather than reject the Principle of Opposites, Socrates proposes that the motivations must be stemming from different parts of the psyche (*Republic*, 436B–441C).

Perhaps the most vocal defender of the divided mind in recent history, however, is Freud, who begins "The Unconscious"—the third of three essays on the topic published in 1915—with some observations about human thought and behavior before issuing the following exhortation against the Cartesian assumptions of the unity and transparency of the mind: "We then have to take the view that it is nothing less than an *untenable presumption* to insist that everything occurring in the psyche must also be known to consciousness" (Freud 1915c / 2005, 50; emphasis original). It was through a culmination of clinical experience and psychoanalytic theory that Freud had become convinced of the divided mind, providing both empirical support and philosophical arguments in defense of it.

Although Hume was among the first Enlightenment thinkers who had taken notice of the fact that unwanted ideas tend to spontaneously insert themselves into our patterns of thought,[160] leading him to believe that it was evidence for the associative theory of mind, Freud's observations went further, arguing that this phenomenon of disrupted thought in fact points to something much stranger about the human mind. From his perspective, we have been focusing on the wrong feature in our mental landscape, always trying to understand the nature of our *conscious* thought processes. By contrast, Freud argues, these disruptions—unwanted thoughts, slips of the tongue, dreams—should be piquing our interest in what we are *not* seeing, namely, the fact that these sorts of things are "of unknown origin and the results of thought processes whose workings remain hidden from us" (Freud 1915c / 2005, 50). It is these things we must understand better if we are to understand ourselves at all. Against Descartes, he continues:

> All of these conscious acts [the disruptions] remain incoherent and incomprehensible if we insist that everything occurring in our psyche must also be experienced through consciousness, whereas they fall into demonstrable patterns if we interpolate unconscious acts that we have inferred. (50)

Like Plato, Freud concluded that positing a divided mind is explanatorily necessary,[161] but unlike Plato, such necessity went beyond even just psychological conflict, extending the theory to help account for even the simplest and most common acts of thinking, including dreaming and recall from memory (51). The mental landscape is vast, and consciousness merely serves the role of perceiving a *part* of it, just as our eyes perceive

---

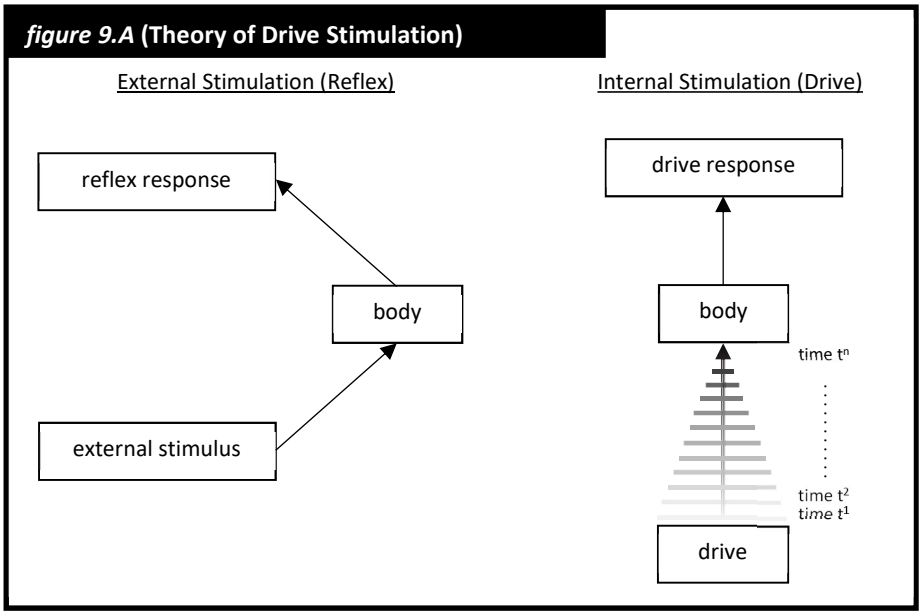
160 See sections 8.4 to 8.5. See also *An Enquiry Concerning Human Understanding*, 3.1.

161 Lear, for instance, remarks how Freud even appears to use something like the Principle of Opposites in support of the unconscious, noting the strangeness in cases of melancholia where sufferers appear to be both judges and judged as they deliver and absorb intense self-criticisms. This leads Freud to argue for the psychological division between *ego* and *superego* (Lear 2005, 168).

only a part of the environment around us (54). At the heart of the unconscious, what gives it its peculiar character, Freud believed, is a nexus of independent drives, each with its own aim, object, and motivational force (or what Freud calls *pressure*, i.e., the intensity of a drive).

Drawing upon it to make sense of his clinical work, the concept of a drive was new for Freud, one that was in need of theoretical fine-tuning, and to elucidate the idea, he sought inspiration from biological mechanisms. Beginning first with the idea of a reflex sketched in the broadest possible terms, in his first essay of 1915 ("Drives and Their Fates"), Freud discovered a conceptual framework that aided him in understanding the drive. A reflex, he believed, can be construed as a kind of behavior directed towards the environment in response to an external stimulus applied to a living body (Freud 1915a / 2005, 14). Drives, likewise, seem to motivate behavior and compel actions, as if there were some kind of *internal* stimulus present, one that is physiological in origin but that influences the psychological as well. Like the reflex, this internal stimulation demands a response, specifically, a satisfaction of a need whose intensity varies in proportion to the amount of time elapsed since it was last gratified (14; 16). Left unsatisfied, the appetite from a drive can become ravenous, strongly inclining an individual to act in the service of its gratification (see figure 9.A).

Freud believed that the two types of drives whose existence seemed most likely (and to which other drives might be reduced) are *self-preservation* and *sexual* drives (Freud 1915a / 2005, 18).



These two drives are diametrically opposed to one another insofar as self-preservation is oriented towards the individual or *ego* and sexuality is oriented towards the *other*. Again seeking inspiration from biology, Freud explains:

> Biology informs us that sexuality is distinct from the other functions of the individual because its purposes—the creation of new individuals, and thus the preservation of the species—transcend the individual. It also shows us that two equally valid ways of conceiving of the relationship between ego and sexuality exist side by side, one in which the individual is paramount, with sexuality regarded as one of its activities and sexual satisfaction as one of its needs, and another according to which the individual is a temporary and transient adjunct to the quasi-immortal germ plasm entrusted to it in the process of reproduction. (19)

Thus, long before Dawkins advanced the idea of a conflict between selfish gene and selfish individual,[162] it was Freud who was all too keenly aware of not only the existence

[162] In chapter thirteen of *The Selfish Gene*, Dawkins begins his first two paragraphs summarizing the problem that emerges from his theory before advertising a possible resolution in his sequel, *The Extended Phenotype*:

> An uneasy tension disturbs the heart of the selfish gene theory. It is the tension between gene and individual body as fundamental agent of life. On the one hand we have the beguiling image of independent DNA replicators, skipping like chamois, free and untrammeled down the generations, temporarily brought together in throwaway survival machines, immortal coils shuffling off an endless succession of mortal ones as they forge towards their separate eternities. On the other hand we look at the individual bodies themselves and each one is obviously a coherent, integrated, immensely complicated machine, with a conspicuous unity of purpose. A body doesn't *look* like

of such a conflict but its profound effects on our behavior and psychic health. Even more impressive, Freud believed that the conflict engendered by these two very different drives could sometimes be resolved by the *same activity*, a phenomenon he referred to as *overlapping* (17). In sexual activity, as noted in the quote above, the needs of both, selfish individual and selfish gene, are able to achieve joint satisfaction through pursuit of the same object and aim: stimulation to completion of the sexual organ (19; 20).

According to Freud, drives never enter into conscious awareness but are only inferred from their objects and aims (Freud 1915c / 2015, 59). Such inferences are made possible by the fact that, early in our developmental phase, drives become associated with psychological representations of the objects that satisfy them, a process known as *fixation*, and after this occurs, the connection between the drive and its object becomes so strong that it is almost immutable (1915a / 2015, 17). If an infant derives pleasure and gratifies the self-preservation drive by suckling on her mother's breast for nourishment, this drive might fixate on a psychological representation of breasts as a way to satisfy a form of hunger, and if the sexual drive comes to overlap with the self-preservation one, the representation of the breast might serve to satisfy sexual hunger as well. While this

---

> the product of a loose and temporary federation of warring genetic agents who hardly have time to get acquainted before embarking in sperm or egg for the next leg of the great genetic diaspora. It has one single-minded brain which coordinates a cooperative of limbs and sense organs to achieve one end. The body looks and behaves like a pretty impressive agent in its own right.
>
> In some chapters of this book we have indeed thought of the individual organism as an agent, striving to maximize its success in passing on all its genes. We imagined individual animals making complicated economic 'as if' calculations about the genetic benefits of various courses of action. Yet in other chapters the fundamental rationale was presented from the point of view of genes. Without the gene's-eye view of life there is no particular reason why an organism should 'care' about its reproductive success and that of its relatives, rather than, for instance, its own longevity.
>
> How shall we resolve this paradox of the two ways of looking at life? (Dawkins 2006, 234; emphasis original)

might seem to contradict the idea that self-preservation drives and sexual drives are in conflict with one another, Freud makes it clear that in their initial stages of development, sexual drives are "dependent on self-preservation drives and become detached from these only gradually; when finding an object, they also follow the paths laid down by the ego drives" (20).

But because there are drives that associate themselves with ideas, even if the drives themselves must remain hidden, does it follow that the representations (i.e. the ideas) with which they associate need necessarily be unconscious as well? How do such ideas stay out of sight? Freud is not exactly clear on this although his theory revolves around the idea of *repression*, that somehow unconscious activity and ideas can be kept from conscious experience.

9.5 | In a War of the Worlds, the Oppressive Regimes Win

One common way to interpret Freud's theory of repression is to posit the existence of a censor that stands at the gates between the conscious and unconscious, permitting some of that activity to enter into conscious thought while rejecting the rest. This interpretation is famously the target of Jean-Paul Sartre's criticism that Freud's theory paradoxically requires a conscious unconscious. According to Sartre, the idea of a censor seems to require that it has an awareness of what needs to be repressed, an awareness of the purpose of repressing, and an awareness of the truth against which it compares the intruding ideas, as if it were a bouncer checking IDs at the door (Sartre 1943 / 1956, 209).

Though Sartre has in mind Freud's later theory (which uses concepts like *ego*, *id*, and *superego*), his criticism can still fairly be leveled against Freud's earlier writings. In "The Unconscious," for example, Freud posits that a halfway house exists between

unconscious contents and conscious ones known as the *preconscious*, consisting of everything that is capable of being conscious without resistance (Freud 1915c / 2015, 56). Quite effortlessly, you may start thinking about whether you would like to enjoy a steak later in the day—such an idea is *capable* of entering conscious thought. But, believed Freud, there are other ideas, such as maybe a fear of impotence deriving from your mother's lack of affection, that cannot be entertained as easily, not without strong resistance. The preconscious thus acts as a kind of foyer to the conscious living quarters, and anyone wishing to enter the foyer is "subject to a kind of inspection (*censorship*)," that determines whether they must remain outside of the house or are permitted to enter (56; emphasis original). Freud even speculates that another kind of censorship might theoretically vet the preconscious guests before allowing them to enter the conscious living quarters, as if there were not one but two doormen (56).

Complicating matters even further is the fact that Freud employs similar metaphorical language in "Repression" when describing the relationship between the unconscious and the conscious:

> The distinction [between removing an unwanted idea from conscious thought and keeping it from entering into conscious thought altogether] is not important; it amounts more or less to the difference between throwing an unpleasant guest out of my drawing-room or hallway, and, having recognized who it is, not letting him through the front door at all. (Freud 1915b / 2015, 41)

And in a footnote to this passage, he extends the metaphor in a way that supports Sartre's understanding:

> This useful analogy for the process of repression can also be extended to include a characteristic of repression mentioned earlier. I just need to add that I also have to put a permanent guard on the door that I have forbidden the guest to enter, otherwise he would force it open. (45, n.1)

In trying to understand the mechanics of repression, then, it certainly seems like Freud requires an intelligent censor, a homunculus who can stand guard at the doors of conscious thought. And yet puzzlingly, elsewhere in these same writings, Freud explicitly remarks on the absurdity of a conscious unconscious. After proposing the idea that unconscious contents might best be understood *as if* they belonged to another mind—insofar as, from an analyst's perspective, they resemble "ideas, aspirations, resolutions, and so on"—Freud cautions that there are important differences to consider:

> A consciousness unknown to its own bearer is something quite different from another person's consciousness, indeed it is questionable whether such a consciousness—devoid of its own most important characteristic—even merits further discussion at all. Anyone rejecting the postulation of a psychic unconscious will certainly not be content to accept an *unconscious consciousness* in its place. (Freud 1915c / 2015, 52; 53; emphasis original)

There is thus reason to presume that Freud's metaphors of guards and censors should not be taken too seriously, undercutting much of Sartre's criticisms. But if this is right, how then is repression supposed to work?

A second popular interpretation has been to argue that repression occurs as a response to intense anxiety, sometimes even construing this as an unconscious yet somehow *intentional* process.[163] Philosopher Bill Hart, who endorses this reading of Freud, argues that the only way it can avoid Sartre's criticism is by taking Freud's economic theory of the mind seriously, that there is a finite amount of psychic energy to go around and it takes a lot of energy to make an idea conscious (Hart 1980, 201). By investing that energy into some ideas, we can repress others as a side-effect, even

[163] Lear, for example, partially supports this interpretation except he argues that repression can be viewed as a simple mechanism that manages anxiety without imputing any sense of agency to it, musing that perhaps someday neuroscientists will have a better explanation for how it works (Lear 2005, 64). Elsewhere, he explicitly states that he believes it requires neither an intelligent censor nor intentionality (108).

intentionally, just as we might intentionally avoid looking at someone in a public space by directing our attention elsewhere. We might, for instance, gaze at some grand advertisement as we walk down the street. The person we are trying to avoid need not even *actually* be there; rather, it is the anxiety created by their *potential* presence that motivates our looking away.

In support of this reading, Freud postulates that a drive can ramp up in psychic energy the longer it goes unsatisfied, and it seeks out gratification by investing psychic energy into an idea, ideally to motivate us into undertaking an appropriate course of action that can gratify it. Even when we are not consciously aware of the demands from these drives, they can still reach satisfaction from what we do, in turn creating a state of pleasure. The problem, however, is that what satisfies one drive might frustrate another, and depending on the displeasure that such frustration creates, we might be inclined to disrupt the satisfaction of the first drive before it has the opportunity to frustrate the aim of the second. Freud, for instance, identifies hunger as a drive stimulus that is incapable of being repressed (presumably a stimulus associated with a self-preservation drive), and while it is true that it cannot be *permanently* repressed,[164] a person might temporarily ignore her hunger by directing her attention elsewhere, especially if doing so proves advantageous to attracting a mate for a sexual encounter, a behavior that makes sense if gratifying the sexual drive is a priority (Freud 1915b / 2015, 36).

---

164 Freud rejects the idea that repression has permanent effects anyway. He writes, "We should not think of the process of repression as a single event with permanent results, as when, say, a living thing is killed and from then on remains dead; repression demands, rather, a constant expenditure of energy and would be undermined if this were relaxed, necessitating a renewed act of repression" (Freud 1915b / 2015, 39). He believes that such energy is relaxed, for instance, during sleep, which is why the repressed contents can make their way into our dreams.

There are two more Freudian ideas that lend weight to this interpretation of repression: *counterinvestment* and *ruling impulse*. Because a drive demands satisfaction and is always at work, albeit at varying intensities, the idea that represents the object on which the drive has fixated is always pressing up against the doors of consciousness. What causes it to be repressed, argues Freud, is some kind of counterpressure of equal or greater force in the opposite direction. The implications this has for his economic model are clear: "Maintaining a repression, then, requires a constant expenditure of energy, whereas lifting it represents, in economic terms, a saving" (39). Consistent with the second interpretation of repression, Freud explains that what is present to consciousness is the result of an *overinvestment* of psychic energy, and so whatever ideas appear in conscious thought are those that are, for lack of a better term, supercharged (Freud 1915c / 2015, 76). In other words, it is psychically costly to hold ideas in conscious awareness, and if there is a finite amount of psychic energy to go around, it follows that promoting or sustaining one set of ideas in conscious thought necessarily deprioritizes and divests some other, unwanted set of ideas. Thus, repression can and does occur indirectly. But what is it that determines which ideas get repressed and which not, if not a censor?

Recall that Freud believes that the drives themselves are never the objects of conscious awareness; instead, they are merely inferred from their effects on thought and behavior. One of the functions of the conscious system that Freud identifies is that it can influence behavior through the activation of motor skills. He observes that when an idea is repressed, it is not only withheld from conscious awareness, "but also from emotional development and muscle activation" (Freud 1915c / 2005, 60–1). What is repressed, therefore, is not just the unwanted *idea* but also the *possibility* for its satisfaction by

inhibiting any behavior that might act in the service of the drive. Why would somebody want to do this?

It turns out that if drives really do compete with one another (as suggested in the previous section), then one obvious reason is that, in fighting for its satisfaction, one drive might interrupt another that conflicts with it. Not only is there room for this possibility within Freud's theory, but Freud himself suggests in several places that a drive can "rule" the conscious system. First, he explains that the content of the conscious system is supplied from two sources: perception and the "life of the drives" (Freud 1915c / 2005, 76). While it is not clear what he means by the phrase, a few paragraphs earlier, he writes:

> Later we shall learn that the process of becoming conscious is restricted by certain aspects of the manner in which attention is directed. There is no simple relationship, then, between consciousness and the various systems, nor between consciousness and repression. The truth is that not only the repressed remains alien to consciousness, but also some of the impulses governing our ego, i.e., that which, in terms of function, stands in the starkest possible contrast to the repressed. (75)

While Freud here reaffirms the role that selective attention plays in causing repression, this also seems to suggest that the conscious ego itself is directed in its activities by some drive impulses without our realizing it.

As a final consideration of the relationship between drives and conscious thought, in what sounds like a restatement of the phenomenon of overlapping, Freud reiterates how repressed impulses can still find satisfaction by working with the ruling impulses:

> Co-operation can occur between a preconscious and an unconscious—even a powerfully repressed—impulse if a situation arises in which the unconscious impulse can work in tandem with one of the ruling impulses. Repression is lifted in this instance and the repressed activity is permitted as a reinforcement of the one intended by the ego. (Freud 1915c / 2005, 77)

What emerges in Freud then is a complex psychological picture in which conscious activity is orchestrated by a ruling drive impulse that is in competition with a host of other impulses for its gratification, repressing like a tyrant anything that conflicts with its aims. Because there is a finite amount of psychic energy and what is conscious is the most costly, other drives constantly work to divest that energy and take a foothold in the conscious system themselves. When denied entry into consciousness, they often resort to a degree of roguishness by disguising themselves as other ideas and intentions, hoping to find some kind of gratification by any means necessary.

Like a psychological "game of thrones," the results of this ongoing war—wherein drives fight for satisfaction, repress the competition, forge temporary alliances, compromise, and infiltrate the conscious system through guile in an effort to depose the ruler—can account for a wide range of bizarre psychological occurrences and disruptions of thought, from awareness of sudden, disturbing fantasies and inclinations to seemingly innocuous trivial pursuits. Freud summarizes both of these extreme possibilities:

> We know, for example, that if a drive representative is removed by repression from the influence of the conscious, it develops more rampantly and exuberantly. It proliferates in the dark, so to speak, and finds extreme forms of expression, which, when translated and presented to the neurotic, not only are bound to appear alien to him, but also frighten him by making the drive seem so extraordinary and dangerous in its intensity. This deceptive intensity is a result of the drive's uninhibited development in fantasy and the build-up caused by lack of satisfaction. The fact that repression has this latter effect points us in the direction of its true significance.
>
> To return to the opposite perspective, though, we should make it clear that repression does not even keep all derivatives of primally repressed material away from the conscious. If they are far enough removed from the repressed representative, whether by distortion or by the number of mediating interpolations, then they can have free access to the conscious. It is as if the resistance against them from the conscious were a function of their remoteness from the originally repressed material. (Freud 1915b / 2015, 37–8)

Because drive impulses do not surrender when repressed and because they are fixated on an object that will satisfy them, they turn to any unconscious content, any repressed ideas that they can recruit in the service of their aims.

Sometimes, however, as Freud recognized, drives might recruit unsavory psychological characters that are downright terrifying to the individual, causing additional repression, before taking a degree of satisfaction in a more tolerable replacement. If one were fixated on the breast, for example, your ruling impulse might repress an inappropriate desire to touch your employer's chest. The drive and its idea, now free to conspire in the depths of the unconscious, associate themselves and charge with psychic energy any breast-like ideas they can, such as one's own chest, a cow's udders, milk, a dartboard and its bullseye, a snow globe, etc. Some of these associations might be able to sneak past the ruling impulse by appearing to pose as no threat to its aims, perhaps in the form of a puzzling fantasy to milk a cow, an innocuous desire to drink a glass of milk, or a disturbing fantasy to sensually caress a snow globe. The anxiety that gives rise to repression, then, is not some tension that is discomforting for the "self" or "ego," in the Cartesian sense of a conscious "I," but the unconscious fear that entertaining other ideas will frustrate the ruling impulse's goals. Contra Descartes, consciousness is merely the theatre in which the ruling impulse puts on its performance, mobilizing the organism into action to gratify its needs.

9.6 | A Theory of Protean Psychology

David Hume took the first step away from the Enlightenment by proposing that *all* of our thought processes were spontaneous, including the seemingly rational ones. Freud developed these ideas even further by taking a closer look at what he believed was

happening behind the scenes of conscious awareness. What he discovered was not just a divided mind, but a *fractured* one, like a territory shared amongst warring kingdoms, each vying for power and limited (cognitive) resources. These insights are important, but Davidson does not embrace all of them, taking instead a more modest approach to save action theory from the tyranny of the irrational.

Returning to the case of Mr. S, the question Davidson wants to answer is: why would someone intentionally act *against* the principle of continence, especially if he avows this principle? Inspired by Freud's arguments against the unity of the mind, Davidson believes that traditional action theory must come to grips with the following psychoanalytic ideas if it wishes to understand irrational phenomena without giving up on the relationship between reason and action (Davidson 1982, 170–1; 184–5):

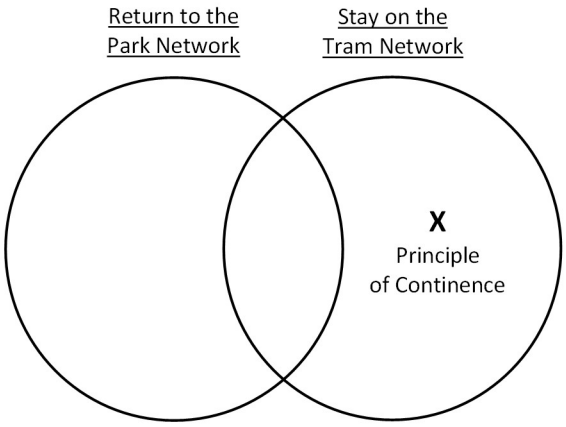
(1) The mind is compartmentalized into smaller, independent structures
(2) Each of these structures is individuated by a network of reasons, beliefs, and desires
(3) There can be non-logical causation between each[165]

This proposal has come to be known as the *Second Mind Hypothesis*, and the basic idea is that the mind is divided into relatively rationally coherent miniature conceptual networks, each capable of making arguments in support of some course of action. Some of these networks overlap, sharing ideas or concepts, but parts of their webs also fan out away from each other. As Davidson explains, while talk of partitions and parts is useful, it is misleading if it is taken to suggest that they cannot share content; instead, he suggests that we think of partition in terms of "overlapping territories" (181, n.6).

[165] In addition to these three psychoanalytic ideas, Davidson entertains in passing that repression could be a fourth psychoanalytic idea to consider. Unfortunately, he gives no defense of this other than saying that while he believes unconscious mental states are "like conscious beliefs, memories, desires, wishes, and fears," it is curious that only some of these states seem readily available to consciousness while others "can become accessible only with difficulty" (Davidson 1982, 171).

According to the Second Mind Hypothesis, there is no need to explain how Mr. R as a whole can act intentionally against the principle of continence. There is not even a need to understand how a unified mind can be inconsistent with itself. Simply put, Mr. R does not have a unified mind. To be sure, each of the parts share some ideas—ideas about parks, branches, other people, etc.—but there is also content that they do not share. Relevant to one part, for instance, is the time and trouble it would take to revisit the park, calculations that do not enter into the part that strongly advocates for correcting a previous action. His psychological division can be depicted with a straightforward Venn diagram (see figure 9.B).

***figure 9.B* (Second Mind Hypothesis)**



How to understand these networks is a matter of theoretical difficulty, for much of Davidson's presentation is speculative and experimental, intended to explore how traditional action theory might come to grips with the problems of irrationality. Considering his words carefully, his theory appears to suggest that we should not think of Mr. S as having a "Return to the Park" network *per se* whose job is to advocate specifically for returning to the park by recruiting any reasons available to accomplish this goal, and yet at the same time, Davidson makes clear that networks are individuated

by the reasons that they use and the actions that they cause (Davidson 1982, 185). Should we suppose that there is a network for every action that we undertake or every *type* of action? Is there a network for every conclusion or every *type* of conclusion?

Whereas some theories of division postulate that the mind is compartmentalized into several semi-permanent structures, Davidson instead appears to be suggesting a radical, fluid approach to compartmentalization. Mr. S does not have a "Return to the Park" network that is constantly at work or even just activating when the opportunity occasions itself, but every now and then, one forms from the primordial psychological ooze of beliefs and desires. Are there organizing principles that direct how and when these networks come into being? Perhaps, but this is not addressed.[166] The more important innovation that this theory recommends is viewing the mind as dynamic, its networks emerging, evolving, competing, and dissolving. Davidson explains that any given *network* is rational, but how those networks interact with one another and how the mind operates as a whole is *not* (Davidson 1982, 181). A network is thus a "second mind" only insofar as it is rationally organized and engages in practical reasoning; contrary to the name of the theory, however, a network is neither a *second* mind nor a *mind* at all.

On the Second Mind Hypothesis, then, psychological conflict enters into the picture precisely when multiple networks are activated and arrive at differing conclusions. One network might contain the principle of continence, using it to arrive at an *all things considered* judgment as to what is best, while another network, oblivious to

[166] In a footnote, Davidson is explicit about this: "I have nothing to say about the number or nature of divisions of the mind, their permanence or aetiology. I am solely concerned to defend the idea of mental compartmentalization, and to argue that it is necessary if we are to explain a common form of irrationality" (Davidson 1982, 181, n.6).

this principle, might recommend the most pleasurable course of action. If it turns out that the courses of action recommended by each network oppose one another, unable to be jointly satisfied, then it leaves the individual with a choice to make. How is she to make her decision?

9.7 | Cultivating the Gardens of the Mind

Drawing from Davidson's account of akrasia in section 6.3, there must be some third thing—e.g. another network, the mind as a whole, or perhaps just some part that is identical to the agent—that motivates the person into action, and this third thing, the *agent*, must have access to the conclusions reached by the conflicting networks even though it may not be privy to *all* of the information that resides in each. The job of the agent is to form an intention and act, and so each network presents it with what it takes to be best. It is on this point that Davidson introduces a social interaction metaphor to explain how this might occur, writing:

> There is, however, a way one mental event can cause another mental event without being a reason for it, and where there is no puzzle and not necessarily any irrationality. This can happen when cause and effect occur in different minds. For example, wishing to have you enter my garden, I grow a beautiful flower there. You crave a look at my flower and enter my garden. My desire caused your craving and action, but my desire was not a reason for your craving, nor a reason on which you acted. (Perhaps you did not even know about my wish.) (Davidson 1982, 181)

What this metaphor elucidates about the individual mind, thinks Davidson, is that the mental states belonging to one network might be able to exert indirect influence through non-rational means. What is left unclear, however, is the subject of this influence. Did Davidson intend to argue that networks can influence *one another*, or might they only influence the agent into action?

If we interpret this passage as suggesting that one network might "tempt" another, it is difficult to see how this theory can avoid the problem of homuncularism, for this would seem to require that a network be able to perceive, think, and evaluate on its own, as if it were a "little agent," something that Davidson explicitly rejects and advises against (Davidson 1982, 185). The alternative instead assigns this power of circumspection to the agent whose job is to make an *all-out judgment* as to the desirability of what is before her, i.e., to form an *intention*. On this reading, a network reaches a conclusion that it presents to the agent, and if that conclusion is enticing enough, the agent will intend to act accordingly.[167]

According to this version of the Second Mind Hypothesis, the explanation for psychological conflict is that a person has multiple networks activated, each recommending a different course of action. In the case of Mr. S, Davidson can argue that there is a "Stay on the Tram" network that happens to be sensitive to the principle of continence and a "Return to the Park" network that never factors the principle into its calculations, each arriving at opposing conclusions. Having the recommendations from both networks in view, Mr. S as agent looks at each to judge which is most desirable and worth putting into action. He has no reason to act against the principle of continence, which is what makes his action irrational—he entertains no argument that the principle

---

[167] Davidson himself is unclear on this topic. He resists the idea that his theory needs a homunculus, but he speaks of the importance of one part of the mind being able to non-rationally influence another as if it were like social interaction (Davidson 1982, 181; 184). Yet elsewhere, he suggests that these causal interactions are more similar in function to naturalistic causes, an idea that seems to undermine his earlier insistence that "blind forces are in the category of the non-rational, not the irrational" (180; 185). All of this is muddied further still by his failure to better explain how a divided mind can produce an action, which he acknowledges (181).

Because his theory is preliminary and speculative, a degree of inconsistency is to be expected, and so in these sections, I am offering a reconstruction that highlights what I take to be the more important points while leaving aside questions of inconsistency.

has qualifications, is unreliable, or even leads to harm—but because it is not a principle shared by both networks, he is free to intentionally act contrary to it. By judging that the "Return to the Park" network's conclusion is most desirable, he effectively ignores the principle of continence residing in the "Stay on the Tram" network, as if he were hoping to avoid an awkward encounter with a disappointed friend by looking elsewhere while he walks in public.

The Second Mind Hypothesis thus represents Davidson's attempt to pivot away from the Cartesian assumption of the unity of the mind while preserving the idea that it is still somehow inherently rational. The virtue of this hypothesis is that it enables us to understand and meaningfully attribute irrationality to people in a way that makes them rational. To recognize that Mr. S is undertaking an action in the Davidsonian sense of the word is to see him as a person, someone who is capable of possessing concepts, reasoning, forming intentions, and acting on them, even if the performance emerges from a conflict between opposing networks. Indeed, on this theory, the very possibility of irrationality requires conceptual networks, each with its own beliefs, desires, emotions, representations, intentional objects, and logical associations. By neither trivializing nor eliminating the possibility for irrationality, Davidson's theory manages to elevate the dignity of personhood by extending a degree of rationality to all people even when they are engaged in the strangest or most perverse of performances.

While the Second Mind Hypothesis is commendable insofar as it strives to maintain that there is a kernel of intelligibility underwriting all forms of human behavior, even the most deviant, this is not to excuse it from any problems. Davidson is well aware that it raises as many questions as it answers. He explains:

> Recall the analysis of akrasia.[168] There I mentioned no partitioning of the mind because the analysis was at that point more descriptive than explanatory. But the way could be cleared for explanation if we were to suppose two semi-autonomous departments of the mind, one that finds a certain course of action to be, all things considered, best, and another that prompts another course of action. On each side, the side of sober judgement and the side of incontinent intent and action, there is a supporting structure of reasons, of interlocking beliefs, expectations, assumptions, attitudes, and desires. To set the scene in this way still leaves much unexplained, for we want to know why this double structure developed, how it accounts for the action taken, and also, no doubt, its psychic consequences and cure. What I stress here is that the partitioned mind leaves the field open to such further explanations. (Davidson 1982, 181)

9.8 | The Failures in Davidson's House of Reason

Though the previous section offered one interpretation of Davidson's Second Mind Hypothesis, it is by no means obvious how the theory can explain action without collapsing into homuncularism or else non-rational, mechanical causation. After all, positing the existence of a metaphysical agent or supernatural will to navigate past both problems seem like solutions Davidson would reject, and so it leaves us with the burden of explaining how agency can occur with a divided mind.[169] This is a problem that is not unique to the Second Mind Hypothesis though, for it burdens even traditional action theory. Is the agent metaphysical or physical? Is it the mind itself? Is the mind identical with the brain? Is there an agent at all? Every action theory must grapple with these questions. As far as irrationality is concerned though, the more pressing issue for the Second Mind Hypothesis is that it does not explain nearly as much as it hopes.

---

[168] See sections 6.2 to 6.4 for a review of his theory of akrasia.

[169] In his article, "Intending," published just four years prior to "Paradoxes of Irrationality," Davidson's opening paragraph derides the idea of explaining intention in terms of "unanalyzed episodes or attitudes like willing, mysterious acts of the will, or kinds of causation foreign to science" (Davidson 1978, 83).

Although Davidson rejects the assumption of the unity of the mind, he is curiously ambivalent when it comes to its transparency, declaring that his thesis can accommodate the idea of an unconscious—an idea that he believes would only make his thesis more explanatorily robust—while also remarking that any philosophical objections to an unconscious can go unmet since his thesis does not require one, using akrasia as a demonstration (Davidson 1982, 185–6). He marvels at the flexibility of the Second Mind Hypothesis:

> It is striking, for example, that nothing in the description of akrasia requires that any thought or motive be unconscious—indeed, I criticized Aristotle for introducing something like an unconscious piece of knowledge when this was not necessary.[170] The standard case of akrasia is one in which the agent knows what he is doing, and why, and knows that it is not for the best, and knows why. He acknowledges his own irrationality. If all this is possible, then the description cannot be made untenable by supposing that sometimes some of the thoughts or desires involved are unconscious. (186)

This raises some theoretical concerns however. If the mind is compartmentalized into networks and those networks are the same *type* (i.e. each is a constellation of rational relations of propositions and concepts), differing only in *content*, then why are they unable to directly communicate? Is this a physiological limitation or a mental one? If they really are unable to communicate, then how does this theory *not* require an unconscious? Does the agent see the contents of both networks and inexplicably fail to relate them—a failure that itself would be irrational and unaccounted for by the Second Mind

---

[170] Using the case of Mr. S, Davidson's criticism of Aristotle's analysis of akrasia reads:

> Aristotle suggested that weakness of the will is due to a kind of forgetting. The akrates has two desires; in our example, he wants to save his time and effort, and also wants to move the branch. He can't act on both desires, but Aristotle will not let him get so far as to appreciate his problem, for according to Aristotle the agent loses active touch with his knowledge that by not returning to the park he can save time and effort. It is not quite a case of a conscious and an unconscious desire in conflict; rather there is a conscious and an unconscious piece of knowledge, where action depends on which piece of knowledge is conscious. (Davidson 1982, 175–6)

Hypothesis? Yet, if the networks *can* communicate with one another, then how are we to understand compartmentalization?

Though it appears as though these concerns arise in response to the question of whether it is necessary to posit an unconscious, they actually point to a more fundamental issue with the Second Mind Hypothesis: its commitment to the Cartesian assumption of the rationality of the mind. Even if Davidson tries to explain division in terms of conscious and unconscious networks, he would still need to address why these networks cannot communicate if they are of the same (i.e. rational) type. To argue that they cannot interact *because* some are conscious and others unconscious is to beg the question.

Jonathan Lear also recognizes this problem. Using the cases of an unhappy couple and of Mr. R (the latter discussed in chapter seven), he argues that while the Second Mind Hypothesis succeeds in accounting for those unconscious motivations that have a rational structure, it ultimately fails to appreciate the complexity of the unconscious as understood by Freud (Lear 2005, 27). Each case is summarized as follows:

> Consider an unhappy couple where each partner has, over the years, built up many reasons to be angry at the other. But, somehow, in order to stay together each has devised a strategy of keeping the reasons for anger out of conscious awareness. Officially and sincerely, each is not angry with the other. But every now and then a vengeful act slips out—though the partner who acts is not really aware of what he or she is doing. (27)

> Mr. R is walking along a road on which he knows his lady-friend will later be traveling in a carriage. He removes a stone from the road so that the carriage will not be damaged. A bit later he feels compelled to go back to replace the stone in the road. (24)

In the first case, each member of the unhappy couple is able to give reasons for feeling angry even though both are unaware of their anger, repressing it from conscious awareness. This web of reasons for feeling angry, Lear explains, "has the same basic

structure as the conscious mind," and the only thing missing is the conscious label "anger" for this structure (27). There is nothing in this example that cannot be accounted for by the Second Mind Hypothesis.

For Mr. R, however, things are a bit different. According to Lear, the reason that he replaces the stone in the road is because he feels angry with his beloved for not reciprocating his interest, and yet, he is oblivious to both, the feeling *and* the reason (Lear 2005, 27–8). How is this possible? It is only through analysis that Freud discovers that the act of replacing the stone is an expression of hostility. From Mr. R's perspective, he takes himself to be performing the right action, replacing a stone that he should not have moved in the first place, and yet this is a rationalization that fails to make sense when all of the particulars are considered. Why is the stone a problem in the first place? Is it really a danger? Why place it back where it was and not simply move it further from the road? How can two opposing actions be motivated by the same concern? For these questions, Mr. R can give no good answer, and as a result, he suffers a reflexive breakdown, failing the test of self-understanding. Lear argues that Mr. R cannot articulate or understand his feelings of anger *because* they lack a supporting structure of reasons; like an infant, Mr. R is only able to *express* his anger, lacking a more cognitively developed, rational understanding of it (26; 33).

9.9 | Chaos Theory

What I have hitherto referred to as the Enlightenment attitude or Enlightenment thinking might be part of a much older story in the history of western thought. While reflecting on the Socratic Thesis—the idea that no one willingly desires what is bad—and how taking it seriously entails denying the existence of akratic actions, Lear concludes that Socrates,

not Descartes, might ultimately be responsible for our "presumption of rationality built into the very ideas of agency, action, and mind" (Lear 1998, 81). Regardless of the origin of this assumption, it is difficult to imagine how one can take human rationality for granted in light of the failures of reason, the spontaneous nature of rational processes, and the theoretical and practical difficulties created by psychological conflict. Davidson was thus right to anticipate that the way forward for action theory was to acknowledge irrationality and embrace a handful of Freudian insights. What he failed to see was that, in addition to the divided mind and non-rational causation between the parts, our theory does in fact demand an unconscious, and there may be no other way to account for the opacity of self-understanding, insights that Hugo Mercier and Dan Sperber attribute to Freud:

> The commonsense confidence in one's ability to know one's mind was, of course, undermined by the work of Sigmund Freud and its focus on what he called "the Unconscious." The existence of unconscious mental processes had been recognized long ago, by Ptolemy or Ibn Al-Haytham,[171] but until Freud, these processes were seen as relatively peripheral. Mental life was regarded, for the most part, as typically conscious, or at least open to introspection. However, Freud made a compelling case that we are quite commonly mistaken about our real motivations. A century later, in a cognitive psychology perspective, the once radically challenging idea of the "Unconscious" seems outdated. Not some, but all mental processes, affective and cognitive, are now seen as largely or even wholly unconscious. (Mercier and Sperber 2017, 114)

A consequence of these ideas—psychological division, non-rational causation, unconscious processes—taken together is thus a welcome decentralization of reason, a turn away from the Cartesian assumptions that have hampered our abilities to better understand human thought and behavior. Doing so allows us to better account for aspects

[171] Mercier and Sperber have in mind Ptolemy's and Ibn Al-Haytham's contributions to the study of optics, both of whom argue that perception is partly and automatically constituted by a subject's knowledge without one awareness of it. Philosopher Gary Hatfield discusses both optical theories in his article, "Perception as Unconscious Inference" (Hatfield 2001, 4–5).

of human nature that otherwise conflict with the Enlightenment optimism concerning our rational natures.

The previous chapter explored the bizarre nature of the so-called rational processes of the mind, and the theory that helped explain the intuitive nature of rational thought as well as its operations and failings is known as the *modularity thesis of mind*. This chapter, by contrast, has focused on the *divided mind*, reviewing how psychological division not only explains but even entails disorder and anarchy in thought. Modularity, too, implies a divided mind, challenging the Cartesian assumptions of the rationality and unity of the mind while preserving the Freudian insight of the mind's chaotic nature.

Recall that the modules available to the desert ant, discussed in section 8.5, each has its own function, and together, they ultimately aid the ant in effortlessly and efficiently navigating back to its nest. This example suggests that modules produce as outputs inferences that are either additional representations or inclinations to action, and these, in turn, are taken up by the activity of other modules that can encourage or disrupt those activities. Similarly, in section 3.7, the research into split-brain patients by Joseph Bogen and Michael Gazzaniga showed how cognitive dysfunction can highlight the compartmentalization of the brain's highly specific operations. In particular, their research discovered that such patients were able to follow directions but utterly failed to give reasonable explanations for what they were doing. The lessons from cognitive dysfunction, as well as other cleverly designed psychological experiments, suggests that mental modules are blind, operating on whatever information is available and suited to their inferential mechanisms. Modules only evolve to optimize their tasks, and it is this complex interplay between many modules that creates conditions for a restless, chaotic

mind, filled with tension and rife with paradox. It is precisely this modularity of the mental that recommends a view of human nature much more similar to the pre-Modern, Christian narrative. As Pinker eloquently explains:

> More generally, the interplay of mental systems can explain how people can entertain revenge fantasies that they never act on, or can commit adultery only in their hearts. In this way the theory of human nature coming out of the cognitive revolution has more in common with the Judeo-Christian theory of human nature, and with the psychoanalytic theory proposed by Sigmund Freud, than with behaviorism, social constructionism, and other versions of the Blank Slate. Behavior is not just emitted or elicited, nor does it come directly out of culture or society. It comes from the internal struggle among mental modules with differing agendas and goals. (Pinker 2002, 40)

But why do we have a restless, chaotic mind? Why, specifically, are the parts as they are? Why do they interact with one another in the way that they do? Why do they fail to interact with one another?

Neither Freud nor the Second Mind Hypothesis can furnish us with many satisfactory answers. This is less of a challenge for the modularity thesis though. From the perspective of natural selection, explanations can be found for why a particular module—such as reason or cheater detection—might have evolved. Mercier and Sperber, for instance, speculate that there is a reason module that is oriented towards social concerns, having evolved to help build, track, and manage reputations, including one's own (Mercier and Sperber 2017, 123–7). The harder question is whether it can be considered advantageous for the *whole* mind to be restless and chaotic. Might this simply be an unintended global side-effect of the evolution of parts of the mind? Perhaps; but it is also possible to make the argument that a restless mind is a good mind.

9.10 | The Gospel of Disruption: Thou Shalt Not Rest

It is always possible that a fractured mind genuinely serves *no* purpose. In evolutionary theory, in addition to *adaptations*—traits that result from natural selection—there are *exaptations* and *nonadaptations*. An exaptation is a side-effect of some adaptation that is eventually co-opted by natural selection to serve a new function for an organism. The evolution of flight discussed in section 8.4 makes for a good example. Flight was a *side-effect* of traits that originally aided thermoregulatory and motor coordination functions, but when flight became possible and proved advantageous for the organisms that developed it, evolution started to select *for* it (Flanagan 2000, 107). A nonadaptation, by contrast, serves no purpose; it too is a side-effect of other traits, but unlike an exaptation, it is not a product of environmental pressures except indirectly. While Owen Flanagan references the color of blood as an example of a nonadaptation—a consequence of oxygenation and immune functions—an even better example comes from philosopher Frank Jackson (106). Jackson explains that evolution selected for polar bears to have thick coats of fur to help them survive in arctic temperatures; however, a consequence of a thick coat of fur is a *heavy* coat of fur, slowing the animal down. It would be short-sighted, he argues, to count this as either a refutation of evolutionary theory or to believe heaviness or slowness were adapted traits. They are simply by-products (Jackson 1982, 134). It is therefore plausible that a chaotic, compartmentalized mind is nothing more than a by-product of a modular mind. But *does* a chaotic mind serve a purpose? Even if it was not directly selected for by evolution, might a such a mind be more like the property of flight, incidentally proving itself very useful?

Reflecting on empirical evidence, clinical experience, and conceptual requirements, Jonathan Lear is one of the few prolific contemporary philosophers who questions the Enlightenment assumption that the mind is fundamentally rational in the first place. In his book *Open-Minded*, he briefly surveys Aristotelian and Davidsonian solutions to the problem of akrasia before arguing that both solutions "assume that Socrates is basically right: that the concept of mind requires rationality."[172] He continues:

> By contrast, I want to argue that it is intrinsic to the very idea of mind that mind must be sometimes irrational. Rather than see irrationality as *coming from the outside* as from an Unconscious Mind which disrupts Conscious Mind, one should see irrational disruptions themselves as inherent expressions of mind. In a nutshell: mind has a tendency to disrupt its own functioning. (Lear 1998, 84; emphasis original)

Why, specifically, does he come to this conclusion and how could it possibly be good for a mind to disrupt itself?

Lear begins his argument by deconstructing the very idea of mind, not in the abstract but in the concrete, human sense, and he concludes that there are two important features that human minds have, features that are seldom considered in conceptual analyses of mindedness: the potential for creativity and embodiment (Lear 1998, 85). If the mind is to have the capacity for creativity, then it needs to be dynamic in its activity, moving from one idea to the next in idiosyncratic and unpredictable ways. This Lear takes to be a form of *restlessness*, defined not by an excessively structured mental sequence that follows along clearly defined logical paths, but, on the contrary, the ability "to make leaps, to make associations, to bring things together and divide them up in all

---

[172] On Lear's interpretation, Aristotle argues that the akrates acts from ignorance, momentarily forgetting what is best, and Davidson's solution is the Second Mind Hypothesis. Lear's treatment of each is brief, but for a fuller discussion see chapter six of this project. Neither are re-introduced here because the point of emphasis is on the Enlightenment assumption of the rational mind, rather than revisiting possible solutions to the problem of akrasia.

sorts of strange ways" (85). Drawing from Freudian contributions to psychology, he finds evidence for this kind of restlessness in human sexuality and dreaming.

Freud was the first in the modern age to argue persuasively that human sexuality is *not* a biological instinct, betraying "great plasticity in its aim and object" as opposed to a relatively rigid behavioral pattern (Lear 1998, 85). As Lear explains in *Freud*, sexuality can be conceptually divided into having an *object* and having an *aim*. Prior to Freud's ingenuity, it was believed that the object of human sexuality was *another human being* with the aim of *reproduction*, but this failed to account for the great variation in how human beings express their sexuality (Lear 2005, 73). It is not uncommon to find sexual objects ranging from oneself to same-sex partners and even inanimate objects; likewise, aims can involve pleasure, sado-masochism, denial, or even just the satisfaction of one's partner, none of which necessarily contribute to reproduction. Lear does not deny that from a biological perspective, human sexuality evolved for reproductive purposes, but because of the *way* it evolved, it now allows for such enormous "variation in activity and object that no particular variation could possibly count as an instance of its breakdown" (73). For him, this signals that what has been getting naturally selected is "an inextricable entanglement of sexuality and imagination," for he sees no other way to explain this sexual plasticity, the consequence of which is the preference for imaginative engagement in sexual acts over reproduction (73). Like the trait of flight, imaginative play in sexuality looks like an exaptation that follows from both a free, creative mind and an urge to reproduce.[173]

[173] Evolution always selects for those traits that have practical results, traits that confer very real advantages for organisms in their environments. While the role of imaginative play in sexuality may be selected for, it can only be *because* it contributes to the fitness of the species—presumably by encouraging *more* sexual activity. Increased production of offspring is the most obvious benefit, but there may be advantages from

Likewise, due to the clinical success in interpreting dreams *as* meaningful, Lear believes that this reveals the existence of a form of symbolically rich mental activity that is not necessarily immediately clear to the subject of the dreams (Lear 1998, 85). Whether or not dreams have any *inherent* meaning has been a subject of debate, but that does not detract from the fact that it reveals a form of alogical, chaotic mental activity that defies the understanding. Indeed, Lear argues that such activity shows that the mind "must be making certain associations among ideas, engaging in symbolization, however elementary," and yet, he continues, "those associations must be opaque to conscious, rational-thinking mind" (85). It is precisely this chaotic mental activity, this restlessness, that creates the conditions for creativity, and it is not hard to imagine how such cognitive flexibility confers advantages for survival in ever-changing environments where food availability and shelter opportunities can quickly shift from abundant to scarce. If we *had* to think along rigid, logical sequences, adapting to change suddenly becomes much more difficult, for (as Hume had realized) there is no logical argument in the absence of experimentation that would enable someone to infer that bread nourishes.[174] Creativity fuels the experimentation and flexibility needed for survival.

The second important feature of having a human mind is being embodied. Embodiment places necessary limitations on what the mind can know, serving as both a literal and metaphorical expression of a limited, fragile existence. Lear believes that it is

---

non-reproductive sexual aims as well, such as social cohesion or increased productivity. For an introduction to some of the sexual practices amongst different species, see Dawkins' chapter, "The Battle of the Sexes," in *The Selfish Gene*.

[174] Section 8.3 reviewed Hume's theory of inference, and in particular, it highlighted how section four of his *Enquiry Concerning Human Understanding* expresses skepticism that rationality is what we use to make inferences regarding the world around us. There, he writes, "Our senses tell us about the colour, weight and consistency of bread; but neither the senses nor reason can ever tell us about the qualities that enable bread to nourish a human body" (Hume *EHU* 4.20).

an important fact about the mind that it is integrated with a body and situated in an environment, unable to exercise complete control over either (Lear 1998, 85). From illness to environmental threats, there is much for the mind to manage and little that it can do, and yet, these are the conditions under which it came to be. It is these conditions that place limits on what can be known, these conditions that, as Lear emphasizes, make omnipotence impossible. At the other end of the spectrum, he has observed in his clinical work with psychotics a correlation between a delusion of omnipotence and actually losing one's mind. In such a state of psychosis, the mind loses touch with reality as input, no longer having anything "to operate on or in relation to," and the result of such a state is severe reflexive breakdown, a loss of all sense and self-understanding (86). Though his clinical work is not so much intended to be an argument as it is a philosophical reflection on omnipotence, reality, and healthy mental functioning, he is firm in his insistence that the mind needs the body, complete with all of the unpredictability and uncertainty that a physical body in a foreign environment brings.

Though Lear is convinced, just how important is embodiment?

9.11 | Embodied Complications

In 1981, Hilary Putnam popularized the idea that there could be a disembodied brain, a "brain in a vat," but such a notion could not be further from what it means to have a mind. Given his research, Antonio Damasio believes that if such a feat were even possible, it would disrupt the feedback loop between the brain, the body, and the environment on which one's sense of being alive depends, a sense that underwrites mental life (Damasio 1994, 228). He invites us to imagine, for instance, walking alone late at night before realizing that somebody is following. What happens during such an

event? Those with a modest familiarity of popular scientific ideas might explain this in terms of threat-detection and fight-or-flight response, but Damasio warns that this is an over-simplification of what actually occurs. Instead, he elaborates:

> The neural and chemical aspects of the brain's response cause a profound change in the way tissues and whole organ systems operate. The energy availability and metabolic rate of the entire organism are altered, as is the readiness of the immune system; the overall biochemical profile of the organism fluctuates rapidly; the skeletal muscles that allow the movement of head, trunk, and limbs contract; and signals about all these changes are relayed back to the brain, some via neural routes, some via chemical routes in the bloodstream, so that the evolving state of the body proper, which has modified continuously second after second, will affect the central nervous system, neurally and chemically, at varied sites. The net result of having the brain detect danger (or any similarly exciting situation) is a profound departure from business as usual, both in restricted sectors of the organism ("local" changes) and in the organism as a whole ("global" changes). Most importantly, the changes occur in *both* brain and body proper. (223–4; emphasis original)

The amount of description that Damasio shares is intended to illustrate just how much the body informs the brain and vice versa in any experience of fear. Myriad changes occur throughout the body, including the brain, in response to potential threat-detection, and the changes are initiated in response to information received not just from the environment but other parts of the body and brain.

What is more, this degree of interdependence between body, brain, and environment is not restricted to exceptional circumstances like those found in experiences of fear. Even in the deceptively passive activity of ordinary perception, the same degree of information exchange takes place, triggering a similar cascade of physiological changes throughout the body. Damasio once more elaborates:

> Think of viewing a favorite landscape. Far more than the retina and the brain's visual cortices are involved. One might say that while the cornea is passive, the lens and the iris not only let light through but also adjust their size and shape in response to the scene before them. The eyeball is positioned by several muscles, so as to track objects effectively, and the head and neck move into optimal

> position. Unless these and other adjustments take place, you actually may not see much. All of these adjustments depend on signals going from brain to body and on related signals going from body to brain. (Damasio 1994, 224)

Thus, even in a standard case of perceiving one's environment, the body is constantly feeding information to the brain, which in turn makes adjustments to the body to optimize that flow of information, ultimately to enable an organism to reproduce and survive. From an evolutionary perspective, this should come as no surprise. Damasio does not deny that the mind emerges from neural activity, but at the same time, he recommends that we remember that "those circuits were shaped in evolution by functional requisites of the organism" (226). The primary jobs of our neural circuits are to regulate the body, monitor environmental stimuli, and mobilize the body into action.

Given these considerations, Damasio has difficulty imagining how such a "brain in a vat" could result in a normal mind—if any mind at all. Were there no body to feed information back to the brain, then the brain would cease performing its most basic functions, the regulation and modification of the body, and if such feedback could be artificially produced, then it would only confirm that bodily inputs are necessary for minds (Damasio 1994, 228).

There are thus good reasons to take Lear's hypothesis seriously. Embodied minds really are characterized by restlessness, and it is only such a mind that permits and empowers organisms to creatively engage with their environments while optimizing their prospects for success (Lear 1998, 88). Consequently, the chaotic life of the modular mind, one that persistently exposes us to spontaneous disruptions in our patterns of thought, turns out to be a beneficial feature insofar as we, as embodied beings, find

ourselves needing to respond to novel challenges and produce creative solutions in an environment that often reminds us that nothing can be taken for granted.

9.12 | Conclusion

Traditional action theory has tended to embrace the Cartesian assumptions that minds are unified, transparent, and fundamentally rational, but this comes at the cost of failing to explain human behavior under less than ideal circumstances, as Davidson rightly recognized in his analysis of akrasia. Part of the motivation for accepting these Cartesian ideas is that they entail that there exists a lawlike connection between beliefs, desires, and actions, a connection that makes human behavior intelligible and normativity possible. It is this lawlike connection, so Enlightenment thinking contends, that forms the bedrock of our social institutions, enabling us to establish and codify norms that dictate correct and incorrect behavior, norms that dictate what is reasonable and unreasonable to do.

Both Freud and Davidson made important psychological and philosophical contributions to our theory of mind and our analyses of human behavior. It was Freud who picked up Hume's anti-Enlightenment banner and, inspired by Plato, took the understanding of the mind in a fresh, non-Cartesian direction. His theory of the unconscious as consisting of a nexus of competing drives that influence what we do not only anticipated the modularity thesis but even provides a framework for how we might understand modules with competing ends. Although Davidson could not fully complete the turn away from Cartesianism, he challenged traditional action theory to entertain the possibility that the human mind is a divided one.

The long-standing difficulty in accepting the thesis that the mind is divided is that it did not make sense against the background of Enlightenment thought. What is

reason's role if not to track truth? What are human beings if not free, rational creatures? What purpose does a divided mind serve? Why would we *not* know what is happening in our own minds? How is that even possible? How could we be anything other than free when it comes to thinking? Even Hume, after all, believed that we exercised autonomy in the kingdoms of our imagination.

Much has changed since 1793, and it was Darwin, specifically, who discovered a new way to conceive of purposes, goals, traits, and behaviors by understanding embodied organisms striving to meet biological demands in changing environments. It is this post-Darwinian perspective that affords us with an alternative to Enlightenment thought. The previous chapter suggested that we consider reason to be just one product of our many mental processes, and this chapter explored how these processes are housed in a complex mind that has evolved to meet natural *and* social demands, a mind that *must* be creative if it is to adapt to unforeseen problems. To ignore these developmental facts is to make real human behavior and decision-making, behavior that occurs in public and in private, more mysterious and less intelligible. Before Darwin, many of these ideas were not only controversial but made little sense, but now we can see how they are far more consistent with our evolutionary story as foragers and social creatures than any Enlightenment alternative rooted in Cartesian assumptions about the mind—assumptions that arose from meditations that we are disembodied, solitary thinkers.

Supposing, then, that the mind is divided, how are we to make sense of who we are? Do we accord one person to each "part" of the mind? Is there no person at all? If not, then what does this mean for responsibility, whether social, legal, ethical, or otherwise? Where do we go from here? What does this mean, above all, for agency?

Chapter 10: The Future of Our Illusions

The solution of Heracleitus's problem,
though familiar,
will afford a convenient approach to some less familiar matters.
The truth is that you *can* bathe in the same *river* twice,
but not in the same river stage.
You can bathe in two river stages which are stages of the same river,
and this is what constitutes bathing in the same river twice.
A river is a process through time,
and the river stages are its momentary parts.
Identification of the river bathed in once with the river bathed in again
is just what determines our subject matter
to be a river process as opposed to a river stage.

— Willard Van Orman Quine, "Identity, Ostension, and Hypostasis"

10.1 | The Absent-Minded Self

In response to the adaptive model of self-deception advanced by Robert Trivers and William von Hippel, one modular theorist, Robert Kurzban, explains that while the evolutionary account of self-deception is provocative, he believes it is ultimately misguided. It is not so much the *idea* that something like self-deception could have evolved to confer strategic advantages that he views as problematic; rather, it is the metaphysical commitments that such a discussion implies. He worries:

> Because I, among others, do not think there is a plausible referent for "the self" used in this way, my concern is that referring to the self at best is mistaken and at worst reifies a Cartesian dualist ontology. That is, when "the self" is being convinced, what, precisely, is doing the convincing and what, precisely, is being convinced? Talk about whatever it is that is being deceived (or "controlled," for that matter) comes perilously close to dualism, with a homuncular "self" being the thing that is being deceived. (Kurzban 2011, 32)

From Kurzban's theoretical point-of-view, there is nothing mysterious about self-deception. Some mental modules have evolved to guide the action of the organism, and so they must create, store, and use representations that are *accurate*; other mental

modules, by contrast, have evolved to aid in persuasion, and so they create, store, and use whichever representations maximize success in social interactions (32).

A module that has been fine-tuned by evolution to aid in persuasion is not going to value accuracy to the same degree as a module fine-tuned to guide action. If a persuasion module happens to use an accurate representation, it is not on account of its *accuracy* that it does so; it is merely a coincidence in the sense that the accurate representation just incidentally happens to be the socially attractive one. Imagine, for example, intending to persuade someone that you know.[175] Depending on who it is, you may adopt different strategies. With a stranger, you might exaggerate the truth to motivate them to help you with a task, but with an employer, you might present exactly what is true while withholding some details that risk tarnishing your reputation at work. If you are feeling desperate, you may even try to outright sell a falsehood. While these strategies work with varying degrees of success in the social world, they prove far more hazardous in a wild and natural environment where survival is the goal. Identifying a fruit as a berry while overlooking the detail that it is poisonous will almost certainly yield a fatal outcome. But how are we to understand conflicts between modules? If there are these conflicting goals between them, is it not the case that one must "conceal" what it "knows" from another? Is this not a return to Sartre's censor or some kind of homunculus?

Censors and homunculi arise as theoretical obstacles whenever we suppose that information within a system is uniform *and* that the system is interconnected. If some

[175] This is not to suggest that modules "think" or "intend" anything; instead, the following examples merely illustrate how contexts and goals dictate strategies. Modules are strategic insofar as they have evolved to perform highly specialized functions, and when the conditions present themselves, they tend to operate automatically and accordingly.

information-processing system exchanges theoretical bits of, say, Φ, and the parts of this system are conjoined, then it becomes challenging to explain why some parts have access to Φ and some do not. For instance, how might part A pass bits of Φ to part B but not part C? Does it "know" which bits of Φ should go to B but not C? Does it make a value judgment that B is better suited for *this* bit of Φ but not *that*? The theories advanced by both Freud and Davidson did not provide any clear answers. There may be elaborate explanations that can finesse this type of objection, but on a modularity thesis, the concern about homuncularism is misplaced.

Thanks to a feature known as *informational encapsulation*, the information that one module uses is not necessarily available to others (Kurzban 2010, 50). If we imagine the mind as composed of modular parts and we accept the thesis that it has evolved under the pressures of natural selection, then there is no reason to suppose that each and every part shares its information with the rest any more than to suppose that lung tissue must be coextensive with brain tissue and kidney tissue. Each module has evolved to perform its function, and if that function is radically distinct from that of another module, they might not even operate at the same time let alone process the same information.

Suppose there is a sleep module which begins to function when physiological and environmental cues present themselves, such as increased melatonin production and decreased detection of higher frequency light waves through the retinas. This type of physiological and environmental feedback triggers the sleep module to prepare the organism for rest, influencing its behavior and goals. Now, why would a different module, such as one that has evolved for reproduction, monitor and process these same cues? Unless a higher level of melatonin or decreased perception of blue light waves

somehow correlated with an increase in reproductive success, a reproductive module would be blind to this information, for these cues are wholly irrelevant to its task. "Evolution must act to connect modules," writes Kurzban, "and it will only act to do so if the connection leads to better functioning" (Kurzban 2010, 50).

On the modularity thesis, not only is there no "commanding" module but there is nothing within the mind that corresponds to a "self" at all. The mind, rather, is an amalgam of modules, each having evolved for its own specific purpose, and what is more, as Freud suspected, psychological conflict is the norm. Provided that it does not interfere with reproductive success, evolution just does not care about psychological conflict.

10.2 | The Curious Self

The economist Steven Landsburg believes that two of the most difficult questions we can ask are: 'Why is there something rather than nothing?' and 'Why do people lock their refrigerator doors?' (Landsburg 2007, 177). The first of these questions has been pondered by the most brilliant minds of every age, beginning with the pre-Socratics in ancient Greece up through contemporary cosmologists and theoretical physicists, such as Lawrence Krauss and Brian Greene. But why is the second question so important? What could it possibly mean?

According to Enlightenment thinking, human beings have Cartesian minds—unified and rational—and while much scientific progress has been made since Descartes, this theory of mind is still very much commonplace.[176] Such a theory of mind

[176] Kurzban, for example, identifies not only Landsburg, but notes how economist Ken Binmore has surveyed his field similarly, finding that adherence to these kinds of Cartesian assumptions—consistent preferences, rational behavior—is still popular (Kurzban 2010, 19; 155; 238 n.4). Economist Richard

recommends thinking about human psychology in terms not only of beliefs and desires but also *preferences*. Do we not, after all, hold some beliefs to be true more strongly than others? Do we not desire some goals more than others? If I *really* believe that broccoli is healthy and I *really* desire to be healthy, then would I not choose broccoli over a donut every single time, especially if I *really* believe that the donut is unhealthy? For the Enlightenment theorist, it follows that predicting or rationalizing behavior is as simple as an economist or action theorist determining which preferences an individual has and what type of syllogism is informing her decision-making.

So why would locking a refrigerator door pose a problem?

Kurzban invites us to imagine the following situation:

> Suppose it's 8:00 PM, and I have just finished eating dinner. I know that I am going to wake up at midnight, and, when I do, that I am going to have to answer the following vexing question: Should I eat the leftover chocolate cake when I find it staring back at me invitingly in the refrigerator, or should I forgo the chocolate cake and go back to bed? (Kurzban 2010, 152)

Traditional action theory ought to analyze this as a relatively straightforward matter. What do you desire more, pleasure or health? Do you have any relevant beliefs about a chocolate cake being healthy or pleasant? Do you have any relevant beliefs about eating dessert late at night?

Let us say that we know the answers to these questions—you really do desire to be healthier and believe that eating chocolate cake will interfere with this goal. We can represent the practical reasoning you might use to guide your action with the following syllogism:

Thaler also shares this view, referring to the creature economists tend to study as *homo economicus*, rather than human beings (Thaler 2015, 4).

P_1 Any action that results in promoting health is desirable
P_2 Avoiding chocolate cake at midnight promotes health
C [Action of avoiding chocolate cake at midnight]

Even if we add a second syllogism that represents a desire for pleasure and a belief that eating chocolate cake is pleasant, as long as it is true that you *also* desire health *more* than pleasure, there should be no conflict. In other words, given all of your beliefs, desires, and preferences, you *will* avoid the cake at midnight.

If the above is correct, then why lock the refrigerator door? If your practical syllogism at 8 PM dictates that you will avoid the cake at midnight, then—supposing that no further beliefs, desires, or preferences have been acquired between 8 PM and midnight—should we not expect this same syllogism to prevail later in the evening?

10.3 | The Inconsistent Self

For some of us, our so-called 8 PM syllogism really does fail at midnight, and much of part two reviewed some of the many different ways in which this form of practical rationality can be hijacked by non-cognitive processes and states, causing it to fail in hitherto unexpected and surprising ways. Psychologists have discovered that we tend to interpret ambiguous information so that it strengthens our pre-existing beliefs, shift our goals in response to changes in emotional or motivational states, and alter our preferences depending on how a problem is framed, to recall just a few of the non-cognitive effects on decision-making. As more research emerges, the Enlightenment view of the mind as unified and rational appears increasingly inadequate, but on the modularity view, once we understand how modules interact, locking a refrigerator door at midnight appears to be a very reasonable thing to do.

Recall how temporal distance can influence decision-making.[177] People, for example, show mixed preferences when deciding between one apple today or two tomorrow, but when presented with one apple in 365 days or two in 366, the overwhelming majority prefer the latter option (Thaler 1981, 202). Why does distance into the future shift our preferences so dramatically even though the choices are roughly equivalent—one apple at time t^1, or two twenty-four hours later?

Similarly, psychologists Amos Tversky, Paul Slovic, and Daniel Kahneman set out to investigate a phenomenon known as *preference reversal* whereby a person displays inconsistency between what she *chooses* and what she *values*. To understand this, take a look at the following two options, a high-probability and low-probability gamble (Tversky, Slovic, and Kahneman 1990, 204):

> *H* bet: 28 / 36 chance to win $10
> *L* bet: 3 / 36 chance to win $100

Now imagine you were selling tickets for a fundraiser that correspond to these odds in some kind of lottery, say *H* tickets and *L* tickets, and imagine further that you get to set the prices for these tickets. For how much would you sell an *H* ticket? What about an *L* ticket?

If you set your *L* ticket at a higher price than *H*, then you—along with the majority of people—must believe that the *L* bet is more valuable than *H* (the payout is higher after all). What happens when you reverse the situation, however? Pretend for a moment that a generous friend has priced these two tickets equivalently, both the *H* ticket and the *L* ticket cost the same amount. Which would you choose?

---

[177] See section 6.9 for a review.

If we are necessarily rational, as Enlightenment theorists contend, then reason dictates that we ought to show a preference for what we believe is most valuable, especially when all else is equal. Why would we choose something *less* valuable under such circumstances? And yet, this is precisely what researchers discovered when they asked participants which of the two options they themselves would choose. Most people *value* the higher payout more but *prefer* to take their chances on the lower payout with the better odds (Tversky, Slovic, and Kahneman 1990, 204).

How can we explain this discrepancy between value and preference with an Enlightenment theory? Do we rationally calculate risks, weigh probabilities, and cast our stones with what is most likely?

Tversky, Slovic, and Kahneman considered this explanation, and so to factor out the variables that participants were performing risk-assessments and probability calculations, they designed a variant of this experiment that excluded the odds of winning a monetary amount, tying it instead to delayed payments and temporal distance. Participants were now presented with the following options:

S: $1,600 payout in 1.5 years
L: $2,500 payout in 5 years

After collating their data, the researchers discovered an extraordinary amount of preference reversal. The short-term (*S*) option was selected over the long-term (*L*) 74% of the time, and yet *S* was valued more than *L* just 25% of the time (Tversky, Slovic, and Kahneman 1990, 213). Why does this happen, even in the absence of probabilities? Why do our preferences not align with our values?

Kurzban believes that a phenomenon such as preference reversal should only come as a surprise if you believe that people are perfectly rational decision-makers

(Kurzban 2010, 158). The modularity view, by contrast, eliminates the mystery. He explains:

> Modularity informs how context matters. Understanding the design of modules, and the features of the environment—internal and external—that they respond to can help explain patterns of choices. (159)

How, then, are modules designed?

One important distinction to make is between *short-term* and *long-term* modules. Short-term modules respond to immediate needs as well as features that are present in the environment here-and-now. In other words, these modules are *opportunistic*. When you are presented with food, for example, it is the short-term modules that exercise considerable weight over your decision-making, inclining you to take advantage of the opportunity before you just in case another opportunity does not present itself for some time. Kurzban, who calls them *impatient* modules, suggests thinking about them as those modules "that make you do things that would make a lot of sense if the world were going to end tomorrow" (Kurzban 2010, 159). The short-term modules look after your survival *right now* and your reproduction opportunities *right now*, and in a world of competition where resources can be scarce and opportunities unpredictable, these modules aim to put you in a position to succeed.

Of course, responding to the here-and-now does not always put us at a strategic advantage if our biological goals are survival and reproduction. There are occasions where rewards are greater if we can find a way to exercise some self-restraint. What if a rival knows that you are brash and has laid a trap for you? What if you discover some food during a time of day when predators are most active? If inclement, dangerous weather approaches, do you take your chances on a potential mate that may very well be

there tomorrow or on your survival? In a modern environment, you might discover that you enjoy a rich dessert on occasion, but were you to consume one with every single meal, as it slowly begins to adversely impact your health, you will begin to incur a significant cost to your long-term survival and mating opportunities. Forgoing some of the rewards of the moment may amplify our opportunities in the long-run, and there are modules designed to aid in precisely this, *long-term* modules.

Long-term modules take into account factors such as probabilities, opportunity availability, and risks / rewards over time. For these reasons, Kurzban refers to them as *patient* modules. Rather than make an inference from the here-and-now, these are the sorts of modules that construct and make inferences from representations of *possible* states-of-affairs, affording us a sense of futurity with which we can plan (Kurzban 2010, 164). It is difficult, for instance, to increase one's physical strength, but with the help of long-term modules, we can endure the present discomfort of exercise for the reward of increased strength in the future.

While we as a culture tend to value long-term, strategic thinking over short-term, impulsive thinking, the truth of the matter is that long-term forecasts matter little if their predictions never come to fruition and rewards are never reaped. For this reason, it is good for a cognitive system such as our own to have a combination of both types of modules at work, taking care to balance the interests between long-term survival and short-term success.

So why do our preferences and values fail to align occasionally? It depends on the module processing the information and influencing our behavior.

Unfortunately, in practice, the interaction between modules is far from an elegant affair. They simply construct or make inferences from the information available to them while inclining us to act on the basis of that information. They do not necessarily coordinate, and the modular life is certainly not a democracy.

10.4 | Macroscopic Self-Composition

Much as Freud anticipated, insofar as goals are in conflict, the activities from all sorts of modules result in inhibitions and promotions of the interests of one another as they scramble to get their way. This does not mean that they *consciously* do this, nor does it follow that they have any kind of *awareness* of this; rather, this conflict emerges as a consequence of their activities. It is akin to working in an office environment, compartmentalized by cubicles, with the exception that as one works in her cubicle, she has no idea that anyone else is present and working in their own cubicles. Any particular employee, an expert at her job, simply works in her cubicle as various files continuously slide across her desk, awaiting her review and recommendations. She never asks from where the files come, but she knows exactly what she is looking for, what she is supposed to do, and where to place the files when she is finished, generally uninterested in the goings-on of others. Unbeknownst to her, some of her recommendations help employees working in the cubicles across from her while others harm the productivity of the employees who work at the other end of the building.

The danger in using metaphors to understand modularity is succumbing to the temptation that there must be some prime mover, some CEO, some decision-maker who takes all of the recommendations into account and executes what she thinks is best. This is not how the modular mind works though. A better idea is to imagine the business as

largely egalitarian, and as each employee reviews her files and makes recommendations, some of their decisions have consequences for how the business itself functions. Jointly, the sum of the activities of every employee influences the direction of the business and which policies it adopts and enforces. It can function just fine without a CEO, provided that the employees have been carefully vetted and groomed by a competent human resources department: natural selection.

So how might conflict resolution occur in the absence of a CEO?

If all modules were active at all times and pulled equal weight, this would indeed pose a problem, and clearly, a persistent inability to function as a result of being torn between opposing motivations would be counterproductive to our fitness. Fortunately, our cognitive system operates in such a way so as to mitigate this type of outcome, and in particular, Kurzban identifies three design features that he believes can facilitate how and to what degree modules can influence our behavior. Modules, he explains, must be *context-sensitive*, *state-dependent*, and *history-dependent*.

A context-sensitive module is one that operates at peak performance whenever the conditions obtain for which that module is especially suited. When nobody is around to observe our behavior, for instance, modules that regulate our concerns for our reputation will tend to be their most quiescent, and when it seems as though there is a very real chance for a mating opportunity, modules that regulate our concerns for reproduction will tend to be hyper-active, even inclining us to engage in potentially risky behavior that under normal circumstances we might never consider. It should come as no surprise that short-term modules will have an edge over long-term modules whenever one encounters

a here-and-now opportunity, or, to put it in Kurzban's words, "seeing cake gives the cake-liking system an advantage" (Kurzban 2010, 163).

But what if you encounter a slice of decadent chocolate cake not long after devouring a rather large meal? So sated from this meal only moments earlier, you are well aware that one more bite of anything could imminently cause you to feel very unwell. Is there still this urge to eat? If not, does this not contradict the context-sensitivity of modules?

It is in circumstances such as these that state-dependence plays a role in regulating the activity of modules. A module should have some way of monitoring the current state of the organism relative to its function and, consequently, upregulate or downregulate its activity appropriately. If the organism's status is "satisfied," then a foraging module should dial back its efforts to incline the organism to seek out food, and likewise, if the status is "starved," that same module should be expected to ramp up its efforts (Kurzban 2010, 162–3). This is what it means for a module to be *state-dependent*.

Still, neither context-sensitivity nor state-dependence can explain why we sometimes abandon our efforts. This is a familiar experience for many people, especially when it comes to dieting. So many find themselves excited about the prospects of how adopting a new diet could enhance their appearance or promote their health, and yet, it is not uncommon for people to quit when it seems as though their weight is not changing quickly enough. On the modularity theory, the explanation is that our modules are *history-dependent*, adapting to reward / effort ratios (Kurzban 2010, 163). If one course of action is not yielding a reward, it is just tying up the amount of time and energy that could be spent pursuing other opportunities, and as it demands more effort while yielding

little to no reward, it should come as no surprise when other modules begin inclining us towards more fruitful courses of action. It would be counterproductive to survival if we lacked a mechanism that encouraged us to call off our efforts when they started to prove futile.

Between these three features of both short-term and long-term modules, we can account for sufficiently complex behavior without appealing to an inner-CEO who is rational and calculating. Just as Freud had envisioned in 1915 with his early theory of drives presented in his essays on the unconscious, psychic life really is chaotic. Modules can be and are in conflict, vying for influence when opportunities are ripe for the taking, and because long-term modules make inferences from representations of possible states-of-affairs, they can even anticipate and sabotage the potential efforts from short-term modules that interfere with their goals. We lock our refrigerator doors at night to stymy the efforts of the modules that will exercise more influence on us later in the evening. Kurzban explains, "Long-sighted modules have different preferences from short-sighted modules," writes Kurzban, "and, as they are able to move first and are capable of planning, they can limit the choices of short-sighted modules" (Kurzban, 164).

What do we do without a CEO? What does modularity mean for being human and being rational? Should we adopt a fatalistic attitude towards the self and life? Should we abandon any aspirations for being reasonable and improving society? Not at all. Modularity damns just one sense of self, the Cartesian kind, but it also invites opportunity to reevaluate the importance of other senses of self.

10.5 | Microscopic Self-Evaluation

An amoeba is one of the simplest living organisms on the planet. Encased by a cellular membrane, it consists largely of jelly-like cytoplasm that both aids in its movement and helps protect its organelles—any structures within a cell that have specialized functions. Two such structures in particular are the vacuole and nucleus. A vacuole is a small cavity in the cytoplasm that can perform a number of functions, from waste management to food digestion, and the nucleus houses the organism's DNA, modulating the expression of its genes and controlling its reproductive activity.

The amoeba, usually living in a pond, slowly navigates its environment using its protean cytoplasm by extending several foot-like shapes known as pseudopodia out in front of itself while retracting those behind it.[178] By using this form of motion, it can propel itself forward in search of a food source, such as bacteria or algae. Once located, the pseudopodia completely surround the food, encapsulating it in a vacuole. Digestion begins the moment enzymes are released into the vacuole, breaking down the food into nutriment and eventually collapsing the vacuole itself.

Beginning with reflections on biological continuity and the nature of living organisms, Antonio Damasio marvels at the life of an amoeba, emphasizing that it "is not just alive but bent on staying alive" (Damasio 1999, 136). Clearly lacking a nervous system, it is highly unlikely that the amoeba possesses even a scintilla of awareness of its existence or any of its functions, but its behavior is saturated with intention inchoate. It navigates its external environment, scavenges for food, regulates its internal cellular environment, and engages in a form of reproduction. Some of the most primitive

[178] For a fascinating look at the movement of an amoeba, see Ralf Wagner's short video, "Amoeba in motion," which can be found here: https://www.youtube.com/watch?v=7pR7TNzJ_pA.

biological urges experienced by nearly every complex organism are being satisfied by this singular, tiny creature whose entire existence is organized around holding itself together and flourishing in a chaotic sea of unceasing change.

What the example of the amoeba illustrates is the quite literal importance of boundaries in defining a singular organism against the rest of the environmental landscape. Damasio reflects, "Life and the life urge exist inside a boundary, the selectively permeable wall that separates the internal environment from the external environment" (Damasio 1999, 137). Within those boundaries that separate one thing from the environment is the body, and this, Damasio argues, affords us with a very real sense of self. But what is it that makes the body a worthy candidate for selfhood in the first place? How can an everchanging body ground a meaningful sense of self?

There is a well-known quasi-fact that the human body recycles itself every seven years. What makes this claim grounded in fact is that so many of the cells in our bodies really are perishable, replaced by genetic copies, but the myth lies in the assumption that there is something special specifically about seven years.[179] Each kind of cell has its own lifespan, some lasting less than a week and others withstanding as much as a decade. As

[179] The story of how it became possible to detect this is itself fascinating. As explained by science journalist Roxanne Khamsi in *Nature*, due to above-ground nuclear testing until 1963, a radioactive form of carbon, carbon-14, had doubled beyond its natural levels in the environment. As a result, higher-than-normal levels of this radioactive atom also made its way into our diets and ultimately our DNA. In 2005, a team of stem cell researchers, lead by medical doctor Jonas Frisén of the Karolinska Institute of Stockholm, realized that this excess amount of carbon-14 could be used to determine a cell's birthdate and to assess turnover rates.

For more on this, see his article, "Retrospective birth dating of cells in humans" (Spalding et al. 2005).

Also see Khamsi's news article: "Carbon dating works for cells," which can be accessed via URL: https://www.nature.com/news/2005/050711/full/news050711-12.html.

far as we currently know, one exception seems to be the neurons housed in the cerebral cortex, but even these undergo their own set of changes. Damasio explains:

> We are not merely perishable at the end of our lives. Most parts of us perish during our lifetime only to be substituted by other perishable parts. The cycles of death and birth repeat themselves many times in a life span—some of the cells in our bodies survive for as little as one week, most for not more than one year; the exceptions are the precious neurons in our brains,[180] the muscle cells of the heart,[181] and the cells of the lens. Most of the components that do not get substituted—such as the neurons—get changed by learning. (Damasio 1999, 144)

Perhaps this does not necessarily appear to be a problem, but consider for a moment a well-known metaphysical paradox known as the Ship of Theseus (Lowe 2002, 25–6). After the ancient Greek hero Theseus had died, his ship was stored in Athens, and over time, it began to rot and decay. Intent on preserving this legendary relic, the Athenians resolved to replace each ruined part so as to not lose the ship to the ravages of time. Eventually, every single part of the ship had been replaced—the mast, the deck, the bow—leaving us to wonder whether this can still count as *the* Ship of Theseus. Gone is the wheel on which he laid his hands, the hull that sailed to Athens from Crete, and the timber on which Theseus sweat. Is this really the Ship of Theseus or an imitation?

---

[180] Frisén and his team have identified, not without controversy, what looks like cell turnover of neurons in the hippocampus. Neurobiologist Moheb Costandi, however, explains why we should be cautious in accepting this claim in his feature article for *New Scientist*, "The mystery of the missing brain cells." In that article, another neurobiologist, Andrew Lumsden, even contends that neurogenesis in adults—the production of new neurons from stem cells—may be costly and inefficient for complex brains such as our own, suggesting that we may have evolved favoring plasticity instead. This is not to deny that adult neurogenesis does occur; on the contrary, neurobiologist and primatologist Robert Sapolsky reviews the history of the discovery of adult neurogenesis in both the hippocampus and cortex in his book *Behave: The Biology of Humans at Our Best and Worst* (Sapolsky 2017, 147–50).

For more on this, see Frisén's article, "Dynamics of hippocampal neurogenesis in adult humans" (Spalding et al. 2013).

Also see Costandi's feature piece in *New Scientist* at the following URL: https://www.newscientist.com/article/mg21328521.600-the-mystery-of-the-missing-brain-cells.

[181] Less controversially, Frisén and his team has also identified adult turnover of muscle cells in the heart as well. See his article, "Evidence for cardiomyocyte renewal in humans" (Bergmann et al. 2009).

Similarly, we can ask of our bodies undergoing substantial replacement and change: is this *the* body of twenty years ago or is it a new one?

The paradox of the Ship of Theseus strikes at the very heart of questions about identity. As Damasio acknowledges, if anything is to count as a *self* as a *you*, there must be continuity of reference. If minds and bodies could be swapped using some new technology, for instance, then using the body as a source for self-reference would be futile, for your body might house *you* one day and Johann or Miranda the next. When I make reference to something you did ten years ago, it needs to in some sense be the same you as today, or else authorship and agency lose their sense of meaning. Ultimately, what underwrites identity and selfhood, contends Damasio, is *stability*; that is, in some sense, you have to be *you* across a long period of time (Damasio 1999, 134–5).

If there is not a Cartesian self to act as an anchor for continuous reference over time and if the body is akin to the Ship of Theseus, each part constantly replaced and renewed, how can the body provide for a sense of self? The very notion of a changing body seems to contradict the stability required for identity and selfhood.

10.6 | Orthoscopic Self-Modeling

Life occurs within a very narrow range of parameters. At the cosmic level, for instance, the conditions for life require a reasonably-sized, rocky planet that has a protective atmosphere and travels within the "Goldilocks Zone" (a distance from a star that keeps the temperature "just right"), amongst many other requirements agreed upon by astrobiologists.

Even stricter parameters are required to sustain the life of each kind of organism. In the case of a human being, our internal temperature—one of many values on which our

life depends—is a relatively stable, approximate 37°C. If this rises above 40°C, we very quickly begin facing the onset of life-threatening complications as the body uses every resource available, from blood vessel dilation to profuse sweating, to sustain itself; and above 42°C, the risk for death increases exponentially in proportion to the duration of fever and any further elevations in temperature. The same is true when our temperatures fall as well, risks beginning around 34°C. Although increases above 42°C in ambient temperature make for a great deal of discomfort for most of us, it is quite remarkable to consider how well we thrive in environments well below 30°C, even without the conveniences afforded by modern life. How is this possible?

Internal temperature is tightly regulated by the operations of the brain as it strives to maintain a unique steady state known as *homeostasis*,[182] which is a continuous fine-tuning of the inner-environment of the body, the *internal milieu*, in response to changes in the external environment, injury, and foreign invaders (Damasio 1999, 138). Temperature along with other bodily processes—such as blood pressure, heart rate, respiration, hormone regulation, and immune system function, among others—contribute collectively to the internal milieu, and each has its own range of optimal functioning; but also like temperature, when values fall outside of those ranges, it can become a matter of life or death. Untreated infections, for instance, often quickly become quite deadly, and so the body cannot allow just any form of bacteria to inhabit the internal environment and proliferate unchecked without risking grave consequences. The same is true for blood pressure (both high or low), the immune system (whether excessively stimulated or

[182] Because of just how much activity is required to maintain favorable conditions within the internal milieu, though Damasio uses the term *homeostasis*, he suggests that we ought to collectively consider transitioning to the term proposed by neuroscientist Steven Rose as a replacement, *homeodynamics* (Damasio 1999, 141).

suppressed), and each of the other processes integral to the functioning of the body. While some changes to the internal milieu are inconsequential and some even welcome, it is also true that there is a very limited range in which such changes can be tolerated without the body making the appropriate adjustments. With the boundaries afforded by having a body, it is this internal environment that distinguishes an organism from the rest of the external environment.

All of this autoregulation is made possible through *somatosensory signaling*,[183] whereby the body uses representations from several subsystems to constantly update different parts of the brain as to their status (Damasio 1999, 149). In response to the information obtained through these status updates, the relevant parts of the brain make their adjustments accordingly. In the case of thermoregulation, a cone-shaped part of the brain known as the *hypothalamus* is responsible for increasing or decreasing the body's temperature. The brain itself is thus a modular system whose joint activities through somatosensory signaling help maintain the internal milieu of the organism. In fact, every single activity from the brain operates with the organism's body in mind, each neural module having evolved for controlling, regulating, activating, and suppressing varying parts and aspects of the body. Damasio explains how this even occurs in the deceptively simple act of crossing a street:

> Imagine yourself crossing a street, and now picture an unexpected car driving fast in your direction. The point of view relative to the car that is coming toward you is the point of view of your body, and it can be no other. A person watching this scene from a window on the third floor of the building behind you has a different

[183] Damasio identifies three subsystems in somatosensory signaling: (1) internal milieu and visceral; (2) vestibular and musculoskeletal; and (3) fine-touch. The first concerns the internal environment of the body and senses chemical changes taking place, while the second reports on the states of the muscles and skeleton and the third on alterations in the skin as we interact with objects in the environment (Damasio 1999, 150–3). Each subsystem uses its own delivery systems (usually different sets of nerve fibers), and this fact, believes Damasio, makes it clear that each of these systems evolved on their own (149–50).

> point of view; that of his or her body. The car approaches, and the position of your head and neck is altered as you orient in its direction, while your eyes move conjugately to focus on the rapidly evolving patterns formed in your retinas. A world of adjustments is in full swing, from the vestibular system, which originates in the inner ear, has to do with balance, and serves to indicate body position in space, to the machinery of the colliculi, which guides eye and head and neck movement with the help of brain-stem nuclei, to the occipital and parietal cortices, which modulate the process from high up. But this is not all. Having a car zooming toward you does cause an emotion called fear, whether you want it or not, and does change many things in the state of your organism—the gut, the heart, and the skin respond quickly, among many others. (146)

Change is thus *necessary* for the body to be able to interact and respond appropriately to changes in the environment. Damasio continues, explaining:

> There is no such thing as a *pure* perception of an object within a sensory channel, for instance, vision. The concurrent changes I have just described are *not* an optional accompaniment. To perceive an object, visually or otherwise, the organism requires both specialized sensory signals *and* signals from the adjustment of the body, which are necessary for perception to occur. (147; emphasis original)

Without these adjustments taking place, the odds of making it to the other side of the street suffer tremendously; for that matter, the odds of being able to do anything without changes taking place are dismal.

Thus, what provides the stability needed for a self is not the *absence* of change, but the continued maintenance of a flexible yet relatively consistent structure, one capable of undergoing and withstanding a degree of changes that make acting in an environment possible. Without a body, there can be no agency, and without a body capable of changing in response to the environment, there can be no action. It is the body, as a carefully maintained structure, that supplies a continuous point of reference.[184]

---

[184] Damasio does not make the argument that I am making, that the body is the ground for the self. Rather, he defends an idea he calls the *proto-self*, arguing that selfhood begins with the aggregation of the somatosensory representations used by the brain (Damasio 1999, 154). The proto-self is thus a kind of representation of the state of the organism as a whole (173). I suspect the reason he does this is because he

What makes the paradox of the Ship of Theseus so challenging is the addition of an external variable that plays an essential role in constituting the identity of the Ship of Theseus. Were there no Theseus, there would be no Ship of Theseus, and it is precisely Theseus' *absence* that fuels the paradox. As the Athenians continue to replace a ship without his presence, it leaves us to wonder when it ceases to be the Ship of Theseus and not just an imitation of it. But when it comes to the body as the ground for the self, there is no analogue to Theseus, no Cartesian mind, no ghost in the machine, no captain of the ship that indispensably constitutes our identities with respect to our bodies. Imagine for a moment that Theseus was not integral to the identity of the Ship of Theseus but instead it was the *structure*, the *design*, and the *function* of the ship, and within the ship was its own raucous environment of sailors tasked with keeping the interior clean and performing other ship-keeping duties. Each part of this ship can be replaced, whether wood or sailor, and it would still persist as the Ship of Theseus provided that there were no interruptions in its continued existence *as* a ship. This is what it is like to have a body, and it is this malleable yet fixed constitution that allows it to ground any sense of self. As Damasio explains:

> Throughout development, adulthood, and even senescence, the *design* of the body remains largely unchanged. To be sure, bodies grow in size during development, but the fundamental systems and organs are the same throughout the life span and the operations that most components perform change little or not at all. (Damasio 1999, 141; emphasis original)

takes more sophisticated forms of the self to begin with consciousness, which he believes arises out of neural activity, and so he seeks some kind of continuity between all forms of the self.

10.7 | Self-Representation

Positing the body as a form of self is to defend it as a point of reference and necessary condition for any other forms of self. It is not to suggest that all organisms are conscious, morally responsible, or even have rich and interesting conceptions of themselves, but without bodies, it is not clear how any of those other fascinating qualities can obtain. In fact, in the case of some organisms who are especially neurobiologically and behaviorally complex, it appears that other, higher forms of self do emerge, and in particular for human beings, one idea that has received a lot of attention is the *autobiographical self*.

The idea of an autobiographical self as it is currently understood was first popularized by philosopher Daniel Dennett in his seminal work *Consciousness Explained*.[185] Similar to the argument made in the preceding section, Dennett too first contends that there is something necessary about a biological self even though he recognizes that it is "porous and indefinite." Home to interlopers both friendly and hostile, were it not for our well-organized, bounded bodies, distinguishing a self from anything else becomes an exercise in futility (Dennett 1991, 414).

But there is more to complex organisms than just bodies. As Dennett explains, beavers make dams, spiders spin webs, bees construct hives, and hermit crabs find shells,

---

[185] Damasio, too, discusses the autobiographical self, but whereas Dennett views the matter from the perspective of information organization and language, Damasio is far more interested in the potential neurobiological underpinnings and requisites, such as memory processes and somatosensory signaling. The two views can be complementary even though there is some disagreement in the details.

In *Descartes' Error*, Damasio explains:

> Daniel Dennett's hypothesis, on the other hand, pertains to the high end of consciousness, to the final products of the mind. He agrees that there is a self, but he does not address its neural basis and focuses instead on the mechanisms by which our experience of a stream-of-consciousness might be created. (Damasio 1994, 244)

For more on Damasio's understanding of the autobiographical self, see *The Feeling of What Happens*, particularly chapters six and seven (Damasio 1999, 168–76; 222–33).

as if each of these creatures had extended their boundaries beyond the body. To be sure, each can still "live" in the absence of these extensions—not live very well and certainly not without a body—but it is precisely this appropriating and fashioning of the environment that enhances their chances for survival and reproduction, leading Richard Dawkins to coin the phrase *extended phenotype* to describe these special kinds of organism-environment interactions (Dennett 1991, 415).

Do human beings have their own extensions?

The answer is a resounding yes when we reflect on tool use, technology, and invention, furnishing us with everything from clothes and computers to automobiles and shelters. And yet, there are also more abstract webs that we spin known as *sentences* and shells that we find known as *ideas*, all of which are made possible through our unique capacity to produce and consume *linguistic representations*. It is language (along with memory) that equips us with the representational flexibility and precision to think about abstract properties like color and currency exchange rates, large and discrete numbers that exceed what can be visualized accurately in the imagination, and fine-grained parts of objects from gears and springs to protons and electrons. Above all, however, language also enables us to represent *ourselves*, as we pull bits and pieces of words and phrases, descriptions of memories and traits, and tales of successes and failures to construct this thing each of us calls *me*.

The *me* that results from this narrative construction is ultimately as discrete and individualized as any number represented by a symbol, made unique by a delicate coordination of words that detail history and possibility, fantasy and reality, hope and

fear, all with the human being that is you at center stage. It is what Dennett calls the *center of narrative gravity* (Dennett 1991, 418). He summarizes his thoughts thusly:

> A self, according to my theory, is not any old mathematical point, but an abstraction defined by the myriads of attributions and interpretations (including self-attributions and self-interpretations) that have composed the biography of the living body whose Center of Narrative Gravity it is. As such, it plays a singularly important role in the ongoing cognitive economy of that living body, because, of all the things in the environment an active body must make mental models of, none is more crucial than the model the agent has of itself. (426–7)

But why do human beings so naturally and effortlessly spin autobiographical selves?

Dennett's response to this question is somewhat unsatisfying. He rightly argues that in order to act as an agent in the world, an organism must know which thing it is. This is not to imply that organisms must be self-aware as much as it is to acknowledge good, basic biological design principles. Thus, he explains, this crude awareness of one's boundaries is something that is "'wired in'—part of the underlying design of the nervous system, like blinking when something approaches the eye or shivering when cold" (Dennett 1991, 427). The proverbial snake that eats its own tail, for example, surely failed to recognize its tail as part of its own body. When it comes to human beings, however, Dennett argues our biological makeup and surrounding environment is more complex, and so we need these narrative selves to identify a *me* and help better survive. We use a self to better organize the information we acquire about the world we inhabit, categorizing the self as something different from the rest of the environment. This, he insists, allows us to locate the needs that correspond to *us* as opposed to someone else, assign moral responsibility, and plan our actions (428–9).

Dennett is not wrong in that we *can* do these things with autobiographical selves and that such activities are useful, but it fails to paint a convincing picture of the

environmental pressure under which an autobiographical self likely evolved and proved useful: the management of one's reputation in a *social* world.

10.8 | Self-Awareness

In the *Enigma of Reason*, Hugo Mercier and Dan Sperber invite their readers to imagine the following scenario:

> You enter the waiting room at your doctor's office. There is already another patient. You both exchange glances and say, "Hello!" You sit down. She is intermittently typing on her smartphone and staring at the screen. You take a magazine. She looks at her watch and sighs. You exchange another glance. (Mercier and Sperber 2017, 96)

What is happening here?

The above scenario is anything but far-fetched, thoroughly saturated with mundanity. Whether it occurs in the setting of a doctor's office, a business suite, or a classroom, it details an experience with which almost every single one of us is all too familiar. And yet, despite its deceptively pedestrian nature, so much is happening between the two persons in this room. Each person, Mercier and Sperber contend, is engaged in a form of *mind-reading*, an activity that takes place effortlessly. Using a Sherlock Holmes style synopsis, they elaborate on what is occurring behind the scenes:

> When you arrived in the waiting room, she knew that you would understand that she would see the doctor before you. You were disappointed that there was already someone waiting, but you tried not to show it, not to let her read your mind on this, but she probably did all the same. You understood that her alternately tapping on her smartphone and staring at it were part of an ongoing interaction with someone else with whom she was chatting at a distance (even if you had no idea what they were chatting about). You guessed that she was looking at her watch because it might already be past the time of her appointment with the doctor, and that she was sighing because she was unhappy to have to wait. You understood from the exchange of glances that she had understood that you had understood her sigh. And so on. (Mercier and Sperber 2017, 97)

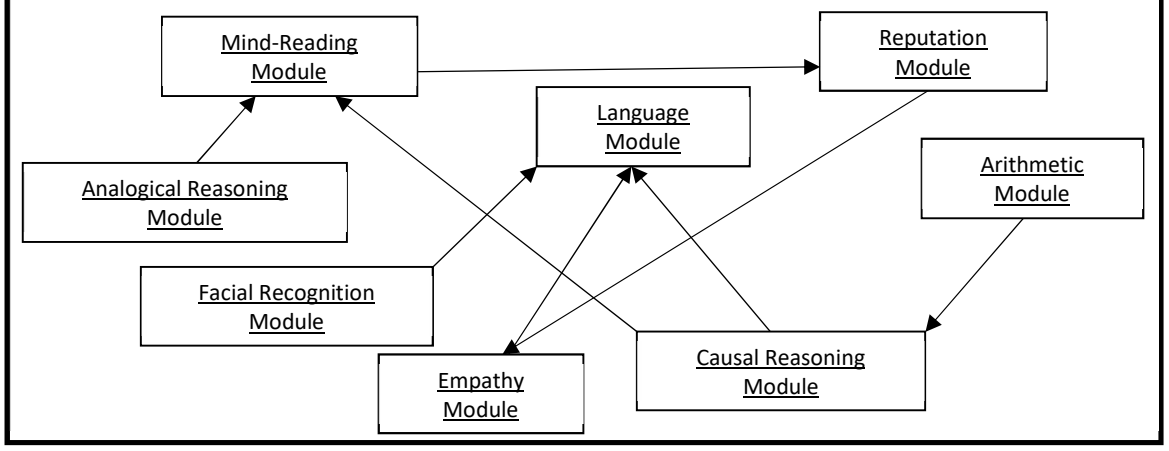
It is as if such mind-reading proclivities are "wired in" to our biological design, as if there is a *mind-reading module*, and indeed, this is precisely what Mercier and Sperber claim.

As social creatures, we have a strong interest in what our kin are doing and thinking, often even speculating what they might be thinking about *us*. To accomplish this weighty task and keep everything neatly organized, Mercier and Sperber suggest that we are using a kind of mental filing system. Each person that we meet causes us to open a new file folder and begin organizing its contents, updating it appropriately when we come across new information or make inferences that lead us to believe otherwise. Depending on the level of importance a person has to our *social* survival, the file may be brief and temporary or lengthy and permanent (Mercier and Sperber 2017, 97–8).

As mentioned in sections 8.4 and 8.5, our modules also jointly co-operate to produce more complex functions and inferences. The mind-reading module, for instance, may work along with an arithmetic module to infer how much money a friend owes, or it may work with a causal reasoning module to infer what an opponent in an athletic competition is trying to do. A web of information exchange takes place between modules, provided that the potential information is sufficiently relevant to activate another module. Figure 10.A (below) illustrates how each module processes information and exchanges the output of its inferences with others, ultimately influencing our behavior and actions.[186]

[186] While a somewhat complex diagram, this is sadly still a vast oversimplification. There are multiple memory modules that correspond with each kind of memory (short-term, long-term, procedural, episodic, associative, and possibly even domain-specific), likely multiple behavioral and bodily function modules, modules geared towards survival, towards reproduction, some specifically towards fear-processing and reward-tracking and so forth. There is simply no way to account for the complexity of the modular mind with a small diagram.

***figure 10.A* (Modular Interaction)**



By cooperating in this information exchange, producing representations, and making inferences, our modules enable us to perform a great variety of mental operations and engage in a broad range of physical tasks.

Even though our mind-reading files are primarily oriented towards the social world, keeping detailed records of others, it is not uncommon for modules to operate beyond the scope of their evolved function, sometimes even getting co-opted by evolution to serve these accidentally advantageous alternative ends as well. The files that we open need not be restricted to *people*, for we obviously sometimes open files about animals, inanimate objects, or even fictional characters (Mercier and Sperber 2017, 97). It is this cognitive capacity—along with a few others—that makes us feel like we *know* a celebrity whom we have never met, *understand* the motivations of a character in our favorite book, or even *feel bad* for a tree that has been uprooted. Although evolutionarily unintended consequences, the advantage of such a quirky design scheme is that it allows for flexible behavior and increased adaptability to changing environments. The last thing we would want is to have a mind-reading module that is so rigid in its operations that it only functions in the presence of human beings, for if we were to evolve beyond human

beings, encounter some other social language-using creature, or develop sentient artificial intelligence, such a module suddenly becomes useless, making for a rather inefficient use of resources.

And yet, the reason that we have mind-reading and language modules at all hearkens back to our foraging roots that have imparted to us a vested interest in trying to figure out what the group wants and expects from us, even how the group thinks about us. These modules not only facilitate mind-reading but also self-promotion within the group and norm enforcement, facts that reveal themselves in our proclivity for gossiping. Louis Cozolino, for instance, cites the work of anthropologist and evolutionary psychologist Robin Dunbar, noting that 60% of our linguistic communication centers around gossip and non-essential personal information,[187] a figure that strongly suggests that some of the mental modules involved in communication have evolved for social purposes such as group cohesion and analysis of social interactions (Cozolino 2002, 155–6). Likewise, Kevin Simler and Robin Hanson explain how gossip can rally a group of people to a cause, helping to overthrow a more powerful individual who otherwise would have had the upper-hand in a one-on-one situation. In these kinds of instances, gossip serves the end of norm enforcement, harming a person's social standing by chipping away at her reputation or even leading to ostracism (Simler and Hanson 2018, 52).

10.9 | Self-Management

If we have the ability to open and manage files on other people, if the operations of modules can be co-opted for other ends, and if our social standing within a community

[187] For more on his work, see *Grooming, Gossip, and the Evolution of Language* (Dunbar 1996).

matters, it should come as no surprise to learn that our modules can even open and manage a file on ourselves,[188] a feat that can be further used to combat reputation-damaging gossip, justify our behavior to the group, and even project to others where we think we ought to stand within the hierarchy of the community. By having a file on ourselves as well as a language module that enables communication, we engage in practices whereby we selectively share our highlights, withhold information about our nadirs, manufacture excuses to salvage our reputation, rationalize why another's reputation is undeserved, and even suppress our true motives in an effort to have the best of both worlds, increased reputation *and* personal gain.

This ability to construct and manage a file on oneself is what makes the autobiographical self possible, and it is through the environmental pressure of reputation management within a social context that it has evolved. Considering our foraging past and mind-reading capabilities, Cozolino's speculation regarding autobiographical identity appears likely: "Evolution may, in fact, have selected for a single self-identity as a means of behavioral continuity within each individual as well as for the necessity of identity, predictability, and accountability within larger groups" (Cozolino 2002, 157). It is the autobiographical self that has become the ultimate tool for self-managing our reputations and negotiating our standing within our communities.

But this is not to suggest that the autobiographical self is causally inert, an epiphenomenon or mere artifact of social interactions. On the contrary, who we take ourselves to be and how we understand ourselves in our autobiographical tale have wide-ranging implications for our behavior and subsequent social interactions.

---

[188] Perhaps there is even an autobiographical *module* whose submodules consist of the mind-reading, language, and reputation modules?

Like Dennett, Douglas Hofstadter also defends the idea that we have an autobiographical self, but he is more explicit in how he understands self-representation. For him, much like the mind-reading files, the self is one more symbolic representation amongst a host of representations that we acquire about the various objects, persons, and qualities in our world. As is the case with any mind-reading file, the information associated with it can be amplified or diminished, added or subtracted, but unlike the files we keep on other objects and persons, it is uniquely self-referential, pointing back to the body and detailing its story (Hofstadter 2007, 181).

So how can an autobiographical self *do* anything?

The autobiographical self is *not* a metaphysical agent, nor is it anything like a Cartesian self. It is very much the equivalent of a file that we have on others except it details our own stories, our accomplishments and failures, talents and short-comings. The information that is in this file, however, influences the sorts of activities in which we engage, types of careers we pursue, and sorts of relationships we cultivate. If in our file are details about being athletic and having good arm strength, then it motivates us to pursue activities that correlate with those details, whether it amounts to leisurely pickup games of basketball or the pursuit of an athletic scholarship in football. And like keeping a file on other people, we also sometimes get the details about ourselves wrong, inclining us to modify the contents in light of the results from our actions. Perhaps you came to believe you had good arm strength because you physically developed quicker than your peers, but now in adulthood, you have come to realize that you have a remarkably average physique?

This back-and-forth testing of our narrative against the backdrop of reality, this continuous modification of the file that we keep on ourselves, is what Hofstadter calls the *strange feedback loop* (Hofstadter 2007, 180). The autobiographical self is an abstraction, existing only as an idea, and yet, it has profound consequences for what we do and how others see us. After relaying an anecdote about his attempt to impress his first-grade peers with a Hopalong Cassidy impersonation, Hofstadter explains:

> What we do—what our "I" tells us to do—has consequences sometimes positive and sometimes negative, and as the days and years go by, we try to sculpt our "I" in such a way as to stop leading us to negative consequences and to lead us to positive ones. We see if our Hopalong Cassidy smile is a hit or a flop, and only in the former case are we likely to trot it out again. (I haven't wheeled it out since first grade, to be honest.)
>
> When we're a little older, we watch as our puns fall flat or evoke admiring laughter, and according to the results we either modify our pun-making style or learn to censor ourselves more strictly, or perhaps both. We also try out various styles of dress and learn to read between the lines of other people's reactions as to whether something looks good on us or not. When we are rebuked for telling small lies, either we decide to stop lying or else we learn to make our lies subtler, and we incorporate our new knowledge about our degree of honesty into our self-symbol. What goes for lies also goes for bragging, obviously. Most of us work on adapting our use of language to various social norms, sometimes more deliberately and sometimes less so. The levels of complexity are endless. (184)

This self-file, this autobiographical self, is engrossed in an endless game of reputation management. It facilitates our navigation of the social environment as we learn in what ways we can flourish, impress members of our communities, and climb the social hierarchy. We continuously try to expand our repertoire of talents, skills, and abilities, refining some and abandoning others all while updating the file that we keep on ourselves.[189] In turn, we hope—sometimes consciously, sometimes unconsciously—that

[189] Lest it be lost in these sections, it would probably be far more accurate to speak of *an* autobiographical self or of autobiographical *selves*. In section 4.9, I alluded to Minsky's theory of multiple self-models and explained how each of us is a member of *many* communities with *multiple* reputations to manage. It

honing these qualities and spinning these self-tales will enable us to get ahead in the world.

It is in the construction of the autobiographical self wherein the possibility for a different kind of agency emerges, *rational* agency.

10.10 | Self-Determination

In the *Timaeus*, Plato departs from his conventional treatment of philosophical topics by ultimately removing his usual protagonist, Socrates, from the conversation. Instead, we along with Socrates are treated to a speech delivered by Timaeus, who waxes philosophically, scientifically, and poetically about the creation and order of the cosmos itself. The cosmos, Timaeus tells us, "is the most beautiful of things born," and yet, he recognizes that this cosmos is in a constant state of change, naturally tending towards chaos and disorder (*Timaeus*, 29a; 30e). How can he lay hold to both of these claims without contradicting himself?

The solution that Timaeus casually introduces, as if it were obvious, is that there must be a *dêmiourgos*, some kind of godlike being who is always at work, meticulously crafting and sculpting the cosmos to make it as fine as possible (*Timaeus*, 28a–b). Like any good craftsmen, the *dêmiourgos* looks to a model for inspiration, one that happens to be perfect in every way. And yet, this being's task is cruelly Sisyphean in nature, for the cosmos, ever prone to decomposition and disarray, requires constant attention and care to remain well-ordered (29e–30b). In spite of the fact that this task seems futile, the

follows from that argument that we also have multiple self-files that we manage. Perhaps a lofty goal is autobiographical integrity, a rendering consistent of our many self-files?

dêmiourgos works every night and day, humbly, willingly, and free from any resentment (29e).

One of the most pressing existential questions each of us faces as a human being is: how should I live? The question itself might be a function of our various modules at work, unconsciously (sometimes even consciously) motivated by our various evolutionary concerns for enhanced reputations, reproductive opportunities, and even just a brute sense of survival. And yet, it is a question that we *can* ask, and a question that we can *try* to answer. It is a question of indispensable importance for who we are and who we come to be, how we interact with others, and which life choices we make.

It was suggested in section 3.4 that our capacity to consciously tend to mental representations equips us with the ability to contrast the present with possible alternatives, a key cognitive feature that aids in planning and problem-solving. An arguably welcome side-effect of this, along with our capacity for self-representation, is that we can also conceive of *possible* selves, who we would like to be. Do you wish to be an entrepreneur? An orator? A statesman? Can you envision a future wherein you take responsibility for a family? Employees? Students? Is your ideal home situated in a rural area? The heart of the city? Or perhaps somewhere in-between? There are multitudes upon multitudes of possibilities that we can entertain as potential candidates for incorporation into our autobiographical stories, provided that we, like the *dêmiourgos*, look to those possibilities as models and work diligently to fashion our lives after them.

And yet, as Plato warns in the *Timaeus*, not just any model will do. The *dêmiourgos* has it easy, for it merely has to decide between the model of *Being* or

Becoming,[190] only one proving to be a fitting blueprint for the cosmos, but in our own cases, there is a nigh-limitless number of such models to consider. Is there just one model appropriate for all of us? Just one unique model for each of us? Or can multiple models suffice for the same person?

In the *Nicomachean Ethics*, Aristotle rhetorically asks what it is that human beings hope to attain in all of their endeavors. His answer is *eudaimonia*, a state of "living well and doing well."[191] This is not simply a state of mind, such as a feeling of happiness or satisfaction, but on the contrary, as philosopher Richard Kraut suggests, it is a matter of evaluation, of determining *objectively* what contributes to and promotes well-being (Kraut 2018, 5).

This subtle distinction between subjective feelings and objective results is important if we are to understand *eudaimonia* in a post-Darwinian world. To reiterate a previous point, evolution always selects for results. If evolution selected for us to strive for *eudaimonia* in some way, it can only be *because* doing so promotes fitness—unless it is little more than a non-adaptive rewarding by-product of fitness-enhancing activities. Were it a feeling alone, not consistently correlated with actions and changes in the external environment, it would do nothing, and if it does nothing, there is no reason for natural selection to care. A good case in point is that of self-esteem. Citing a systematic review from researchers Thomas Scheff and David Fearon,[192] Robert Kurzban explains

---

[190] See *Timaeus*, 28c–29a.

[191] In what follows, I am not attempting to offer an authoritative account of *eudaimonism* or, in the next section, *deliberation*; rather, my intent is to borrow some Aristotelian-inspired ideas in an effort to contribute to my theory of agency and the importance of our autonomy in autobiography.

[192] See "Cognition and emotion? The dead end in self-esteem research" (Scheff and Fearon 2004, 73–90).

how over 15,000 studies on self-esteem (a subjective feeling) yielded inconsistent and inconclusive findings, suggesting that self-esteem is not demonstrably a cause of anything (Kurzban 2010, 137–8). With a theory of evolution in hand, therefore, *eudaimonia* must be interpreted as something objective and results-oriented, such that conceiving of it is to conceive of what it means to live in the world and live well. Interpreted in this way, it might be exactly the evolutionary carrot that self-representing creatures need, especially if part (or all) of it involves social flourishing.[193]

What is it that promotes the well-being characteristic of *eudaimonia*?

The answer to this question was as much a debate in ancient philosophy as it is today. Philosopher Miira Tuominen, following the ancient commentator Aspasius, believes that the best way to cultivate *eudaimonia* is to consciously recognize it *as* the goal, subjecting all of our decisions and choices to it, rather than letting ourselves be distracted by secondary goals that may or may not contribute to it. To do this, we must seek out a *eudaimonistic* model after which we can fashion ourselves; it is a model that attends to our needs as human beings (Tuominen 2009, 244). Though we may not be able to answer precisely what promotes *eudaimonia* in every context for each human being, we can at least identify what makes for a good *eudaimonistic* self-model.

It was not an uncommon view in ancient Greek philosophy to assume that being virtuous is necessary for attaining a state of *eudaimonia* (Bobonich 2002, 210). A common misunderstanding today, however, is to assume that being virtuous is equivalent to being *moral*. While one might conclude that *part* of being virtuous is being a morally good person, this is far too narrow a restriction on virtue than the ancients intended

[193] I do not think it is even possible to conceive of what it means for a social organism to live well in the absence of a community.

(Tuominen 2009, 241–2). Virtue, or *aretē*, is literally translated as *excellence*, and so for Aristotle, a virtuous human being is an *excellent* one, someone who exercises well what it means to be who they are (*Nicomachean Ethics*, $1097^{b}23$–29). Just as Luciano Pavarotti had exemplified what it means to be an excellent operatic tenor, Aristotle wonders whether and how a person can exemplify what it means to be an excellent human being in general.[194]

Perhaps there are multiple ways to attain a state of *eudaimonia*? Might an orator, a statesman, or a physician each have their own idiosyncratic ways of being excellent, ways that hinge on social, professional, and personal contexts? Might virtue, in other words, be a matter of determining what it takes to be excellent *as* an orator, *as* a statesman, or *as* a physician?[195] How does one even begin to divine in what the fullness of a life consists?

---

[194] On this topic, Aristotle himself is regrettably unclear. Is it *eudaimonistic* to live a life of action, one in which we let prudence, courage, temperance, and justice guide our decisions so as to help secure the best outcomes in our practical affairs? Or perhaps to live a life of contemplation, to be like the gods who have no concern for such petty mortal affairs but who instead spend their time exercising their intellects alone and thinking about what is good?

Both Tuominen and Jonathan Lear summarize this tension in Aristotle's work. According to Tuominen, Aspasius concludes that human beings must be excellent in both ways in order to attain *eudaimonia* and be virtuous human beings (Tuominen 2009, 246–7). Lear, however, reads Aristotle as advancing the cognitivist thesis that excellence in practical affairs is only to be desired for the sake of maximizing our time to do what we really want to do: contemplate (Lear 2000, 43–8).

Keeping in mind that evolution selects for results, the cognitivist interpretation is almost wholly unappealing, regardless of whether it was what Aristotle really intended.

[195] This is somewhat how philosopher Martha Nussbaum understands *eudaimonia*. She writes:

> In a eudaimonistic ethical theory, the central question asked by a person is, "How should a human being live?" The answer to that question is the person's conception of *eudaimonia*, or human flourishing, a complete human life. A conception of *eudaimonia* is taken to be inclusive of all to which the agent ascribes intrinsic value: if one can show someone that she has omitted something without which she would not think her life complete, then that is a sufficient argument for the addition of the item in question. (Nussbaum 2001, 31–2)

Part of the answer lies in selecting a suitable self-model to use as a guide for living one's life, but accomplishing this requires appropriating and adjusting another one of Aristotle's ideas: *deliberation*.

10.11 | Self-Modification

There are many things that do not count as objects of deliberation. We do not deliberate about the nature of the cosmos, the laws of the known universe, principles of mathematics, things that happen by necessity or chance, or even the affairs of others (*Nicomachean Ethics*, 1112^{a}20–35). Instead, Aristotle explains, "We deliberate about things that are up to us and are matters of action" (1112^{a}31–32). Does this mean, then, that we deliberate about whatever it is that we can *do*? Not quite.

Aristotle makes two additional qualifications.

First, he explains, we deliberate about those things for which there is more than one way to do them, such as "things done by medical skill or skill in business" or even "piloting a ship" (*Nicomachean Ethics*, 1112^{b}5–6). Whether you are plotting the direction of your business or the route to your favorite destination, it is clear that this is quite unlike spelling a word that you have in mind, for (as he gently reminds us) there is only one way to go about spelling a word—the letters must be arranged in the correct order (1112^{b}2–3). It is pointless to spend any time weighing and evaluating alternatives when there are none. Thus, deliberation requires that there are multiple possibilities to entertain and consider.

Second, Aristotle elaborates that we "deliberate not about ends, but about what promotes ends" (*Nicomachean Ethics*, 1112^{b}12). To clarify what he has in mind, he explains that a physician *already* has a goal to heal. The physician does not need to

calculate whether or not she *ought* to heal—that is part of her identity, who she is—rather, the object of her deliberation ought to be *how* to heal. Does she prescribe medication or therapy? Does her patient need some sort of diagnostic testing? Does she need to surgically intervene to promote her patient's health?

> The same is true for any other vocation (*Nicomachean Ethics*, $1112^{b}13$–14).

The orator, for instance, *already* has a goal to persuade. The question the orator is asking whenever she hopes to be a good orator is how she can go about promoting this goal. Which topic does she wish to present to her audience? Does her argument need to be rearranged in any way? Should she begin with an anecdote or present some provocative statistics?

Likewise, the politician has a goal of proper governance, and so she ought to deliberate about *how* to govern. Does she promote the ideals of a democracy or a monarchy? Is it better to focus the state's resources on domestic affairs or foreign? Should she adopt a diplomatic approach to rival nations or project an image of strength and military might?

Necessarily biologically programmed with the end of *eudaimonia* as a concern, each of us deliberates, wittingly or unwittingly, about how to attain such a state of affairs. To do so intentionally requires systematically setting down additional goals as we entertain the myriad possibilities that seem open to us.[196] You might narrow down your choices by concluding that living well consists in becoming a physician, an orator, or a

---

196 In what follows, I will be taking cues from philosopher David Wiggins' interpretation of deliberation. In his article "Deliberation and Practical Reason," he suggests that deliberation is a process that begins once we have already settled on a goal or end (*telos*), but it does not follow that we can *never* deliberate about the end itself—although there is one exception: *eudaimonia*. While it is engrained in our nature to strive for it, *how* we do so is indeed up to us, requiring that we deliberate about increasingly fine-grained goals that support that end (Wiggins 1980, 227).

statesman, and in so doing, think through what each of these choices requires given your current set of circumstances. If you wish to be a physician, what is it that you need to *do* right now? How, furthermore, can you become an *excellent* physician? Deliberation is thus ultimately a matter of searching for the good life of *eudaimonia*, moving from a vague conception of what it means to live well to something that can influence our decision-making, that guides us in what we do and how we tell our life stories.

Perhaps this is the sense that Aristotle intends when he tell us that the orator, physician, and statesman already have goals that are beyond deliberation? Philosopher David Wiggins suggests:

> Each of these three gentlemen, the orator, doctor, or statesman, has *one* telos (for present purposes). He is already a doctor, orator, or statesman and already at work. That is already fixed (which is not to say that it is absolutely fixed), and to that extent the form of the eudaimonia he can achieve is already conditioned by something no longer needing (at least at this moment) to be deliberated. (Wiggins 1980, 226; emphasis original)

Neither the orator, nor the doctor, nor the statesman deliberate about their goals of persuasion, healing, and governance, respectively, because each of those aims specifies what is required to be virtuous, to be excellent *as* an orator, doctor, or statesman. If the orator fails to be persuasive, for example, she will not find satisfaction *as* an orator, and so the goal of persuasion for anyone who insists on living her life as an orator is non-negotiable.

Aristotle warns that deliberation, if it is done well, is as challenging as discovering a mathematical proof, but no matter how difficult, how laborious, how, perhaps, even Sisyphean a task, there is little doubt that this is what it means to be a rational agent in the fullest sense of the word (*Nicomachean Ethics*, $1112^{b}20–23$). Deliberating is to take seriously our form of existence as autobiographical, self-

determining beings. Parts of our stories will no doubt include bouts of madness, fits of rage, nonsensical worries, and perverse activities, but such are the consequences of having chaotically beautiful, beautifully chaotic minds. This does not, however, abdicate us of our responsibility to *try* to make sense of our lives and to be excellent in the myriad ways demanded of us. We are, after all, social creatures, creatures who have evolved with a desire to be a part of a community. It is deliberation for the sake of *eudaimonia* that challenges us to make sense of it all, to integrate it into a *good* story.

10.12 | Conclusion

The modular thesis of the mind appears at first glance to recommend that we discard any notion of a self, regarding it as yet another empty term without a reference, as if it were a mere metaphysical fiction concocted by thinkers such as Descartes. Modules, having been carefully selected for and refined by the processes of natural selection, guide the organism in its efforts to navigate the environment. Given the way that they operate, not all information is shared between them, and so the divided mind is more akin to being hopelessly fractured than it is compartmentalized in the way that Davidson had imagined.

It would be short-sighted, however, to dispense with the idea of a self altogether.

Modules did not evolve in a vacuum. Rather, they have evolved as part and parcel of a living body whose fitness they promote. Their operations ultimately contribute to the successes or failures of the organism within its ceaselessly changing environment, and when they work well, they are refined and carried over into the next generation of organisms. Though navigating the external environment is one job, another is to manage the *internal* environment, the carefully calibrated internal milieu whose operational parameters are far narrower than the environmental conditions in which the organism

itself lives and moves. This internal milieu, encapsulated by bodily tissue, is what individuates any organism from the rest of the world, and if anything can provide for the continuity of reference required for a self, it is the body proper.

Some organisms, however, are also capable of extending the boundaries of themselves into their environments, appropriating portions of it to make it their own, and in the case of human beings, there are abstract extensions of our bodies known as autobiographical selves, self-representations composed of stories and ideas that tell the tale of the body, giving it a history and, by extension, a reputation that aids in justifying just who one is to the members of her community. Because such self-identity and self-understanding has implications for what we do and how we relate to others, we have a responsibility to acknowledge it and take ownership for it, to fashion it as best we can in the service of living well.

Conclusion: Rational Creatures and Social Organisms

My dear Cratylus,
I have long been suspicious of my own wisdom,
I cannot trust myself.
This is why I think it's necessary to stop and ask myself: "What am I saying?"
For there is nothing worse than self-deception,
when the deceiver is always right at home with you and never leaves.
It's quite terrible, and therefore, I think
we must retrace our steps and test them
by looking, 'fore and aft,' in the words of Homer.
Now let me see;
where *are* we?

— Socrates, *Cratylus*

11.1 | Dare to Dream

The historical Enlightenment ended in 1793, even though its spirit lives on. The essence of this spirit is a daring embrace, a blind faith, in a handful of romanticized ideals—peace, prosperity, progress, and reason. There is little doubt that such pursuits are noble, motivated by a hope that our commitment to them will bring about a new age of human flourishing, and to an extent, though it is far from perfect, the quality of human life around the world has improved immensely from social and political movements that have cherished some of these ideas. Some of the basic necessities of life in a modern world—food, medical care, transportation, education—is more readily available to people today than ever before. As a first-hand witness to the beginning of these social changes and scientific discoveries, Kant was careful to qualify his opinion on the matter:

> The question may not be put: Do we live at present in an enlighten*ed* age? The answer is: No, but in an age of enlighten*ment*. (Kant 1784, 140; emphasis added)

In 1784, he cautioned, we are not quite there.

And yet, in spite of these momentous achievements, our commitments to these ideals for more than 200 years have failed to deliver us to the promised utopia supposedly

awaiting us in the near future. The reason why, Enlightenment defenders will say, is that we have not collectively embraced the ideals *enough* or that we need *more time*. Hegel, however, pointed to a different reason: such perfect ideals are not suited for imperfect creatures. Ideals are, of course, necessary insofar as they play an indispensable role in the evolution of our stories, but at some point, such ideals wear out their usefulness, requiring that we look elsewhere.

Perhaps a contradiction is buried, waiting to be uncovered, in our pursuits of peace and prosperity or our beliefs in equality and progress? Whatever the answer to that question may be, this much is certain: it can no longer be denied that our faith in reason has been deeply flawed. It is this ideal that unifies so much of Enlightenment thought, every success attributed to its well-exercised application and every failure explained away, and it is this ideal that continues to perpetually frustrate us, making the actual world of human experience, a world far removed from the ideal, appear so bewildering and strange.

Should we persist in such a self-defeating task?

One of the fears of abandoning the Enlightenment is that it somehow implies a return to a pre-scientific dark age, but this could not be further from the truth. The Scientific Revolution itself was well underway before the Enlightenment began, and it continues to this day, long after the historical Enlightenment ended. In fact, much of the progress in living conditions attributed to the Enlightenment might be better attributed to our faith in the scientific enterprise. Our species has been blessed with a number of outstanding scientific minds who have contributed to our understanding of the world from Copernicus, Kepler, Galileo, and Boyle to Mendel, Curie, Einstein, and Hawking,

and considering the periods during which each lived, it can hardly be argued that science has made progress both because of *and* in spite of the Enlightenment.

11.2 | From Gears and Sprockets to Germs and Seeds

Since 1793, one of the most important scientists who fundamentally changed the way that we understand the world around us was none other than Charles Darwin.

Darwin showed us that we belong, *all* organisms, to a much larger web of life, a web spun by an everchanging environment that challenges us to respond in kind, and though Enlightenment thinking might incline us to believe that infallible reason separates us from the rest of the animal kingdom, nothing could be further from the truth.

A story of human action in a post-Darwinian world must take these scientific insights into account. Human action does not spring forth from some mysterious power that supervenes on the natural world, but is a product of its richness and dynamism. If Darwin's insights mean anything whatsoever, then we should expect to find analogues to our complex behavior and mental processes throughout the rest of the biological kingdom, from the large to the small.

And indeed we do.

Rats appear to use a form of mathematical cognition, raccoons and crows exhibit a proclivity for cleverness in problem-solving, ants effortlessly keep track of the location of their home, and bumble bees show that they are capable of learning where predators may be hiding. In so many of these creatures we find, writ small, a number of meaningful behavioral and cognitive capacities that bear remarkable similarities to our own skills and abilities.

In an effort to preserve the importance of human reason, Enlightenment thinking writes these observations off. It argues that such creatures are little more than biological machines, programmed to elicit fixed responses to environmental stimuli. These non-rational creatures, we are encouraged to believe, are causally determined in their behavior from beginning to end, as if somehow the actions produced by human beings are miraculously exempt from the unforgiving grind of the wheels of determinism. Unfortunately, herein lies one of the usual paradoxes for Enlightenment: determinism or freedom? Without one's worldview collapsing into Cartesian dualism, it is unclear how one can maintain that human beings have metaphysical freedom while simultaneously insisting that the activity from the rest of the biological kingdom is a product of mechanical forces.

Here, however, Darwin once again provides a solution if we look closely enough. If we consider that novel properties can emerge out of the complex interactions and organizations of matter—if we do not *assume* that matter is merely mechanical in nature—the difficulty is removed. And there is every reason to embrace this non-determinist theory of nature. Never mind how bizarre matter behaves at the quantum level, for even at the macro level, it is evident that as organisms scale in biological sophistication, there is a transition from an aggregate of parts acting independently to the unity of an organism acting as a *whole*, a marvel made possible through the neurobiological mechanisms of *cognitive integration*. This, I have argued, is the seed from which action springs.

So what it is that makes *human* action possible? Not reason *per se*, but *mental representation*.

11.3 | A Reason to Reconsider

Davidson, following in the tradition of Enlightenment thinking, defended the idea that intentionally acting is acting *for a reason*, that human action is as simple as having a desire, having a related set of beliefs, and putting those beliefs into action. Coupled with his assumption of lingualism, that beliefs must be linguistic in nature, we arrived at the conclusion that action is exclusive to language-users, a conclusion that is inconsistent with our best understanding of animal behavior (especially, for example, de Waal's Capuchin monkeys) and neurobiology.

But, setting aside those scientific insights, suppose for a moment that Davidson is correct. Then what?

From observations of the behavior of patients with brain injuries as well as findings from cleverly designed experiments by psychologists and scientists (such as Norman Maier and Benjamin Libet), from the phenomena of habits to states of flow, it appears that *most* of our behavior happens in the absence of any rational planning as imagined by Davidson. If his theory of action is correct, then it looks like we are seldom acting intentionally. But, unless determinism is true after all, this seems absurd.

Instead, the solution I had proposed was to reconsider our understanding of intentional action. Rather than defining intentions in terms of reasons, we should think of intentions as undertaking a commitment to any kind of mental representation that can be consciously entertained, whether it is an imagined state-of-affairs or even a vague, emotional inclination. Following Crick and Koch, one of the functions of consciousness appears to be to provide an organism with an executive summary of the here-and-now, allowing for opportunities to selectively highlight features of the environment so as to

respond accordingly. In addition, it was argued, mental representations provide a point of contrast to this here-and-now experience, enabling the organism to conceive of alternatives and not just respond but *creatively* respond to its environment. This is the essence of an intentional action.

What is reason's role?

There is no doubt that the possession of language makes for an additional kind of mental representation, *linguistic* representations, and reasons, as constituents of *arguments*, certainly belong to the category of linguistic representations. But considering our evolutionary roots and the fact that the use of reason tends to occur *ex post facto*—that is, *retrospectively* rather than *prospectively*—it begins to look as though reason's role in the production of actions is far more limited than we had hitherto traditionally supposed.

If we are not wholly or even primarily rational creatures, then who are we? To answer this question, we turned to the dark side of human behavior, the irrational. Not only must we reconcile our self-identity with the fact that reason is not as fundamental as we have long believed, but there is also a need to recognize that we often fail in our efforts to make sense of ourselves and others.

11.4 | Unreasonable Expectations

Thinking and acting irrationally is not the same thing as having a mental illness. Mental illnesses might very well *cause* one to think or act bizarrely, but we can also do this under otherwise normal and healthy circumstances. Augustine, for instance, marveled at how difficult it was to understand his own behavior at times. This is a function of the

complexity of our minds, and this is something that a modularity thesis explains rather elegantly.

Enlightenment thinking, by contrast, has a storied history of struggling to make sense of our unusual thoughts and behavior. The traditional approach has been to assume that there is some kind of mechanical malfunctioning or a mysterious secondary process that works alongside a rational one. When considering the nature of reason, Locke, for instance, wonders how it is possible that otherwise rational people could make errors and draw bizarre conclusions. His only solution is to posit the existence of non-rational processes that operate alongside reason, but this only introduces a paradox, for it remains unclear how these two processes, one mechanical and the other autonomous, could possibly interact. Other thinkers, such as Davidson, attempt a reductionist approach, arguing that irrationality is just rationality gone astray, as if there is a good explanation for why that occurs. None of these proposed solutions work very well, each serving to undermine the Enlightenment's commitments to human freedom and the power of human reason. When we acknowledge the existence of irrational behavior, reason begins to appear impotent in the face of these irrational processes, a conclusion that even troubled Locke.

One of the obstacles to understanding irrational phenomena has thus been the uncritical acceptance of Enlightenment ideas regarding the nature of human beings, especially reason. Neither wishful thinking nor self-deception, for example, make sense as far as truth and practical success are concerned. Why would we persist in our beliefs that something is true when it is actually false? Why would we adopt a high-risk / high-reward strategy of misrepresenting what we believe when it proves so costly in the event

that our bluff is called? Even from an evolutionary perspective, neither strategy is very productive if we are thinking in terms of survival in a hostile natural environment; but in a *social* environment, where reputations are meaningful, all of a sudden such behaviors appear far more sensible. We engage in wishful thinking to enhance our reputations within communities and self-deception to avoid potential fears; and sometimes, such strategies backfire, making our lives all the worse. Neither of these strategies are deployed consciously or deliberately; they are part and parcel of how our minds work, products of multiple modular processes, some with competing and others with shared interests.

Perhaps the clearest failure of Enlightenment assumptions, though, can be found in its handling of akrasia. Tasked with needing to explain why reason is failing, such thinkers retreat to the position that an akrates is little more than a person who does not *really* know what is best when she acts against her better judgment, that the akrates, in other words, does not know *enough*. Yet, psychological research shows that there exist nonrational states and processes that exercise quite a bit of influence on our decision-making. From hindsight and information-processing biases to bodily states such as hunger or thirst, from how situations are framed to the here-and-now presence of the things we desire, our decision-making can be profoundly altered in ways that incline us away from our avowed beliefs in otherwise normal circumstances. To his credit, even Davidson had difficulty accepting the traditional position, ultimately proposing that we needed to divide the mind—a rather non-Enlightenment solution.

11.5 | Adapting to Norms

So many of these difficulties stem from embracing Cartesian assumptions regarding the mind. It was Descartes who principally introduced the Enlightenment to the ideas that the mind is a seamless unity, fundamentally rational, and wholly transparent to itself. Beginning with these axioms, it becomes hard to imagine how human action can be anything other than that which is done for a reason, and yet, these three assumptions, taken together, have only made human behavior *less* intelligible, not more. There is nothing short of irony, then, in the fact that Enlightenment thinking contends that these Cartesian ideas are necessary to establish the lawlike connection between beliefs and actions that makes behavior intelligible, thereby grounding the normativity required for our social practices.

But does the alternative, the modular thesis, fare any better? Might proposing a divided, chaotic mind issue in its own variation of Kant's Third Antinomy, leaving us to wonder whether anything is truly gained from a transition to a post-Darwinian perspective?

Recall that the modular thesis celebrates the creativity that a divided mind makes possible. It is this creativity, I had argued, that equips us with the ultimate tool to respond to unpredictable environments and novel situations. Although Kant's Third Antinomy illustrates the logical tension between necessity and freedom, it appears that the modular thesis results in its own logical tension between necessity and *creativity* (see figure 11.A). If the mind has evolved for creativity, then does it not follow that there really is no lawlike connection between beliefs, desires, and reasons? What does this mean for normativity and our social practices of accounting for oneself and holding others

responsible? Though the modular thesis gains the advantage in explaining *individual* behavior, does it come at the cost of forfeiting the normative practices that underwrite so many of our social institutions?

What, in other words, does this post-Darwinian solution mean for normativity?

When it comes to reason, after reflecting on its failures, the fact that it tends to be employed retrospectively, and that when it is used prospectively, it tends to be a matter of anticipating a need to justify oneself to others (rather than thoughtfully intending to perform an action), Mercier and Sperber suggest that we ought to think of reason's role as primarily *persuasive* rather than *truth-oriented* (Mercier and Sperber 2017, 112; 138). They reject the idea that people have implicit reasons for acting on the grounds that such talk introduces an empty term used to protect the assumption that people, in general, must be thinking and acting on the basis of reasons. On the contrary, they argue, it is the way in which modules have evolved that helps ensure that they tend to function well and produce good inferences under normal circumstances, and so it is on account of *that* that behavior can be rationalized or reasons discovered *ex post facto* (118). Our modules, in other words, evolved to efficiently address the demands that our environments place on us, and so it is no wonder that we can often describe our behavior in terms of reasons and construe it as reasonable. We very often successfully and efficiently overcome real-world obstacles because our modules evolved to do just that.

It is true that the practice of giving and asking for reasons is a necessary precondition for establishing and enforcing social norms, but all that is required for this practice to flourish is *persuasion*, not *truth*. This practice does not necessitate that we act *for* reasons; it merely requires that we be able to *justify* ourselves, to make ourselves

appear reasonable to others, regardless of whether such justifications occur *ex post facto*, as they so often seem to do. Erroneously believing that reasoning for the sake of justification entails that we act *on account of reasons* at the time of acting is what makes human behavior so puzzling, strange, even unbelievable.

Nothing, therefore, is lost in the post-Darwinian transition. As social organisms, we have evolved to establish and participate in normative practices, and reason (along with other modules) has evolved to help us do just that.

11.6 | We Live in a Place Between Being and Not-Being

While there is much to commend in Lear's proposal that an irrational mind is a good mind insofar as it affords an organism more flexibility to adapt to a changing environment than is available to a rigid and rational one, he ultimately overstates its irrationality. It is true that our mind is a messy one that often disrupts itself, and this is a consequence of the mind's modularity. Mental modules, however, are not *irrational*. It would be absurd for a facial recognition module to evolve to be bad at what it does. On the contrary, modules have evolved to perform their tasks with great efficiency (the bad modules were passed over by natural selection). The alternative, however, is not a *rational* mind, an idea that is simply untenable when we consider the many competing aims of our modules as well as informational encapsulation—the fact that modules do not share their information with one another. Where does this leave the picture of the mind?

In trying to understand how an unconscious mind works, the psychiatrist Linda Brakel introduces the term *arational*. Following Freud, she believes that unconscious mental activity operates associatively without any concern for the logical, sequential thinking common of conscious mental activity (Brakel 2009, 7–8). The reason she opts

for *arational* instead of *irrational* is because she believes, following Davidson, that irrationality is "rationality gone astray." But, she explains, this is not a category that applies to the "not-yet-rational" infants or the "never-to-be-rational / but good enough for survival" mental states of animals. Both, she insists, are under the sway of the activities of the unconscious (7–8, n.6). The advantage of introducing such a term is that it allows for a form of mental activity that can produce a variety of outcomes: irrational, rational, or even neither.

Although modularity deviates quite a bit away from the dual-process thinking Brakel seems to endorse, arational is perhaps the best way to describe the activities of a modular mind as well. Modules do what they do well, and they have no concern for whether their work hinders or helps the abilities of other modules to do what they are trying to do. Through their evolution under environmental pressures, they *tend* to work together, but such constant, universal cohesiveness is not what has been selected for. As we have seen in those who suffer from traumatic brain injuries, such as Phineas Gage, a mind composed of independent modules that work on their own is more advantageous to survival. And, as Lear argues, a chaotic mind is what makes creativity possible.

The unhappy consequence of a modular mind, however, is that this also creates the conditions for regular contradictions and teleological conflict. Beliefs engendered by the activities of some modules could very well contradict beliefs engendered by others without an individual necessarily realizing that she is committed to incompatible beliefs. Likewise, the inferential activity from a risk / reward module could incline an individual under the right circumstances to undertake a course of action that is otherwise against her better judgment. When the Enlightenment faith in man's fundamental rationality is

abandoned for the evolutionary necessity of man's arational mind, it becomes clear how and why irrationality is simply a part of being human.

Does this imply that we are helpless, at the mercy of modules?

Only if we remain committed to the Cartesian idea that each of us is a *captain* of the ship rather than the ship itself. There is no distinction to be made between our modules and ourselves any more than there is a distinction to be made between our organs and our bodies. We are, each of us, *one* organism in spite of our many cells individually performing their own activities. Though our thoughts hail from all sorts of unknown places, conscious awareness tends to filter out the extraneous information, providing us with an experience that appears continuous and ordered; and the narratives that we construct to aid in self-understanding tend to portray us as integrated, whole persons.

Modularity, however, alerts us to the numerous unconscious influences on our decision-making, and it leaves us with the question as to whether we can even do anything about these influences, whether we can ever learn to see through our intellectual blind spots. In Fischoff's experiments on hindsight bias, for example, those made aware of these unconscious influences had still managed to succumb to them.[197] Likewise, when alerted to the above-average effect, social psychologist Emily Pronin and colleagues discovered that participants continued to evaluate themselves at being better than average at being susceptible to it.[198]

---

[197] See section 6.6.

[198] See section 4.9.

Yet, if anything, the lesson from modularity, neurobiology, and numerous psychological experiments on information-processing and thinking is not so much that we are *rational* creatures but that we are *social organisms*. We may not be able to, on our own, detect our intellectual blind spots; however, individual improvement comes not through Cartesian meditation but social interaction. It is from the perspective of others that we can become aware of our inconsistency, our norm-violating behaviors, and our shortcomings, and in the interest of flourishing within a community, we have every incentive to respond accordingly. Part of this involves, at times, experiencing a reflexive breakdown, at utter loss of words and failure in self-understanding that shatters any misconceptions that we have of ourselves, and though at times painful, it frees us to move forward and begin again the process of integrating our disparate behavioral seams into a coherent story.

Still, even though we are not *fundamentally* rational, there is a sense in which we can be reasonable through making ourselves intelligible to others. This occurs by caring about what we do, how we live, how we understand ourselves—by, ultimately, participating in writing our autobiographical stories. Our considerations, like those of Timaeus' *dêmiourgos*, ought to take into account what is best so that we can discover how to live well. It is in this unique form of taking responsibility for who we are that rational agency takes on a new meaning (and perhaps the only meaning). Provided that our efforts do not give way to a neurotic obsession with inhuman perfection, such as how the zealous devotion to Enlightenment ideals culminated in *The Terror*, our careful, deliberate pursuit of the good life in writing our autobiographies, tempered by an

understanding of our wild and wonderful humanity affords us with the best chance to discover that for which we strive in all of our endeavors: *eudaimonia*.

Bibliography

Ainslie, George. 2001. *Breakdown of Will*. New York: Cambridge UP.

Alicke, Mark and Govorun, Olesya. 2005. "The better-than-average effect," *The Self in Social Judgment* 1: 85–106.

Alvarenga, Débora, Kelly Lambert, Stephen Noctor, Fernanda Pestana, Mads Bertelsen, Paul Manger, and Suzana Herculano-Houzel. 2017. "Dogs have the most neurons, though not the largest brain: Trade-off between body mass and number of neurons in the cerebral cortex of large carnivoran species," *Frontiers in Neuroanatomy* 11: 118.

Ambrose of Milan. 2003. *Hexameron, Paradise, Cain and Abel*. Translated by John Savage. Washington, DC: The Catholic University of America Press.

Anscombe, G.E.M. 1957. *Intention*. Cambridge, MA: Harvard UP 2000.

Aquinas, Thomas. 1995. *Commentary on Aristotle's Metaphysics*. Translated by John Rowan. Notre Dame, IN: Dumb Ox Books.

Aristotle. 1979. *Aristotle's Metaphysics*. Translated by Hippocrates Apostle. Grinnell, IA: Peripatetic Press.

——— 1999. *Nicomachean Ethics*. 2nd Edition. Translated by Terence Irwin. Indianapolis, IN: Hackett Pub.

——— 2002. *Nicomachean Ethics*. Translated by Joe Sachs. Newburyport, MA: Focus Pub.

Armstrong, David. 1973. *Belief, Truth and Knowledge*. New York: Cambridge UP.

——— 1980. "The Nature of Mind," in *The Nature of Mind and Other Essays*. Ithaca, NY: Cornell Publishing, 295–302.

Augustine of Hippo. 1960. *The Confessions of St. Augustine*. Translated by John Ryan. New York: Doubleday Religion.

——— 2008. *The Augustine Catechism: The Enchiridion on Faith, Hope, and Charity*. Translated by Bruce Harbert. Hyde Park, NY: New City Press.

——— 2012. *The Confessions: Saint Augustine of Hippo*. Ignatius Critical Edition. Edited by David Meconi. Translated by Maria Boulding. San Francisco: Ignatius Press.

Aylesworth, Gary. 2015. "Postmodernism," in *The Stanford Encyclopedia of Philosophy* (Spring 2015 Edition). Edited by Edward N. Zalta. https://plato.stanford.edu/archives/spr2015/entries/postmodernism/.

Baumhart, Raymond. 1968. *An honest profit: What Businessmen Say about Ethics in Business*. New York: Prentice-Hall.

Bell, David. 2012. "The conductor," (review of Peter McPhee, *Robespierre: A Revolutionary Life*), an online review at *The New Republic*. https://newrepublic.com/article/100710/robespierre-revolutionary-life-peter-mcphee.

Bennett, Jonathan. 1999. "Introduction to events," in *Metaphysics: Contemporary Readings*. 1st Edition. Edited by Steven D. Hales. London, UK: Wadsworth, 319–324.

Bergmann, Olaf, Ratan Bhardwaj1, Samuel Bernard, Sofia Zdunek, Fanie Barnabé-Heider, Stuart Walsh, Joel Zupicich, Kanar Alkass, Bruce Buchholz, Henrik Druid, Stefan Jovinge, and Jonas Frisén. 2009. "Evidence for cardiomyocyte renewal in humans," *Science* 324(5923): 98–102.

Bermúdez, José Luis. 2003. *Thinking Without Words*. New York: Oxford UP.

Binmore, Ken. 2007. *Playing for Real: A Text on Game Theory*. New York: Oxford UP.

Biran, Iftah and Chatterjee, Anjan. 2004. "Alien hand syndrome," *Archives of Neurology* 61: 292–294.

Biran, Iftah, Tania Giovannetti, Laurel Buxbaum, and Anjan Chatterjee. 2006. "The alien hand syndrome: What makes the alien hand alien?," *Cognitive Neuropsychology* 23: 563–582.

Bird, Christopher and Emery, Nathan. 2009. "Rooks use stones to raise the water level to reach a floating worm," *Current Biology* 19: 1410–1414.

Bobonich, Christopher. 2004. *Plato's Utopia Recast: His Later Ethics and Politics*. New York: Oxford UP.

Bogen, Joseph, Fisher, ED, and Vogel, Philip. 1965. "Cerebral commissurotomy: A second case report," *Journal of the American Medical Association* 194: 1328–1329.

Bogen, Joseph and Gazzaniga, Michael. 1965. "Cerebral commissurotomy in man: Minor hemisphere dominance for certain visuospatial functions," *Journal of Neurosurgery* 23: 394–399.

Bollas, Christopher. 2009. *The Evocative Object World*. New York: Routledge.

——— 2009. *The Infinite Question*. New York: Routledge.

Brand, Myles. 1984. *Intending and Acting*. Cambridge, MA: MIT Press.

Brandom, Robert. 1994. *Making it Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, MA: Harvard UP.

——— 2000. *Articulating Reasons: An Introduction to Inferentialism*. Cambridge, MA: Harvard UP.

——— 2002. *Tales of the Mighty Dead: Historical Essays in the Metaphysics of Intentionality*. Cambridge, MA: Harvard UP.

Brakel, Linda. 2009. *Philosophy, Psychoanalysis, and the A-rational mind*. New York: Oxford UP.

Brehmer, Berndt. 1980. "In one word: Not from experience," *Acta Psychologica* 45: 223–241.

Brentano, Franz. 1874 / 2005. "The Distinction between mental and physical phenomena," reprinted in *Metaphysics: Classic and Contemporary Readings*. 2nd Edition. Edited by Ronald Hoy and Nathan Oaklander. Translated by Dailey Terrell. Belmont, CA: Wadsworth Cengage Learning.

Bristow, William. 2017. "Enlightenment" in *The Stanford Encyclopedia of Philosophy*. (Fall 2017 Edition). Edited by Edward N. Zalta. https://plato.stanford.edu/archives/fall2017/entries/enlightenment/.

Brock, Timothy and Balloun, Joe. 1967. "Behavioral receptivity to dissonant information," *Journal of Personality and Social Psychology* 6(4.1): 413.

Brosnan, Sarah and de Waal, Frans. 2003. "Monkeys reject unequal pay," *Nature* 425: 297–299.

Buckle, Stephen. 2004. *Hume's Enlightenment Tract: The Unity and Purpose of An Enquiry Concerning Human Understanding*. New York: Oxford UP.

Buss, David. 2011. "Domains of deception," *Behavioral and Brain Sciences* 34 (1): 18.

Byrne, Ruth. 1989. "Suppressing valid inferences with conditionals," *Cognition* 31(1): 61–83.

Carone, Gabriela. 2001. "Akrasia in the Republic: Does Plato change his mind?," *Oxford Studies in Ancient Philosophy* 20: 107–148.

Carruthers, Peter. 1989. “Brute Experience,” *The Journal of Philosophy* 86(5): 505–516.

Carver, Charles, and Scheier, Michael. 2002. "Optimism," in *Handbook of Positive Psychology*. Edited by Charles Snyder and Shane Lopez. New York: Oxford UP, 231–43.

Casati, Roberto and Varzi, Achille. 2014. “Events,” in *The Stanford Encyclopedia of Philosophy*. (Winter 2015 Edition). Edited by Edward N. Zalta. https://plato.stanford.edu/archives/win2015/entries/events/.

Cassirer, Ernst. 1932 / 1979. *The Philosophy of the Enlightenment*. Translated by Fritz Koelln and James Pettegrove. Princeton: Princeton UP.

Cavell, Marcia. 1993. *The Psychoanalytic Mind: From Freud to Philosophy*. Cambridge, MA: Harvard UP.

——— 2006. *Becoming a Subject: Reflections in Philosophy and Psychoanalysis*. New York: Oxford UP.

Cavendish, Richard, “The birth of Robespierre,” *History Today* 58(5). https://www.historytoday.com/richard-cavendish/birth-robespierre.

Cheke, Lucy, Elsa Loissel and Nicola Clayton. 2012. “How do children solve Aesop’s fable?,” *PLoS ONE* 7(7): e40574.

Chalmers, David. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford UP.

Charlier, Philippe and Froesch, Philippe. 2013. “Robespierre: The oldest case of sarcoidosis?,” *The Lancet* 382(9910): 2068.

Chisholm, Roderick. 1964. “Human freedom and the self,” The Lindley Lecture, University of Kansas, 23 April.

——— 1970. “Events and propositions,” *Noûs* 4(1): 15–24.

——— 1990. “Events without times: An essay on ontology,” *Noûs* 24(3): 413–427.

Cloninger, Robert. 2004. *Feeling Good: The Science of Well-Being*. New York: Oxford UP.

Colvin, Randy, Jack Block, and David Funder. 1995. "Overly positive self-evaluations and personality: Negative implications for mental health," *Journal of Personality and Social Psychology* 68: 1152–1162.

Coope, Ursula. 2013. “Aristotle,” in *A Companion to the Philosophy of Action*. Edited by Timothy O’Connor and Constantine Sandis. Malden, MA: Blackwell, 439–446.

Cozolino, Louis. 2002. *The Neuroscience of Psychotherapy: Building and Rebuilding the Human Brain*. New York: W.W. Norton.

Cross, Patricia. 1977. “Not can but will college teaching be improved,” *New Directions for Higher Education* (17): 1–15.

Croyle, Robert, Elizabeth Loftus, Steven Barger, Yi-Chun Sun, Marybeth Hart, and JoAnn Gettig. 2006. “How well do people recall risk factor test results? Accuracy and bias among cholesterol screening participants,” *Healthy Psychology* 25(3): 425.

Csizszentmihalyi, Mihaly. 1991. *Flow: The Psychology of Optimal Experience*. New York: HarperCollins.

Damasio, Antonio. 1994. *Descartes’ Error: Emotion, Reason, and the Human Brain*. New York: Penguin.

——— 1999. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York: Harcourt.

Darley, John and Latané, Bibb. 1968. “Bystander intervention in emergencies: Diffusion of responsibility,” *Journal of Personality and Social Psychology* 8(4): 377–383.

Darwin, Charles. 2001. *Darwin: A Norton Critical Edition*. 3rd Edition. Edited by Philip Appleman. New York: W.W. Norton and Co.

Davidson, Donald. 1963. "Actions, reasons, and causes," reprinted in *Essays on Actions and Events*. Oxford: Oxford UP 2001, 3-20.

——— 1969a. “The individuation of events,” reprinted in *Essays on Actions and Events*. Oxford: Oxford UP 2001, 163–180.

——— 1969b. “True to the facts,” reprinted in *Inquiries into Truth and Interpretation*. Oxford: Oxford UP 2009, 37–54.

——— 1969c. “How is weakness of the will possible?,” reprinted in *Essays on Actions and Events*. Oxford: Oxford UP 2009, 21–42.

——— 1970. “Events as particulars,” reprinted in *Essays on Action and Events*. Oxford: Oxford UP 2001, 181–187.

——— 1975. “Thought and talk,” reprinted in *Inquiries into Truth and Interpretation*. Oxford: Oxford UP 2009, 155–170.

——— 1977. “Reality without reference,” reprinted in *Inquiries into Truth and Interpretation*. Oxford: Oxford UP 2009, 215–225.

——— 1978. “Intending,” reprinted in *Essays on Actions and Events*. Oxford: Oxford UP 2001, 83–102.

——— 1982. “Paradoxes of irrationality,” reprinted in *Problems of Rationality*. Oxford: Oxford UP 2004, 169–187.

——— 1985a. “Incoherence and irrationality,” reprinted in *Problems of Rationality*. Oxford: Oxford UP 2004, 189–198.

——— 1985b. “Reply to quine on events,” reprinted in *Essays on Actions and Events*. Oxford: Oxford UP 2001, 305–311.

——— 1986. “Deception and division,” reprinted in *Problems of Rationality*. Oxford: Oxford UP 2004, 199–212.

——— 1997. “Who is fooled?,” reprinted in *Problems of Rationality*. Oxford: Oxford UP 2004, 213–230.

Davis, Philip and Hersh, Reuben. 1986. *Descartes’ Dream: The World According to Mathematics*. Mineola, NY: Dover Publications.

Dawkins, Richard. 2006. *The Selfish Gene*. 30th Anniversary Edition. New York: Oxford UP.

——— 1999. *The Extended Phenotype: The Long Reach of the Gene*. Revised Edition. New York: Oxford UP.

Dawson, Erica, Kenneth Savitsky, and David Dunning. 2006. “‘Don’t tell me, I don’t want to know’: Understanding people’s reluctance to obtain medical diagnostic information,” *Journal of Applied Social Psychology* 36(3): 751–768.

Dennett, Daniel. 1991. *Consciousness Explained*. Boston, MA: Little, Brown and Co.

Dennis Jr., Edward. 1993. “Evaluation of the handling of the Branch Davidian stand-off in Waco, Texas February 28 to April 19, 1993,” The United States Department of Justice Archives. Washington, DC: Department of Justice. https://www.justice.gov/archives/publications/waco/evaluation-handling-branch-davidian-stand-waco-texas-february-28-april-19-1993.

Descartes, René. 1637. *Discourse on the Method*, reprinted in *Selected Philosophical Writings*. Translated by John Cottingham, Robert Stoothoff, and Dugald Murdoch. Cambridge: Cambridge UP 1999, 20–56.

——— 1641. *Meditations on First Philosophy*, reprinted in *Selected Philosophical Writings*. Translated by John Cottingham, Robert Stoothoff, and Dugald Murdoch. Cambridge: Cambridge UP 1999, 73–122.

Deweese-Boyd, Ian 2017. "Self-deception," in *The Stanford Encyclopedia of Philosophy*. (Winter 2017 Edition). Edited by Edward N. Zalta. https://plato.stanford.edu/archives/fall2017/entries/self-deception/.

Dial, Kenneth. 2003. "Wing-assisted incline running and the evolution of flight," *Science* 299(5605): 402–404.

Ditto, Peter, James Scepansky, Geoffrey Munro, Anne Marie Apanovitch, and Lisa Lockhart. 1998. "Motivated sensitivity to preference-inconsistent information," *Journal of Personality and Social Psychology* 75(1): 53–69.

Drewes, Charles. 1984. "Escape reflexes in earthworms and other annelids" in *Neural Mechanisms of Startle Behavior*. Edited by Robert Eaton. Boston, MA: Springer.

Dunbar, Robin. 1996. *Grooming, Gossip, and the Evolution of Language*. Cambridge, MA: Harvard UP.

Dunning, David, Judith Meyerowitz, and Amy Holzberg. 1989. "Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability," *Journal of Personality and Social Psychology* 57(6): 1082–1090.

Epley, Nicholas, and Whitchurch, Erin. 2008. "Mirror, mirror on the wall: Enhancement in self recognition," *Personality and Social Psychology Bulletin* 34: 1159–70.

Evans, Jonathan. 1972. "Interpretation and matching bias in a reasoning task," *Quarterly Journal of Experimental Psychology* 24(2): 193–199.

——— 1989. *Bias in Human Reasoning: Causes and Consequences*. Hillsdale, NJ: Lawrence Erlbaum.

——— 2005. "Deductive reasoning," in *The Cambridge Handbook of Thinking and Reasoning*. Edited by Keith Holyoak and Robert Morrison. Cambridge: Cambridge UP, 169–184.

Ewing, Alfred Cyril. 1967. *A Short Commentary on Kant's Critique of Pure Reason*. Chicago, IL: University of Chicago Press.

Fazio, Russell. 1995. "Attitudes as object-evaluation associations: Determinants, consequences, and correlates of attitude accessibility," in *Attitude Strength: Antecedents and Consequences*. Edited by Richard Petty and Jon Krosnick. Mahwah, NJ: Erlbaum, 247–282.

Fenichel, Otto. 1995. *The Psychoanalytic Theory of Neurosis*. New York: W.W. Norton.

Fingarette, Herbert. 1969. "Self-deception and the splitting of the ego," selection from *Self*-Deception, reprinted in *Philosophical Essays on Freud*. Edited by Richard Wollheim and James Hopkins. Cambridge: Cambridge UP 1982, 212–227.

Fischhoff, Baruch. 1975 / 2003. "Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty," *Quality and Safety in Health Care* 12: 304–312.

Fischhoff, Baruch and Beyth, Ruth. 1975. "'I knew it would happen': Remembered probabilities of once-future things," *Organizational Behavior and Human Performance* 13: 1–16.

Flanagan, Owen. 2000. *Dreaming Souls: Sleep, Dreams, and the Evolution of the Conscious Mind*. New York: Oxford UP.

Fodor, Jerry. 2003. *Hume Variations*. New York: Oxford UP.

Frederick, Shane, George Loewenstein, and Ted O'Donoghue. 2002. "Time discounting and time preference: A critical review," *Journal of Economic Literature* 40: 351–401.

Freud, Sigmund, and Breuer, Joseph. 2004. *Studies in Hysteria*. Translated by Nicola Luckhurst. New York: Penguin.

Freud, Sigmund. 1909a. "Some remarks on a case of obsessive-compulsive neurosis [The 'Ratman']," reprinted in *The 'Wolfman' and Other Cases*. Translated by Louise Adey Huish. New York: Penguin 2002, 123–202.

——— 1915a. "Drives and their fates," in *The Unconscious*. Translated by Graham Frankland. New York: Penguin 2005, 11–31.

——— 1915b. "Repression," in *The Unconscious*. Translated by Graham Frankland. New York: Penguin 2005, 33–45.

——— 1915c. "The unconscious," in *The Unconscious*. Translated by Graham Frankland. New York: Penguin 2005, 47–85.

——— 1917. "Mourning and melancholia," in *On Murder, Mourning, and Melancholia*. Translated by Shaun Whiteside. New York: Penguin 2005, 201–18.

Fridland, Ellen. 2011. "Reviewing the logic of self-deception," *Behavioral and Brain Sciences* 34(1): 22–23.

Furet, François. 1989. “Terror” in *A Critical Dictionary of the French Revolution*. Edited by François Furet and Mona Ozouf. Translated by Arthur Goldhammer. Cambridge, MA: Belknap of Harvard UP.

Gallistel, Charles, and Gelman, Rochel. 2005. “Mathematical cognition,” in *The Cambridge Handbook of Thinking and Reasoning*. Edited by Keith Holyoak and Robert Morrison. Cambridge: Cambridge UP, 559–588.

Gardner, Sebastian. 1993. *Irrationality and the Philosophy of Psychoanalysis*. Cambridge: Cambridge UP 2006.

——— 1999. *Kant and the Critique of Pure Reason*. New York: Routledge.

Gaukroger, Stephen. 1995. *Descartes: An Intellectual Biography*. New York: Oxford UP.

——— 2008. *The Emergence of a Scientific Culture: Science and the Shaping of Modernity 1210–1685*. New York: Oxford UP.

Gazzaniga, Michael and LeDoux, Joseph. 1978. *The Integrated Mind*. New York: Plenum.

Gillespie, Michael. 2008. *The Theological Origins of Modernity*. Chicago, IL: University of Chicago Press.

Glock, Hans-Johann. 2013. “Animal agency,” in *A Companion to the Philosophy of Action*. Edited by Timothy O’Connor and Constantine Sandis. Malden, MA: Blackwell, 384–392.

Goodale, Melvyn and Milner, David. 2004. *Sight Unseen: An Exploration of Conscious and Unconscious Vision*. New York: Oxford UP.

Greer, Donald. 1935. *The Incidence of the Terror During the French Revolution: A Statistical Intepretation*. Cambridge, MA: Harvard UP.

Gross, Daniel. 2014. “This is your brain on silence,” in *Nautilus*. Edited by Michael Segale. http://nautil.us/issue/16/nothingness/this-is-your-brain-on-silence.

Gusinde, Martin. 1937. *The Yahgan: The life and thought of the water nomads of Cape Horn*. Translated by Frieda Schütze. New Haven, CT: Human Relations Area Files.

HC Deb. 1934. “King’s Speech: Debate on the Address.” Speech, November 28, vol. 295, c862. https://api.parliament.uk/historic-hansard/commons/1934/nov/28/debate-on-the-address.

Haack, Susan. 1978. *Philosophy of Logics*. Cambridge: Cambridge UP.

Haidt, Jonathan. 2006. *The Happiness Hypothesis: Finding Modern Truth in Ancient Wisdom*. New York: Basic Books.

Halford, Graeme. 2005. "Development of thinking," in *The Cambridge Handbook of Thinking and Reasoning*. Edited by Keith Holyoak and Robert Morrison. Cambridge: Cambridge UP, 528–558.

Halford, Graeme, William Wilson, and Steven Phillips. 1998. "Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology," *Behavioral and Brain Sciences* 21(6): 803–831.

Harcourt, Edward. 2013. "Action explanation and the unconscious," in *A Companion to the Philosophy of Action*. Malden, MA: Blackwell, 166–173.

Harman, Gilbert. 1973. *Thought*. Princeton: Princeton UP.

Hart, William. 1982. "Models of repression," in *Philosophical Essays on Freud*. Edited by Richard Wollheim and James Hopkins. New York: Cambridge UP, 180–202.

Hatfield, Gary. 2001. "Perception as unconscious inference," *IRCS Technical Reports Series* 9: 1–48.

Haydon, Colin and Doyle, William. 1999. *Robespierre*. New York: Cambridge UP.

Hebb, Donald. 1946. "Emotion in man and animal: An Analysis of the intuitive processes of recognition," *Psychological Review* 53: 88–106.

Hegel, Georg W.F. 1837. *The Philosophy of History*. Translated by John Sibree. New York: Dover Publications (1956).

Higgins, Edward. 1996. "Knowledge activation: Accessibility, applicability, and salience," in *Social Psychology: Handbook of Basic Principles*. Edited by Edward Higgins and Arie Kruglanski. New York: Guilford Press, 133–168.

Hobbes, Thomas. 1651. *Leviathan*. Edited by Edwin Curley. Indianapolis, IN: Hackett Publishing 1994.

Hofstadter, Douglas. 2008. *I am a Strange Loop*. New York: Basic Books.

Holbach, Paul-Henri Thiry. 1770 / 2001. *The System of Nature or Laws of the Moral and Physical World*. Volume 1. Translated by Henry Robinson. Kitchener, Ontario: Batoche Books.

Homer. 1996. *The Odyssey*. Translated by Robert Fagles. New York: Viking Penguin.

Howard Hughes Medical Institute. 2018. “Mosquito brain atlas aims to reveal neural circuitry of behavior,” in *Science Daily* (March 7). www.sciencedaily.com/releases/2018/03/180307095258.htm.

Hume, David. 1748. *An Enquiry Concerning Human Understanding*. Edited by Tom Beauchamp. New York: Oxford UP 1999.

Ings, Thomas and Chittka, Lars. 2008. “Speed-accuracy tradeoffs and false alarms in bee responses to cryptic predators,” *Current Biology* 18: 1520–1524.

Irwin, Terence. 1995. *Plato’s Ethics*. New York: Oxford UP.

Jackson, Frank. 1982. “Epiphenomenal qualia,” *The Philosophical Quarterly* 32(127): 127–136.

Jelbert, Sarah, Alex Taylor, Lucy Cheke, Nicola Clayton, and Russell Gray. 2014. “Using the Aesop’s fable paradigm to investigate causal understanding of water displacement by new Caledonian crows,” *PLoS ONE* 9(3): e92895.

John, Oliver, and Robins, Richard. 1994. "Accuracy and bias in self-perception: Individual differences in self-enhancement and narcissism," *Journal of Personality and Social Psychology* 66: 206–19.

Kahneman, Daniel. 2003. "Objective happiness," in *Well-Being: Foundations of Hedonic Psychology*. Edited by Daniel Kahneman, Ed Diener, and Norbert Schwarz. New York: Russell Sage Foundation, 3–25.

Kahneman, Daniel and Lovallo, Dan. 1993. "Timid choices and bold forecasts: A cognitive perspective on risk taking," *Management Science* 39(1): 17–31.

Kahneman, Daniel and Tversky, Amos. 1984. "Choices, values, and frames," *American Psychologist* 39(4): 341–350.

Kant, Immanuel. 1781. *Critique of Pure Reason*. Translated by Norman Kemp Smith. New York: Palgrave Macmillan 2003.

——— 1784. “Answer to the question: What is Enlightenment?,” in *Basic Writings of Kant*. Edited by Allen Wood. Translated by Thomas Abbott. New York: Modern Library 2001.

Keinan, Giora. 1987. "Decision making under stress: Scanning of alternatives under controllable and uncontrollable threats," *Journal of Personality and Social Psychology* 52(3): 639–644.

Kim, Heejung and Markus, Hazel Rose. 1999. "Deviance or uniqueness, harmony or conformity? A cultural analysis," *Journal of Personality and Social Psychology* 77(4): 785–800.

Kim, Jaegwon. 1971. "Causation, nomic subsumption, and the concept of event," *The Journal of Philosophy* 70(8): 217–236.

——— 1976. "Events as property exemplifications," reprinted in *Metaphysics: Contemporary Readings*. 1st Edition. Edited by Steven D. Hales. London, UK: Wadsworth 1999, 336–348.

Klein, Jacob. 1989. *A Commentary on Plato's Meno*. Chicago, IL: The University of Chicago Press.

Koch, Christof. 2004. *The Quest for Consciousness: A Neurobiological Approach*. Englewood, CO: Roberts & Company Publishers.

Konorski, Jerzy. 1967. *Integrative Activity of the Brain*. Chicago: University of Chicago Press.

Kornhuber, Hans and Deecke, Lüder. 1976. "Hirnpotentialänderungen bei Willkürbewegungen und passive Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale," *Pflügers Archiv fir dir gesamte Physiologie des Menschen und Tieren* 284: 99–119.

Kraut, Richard. 2018. "Aristotle's Ethics," in *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition). Edited by Edward N. Zalta. https://plato.stanford.edu/archives/sum2018/entries/aristotle-ethics/.

Kreiman, Gabriel. 2001. *On the neuronal activity in the human brain during visual recognition, imagery and binocular rivalry*. Ph.D. Thesis. Pasadena: California Institute of Technology.

Kreiman, Gabriel, Itzhak Fried, and Christof Koch. 2002. "Single-neuron correlates of subjective vision in the human medial temporal lobe," *Proceedings of the National Academy of Sciences USA* 99: 8378–8383.

Kreiman, Gabriel, Christof Koch, and Itzhak Fried. 2000a. "Category-specific visual responses of single neurons in the human medial temporal lobe," *Nature Neuroscience* 3: 946–953.

——— 2000b. "Imagery neurons in the human brain," *Nature* 408: 357–361.

Kunda, Ziva. 1987. "Motivated inference: Self-serving generalization and evaluation of causal theories," *Journal of Personality and Social Psychology* 53: 636–647.

——— 1990. “The case for motivated reasoning,” *Psychological Bulletin* 108: 480–498.

Kurzban, Robert. 2010. *Why everyone (else) is a hypocrite*. Princeton: Princeton UP.

——— 2011. “Two problems with ‘self-deception’: No ‘self’ and no ‘deception’,” *Behavioral and Brain Sciences* 34(1): 32–33.

Kusch, Martin. 1995. *Psychologism: A Case Study in the Sociology of Philosophical Knowledge*. New York: Routledge.

——— 2015. “Psychologism,” in *The Stanford Encyclopedia of Philosophy* (Winter 2015 Edition). Edited by Edward N. Zalta. https://plato.stanford.edu/archives/win2015/entries/psychologism/.

LaPlanche, J. and Pontalis, J.-B. 1967. *The Language of Psychoanalysis*. Translated by Donald Nicholson-Smith. New York: W.W. Norton 1973.

Landsburg, Steven. 2008. *More Sex is Safer Sex: The Unconventional Wisdom of Economics*. New York: Free Press.

Larwood, Laurie. 1978. “Swine flu: A field study of self-serving biases,” *Journal of Applied Social Psychology* 18: 283–289.

Larwood, Laurie, and Whittaker, William. 1977. “Managerial myopia: Self-serving biases in organizational planning,” *Journal of Applied Psychology* 62: 194–198.

Latané, Bibb and Rodin, Judith. 1969. “A lady in distress: Inhibiting effects of friends and strangers on bystander intervention,” *Journal of Experimental Social Psychology* 5(2): 189–202.

LeBoeuf, Robyn and Shafir, Eldar. 2005. "Decision making," in *The Cambridge Handbook of Thinking and Reasoning*. Edited by Keith Holyoak and Robert Morrison. Cambridge: Cambridge UP, 243–265.

LeDoux, Joseph. 1996. *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York: Simon & Schuster Paperbacks.

——— 2002. *Synaptic Self: How Our Brains Become Who We Are*. New York: Penguin.

Lear, Jonathan. 1995. “The Heterogeneity of the mental,” in *Mind* 104(416): 863–879.

——— 1998. *Open Minded: Working Out the Logic of the Soul*. Cambridge, MA: Harvard UP.

——— 1999. *Love and Its Place in Nature: A Philosophical Interpretation of Freudian Psychoanalysis*. New Haven, CT: Yale UP.

——— 2000. *Happiness, Death, and the Remainder of Life*. Cambridge, MA: Harvard UP.

——— 2003. *Therapeutic Action: An Earnest Plea for Irony*. New York: Karnac Books.

——— 2005. *Freud*. New York: Routledge.

Leffler, Phyllis. 1976. "The 'Histoire Raisonnee,' 1660–1720: A pre-Enlightenment genre," *Journal of the History of Ideas* 37(2): 219–40.

Leibniz, G.W. 1704. *New Essays Concerning Human Understanding*. Translated by Peter Remnant and Jonathan Bennett. Cambridge: Cambridge UP 2000.

Lerman, Caryn, Robert Croyle, Kenneth Tercyak, and Heidi Hamann. 2002. "Genetic testing: Psychological aspects and implications," *Journal of Consulting and Clinical Psychology* 70: 784–97.

Lerner, Jennifer and Keltner, Dacher. 2001. "Fear, anger, and risk," *Journal of Personality and Social Psychology* 81(1): 146–159.

Libet, Benjamin. 1965. "Cortical activation in conscious and unconscious experience," *Perspectives in Biology and Medicine* 9: 77–86.

——— 1985. "Unconscious cerebral initiative and the role of conscious will in voluntary action," *The Behavioral and Brain Sciences* 8: 529–566.

Libet, Benjamin, Gleason, Curtis, and Wright, Elwood. 1983. "Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential." *Brain* 106: 623–642.

Linton, Marisa, "Robespierre and the terror," *History Today* 56(8). https://www.historytoday.com/marisa-linton/robespierre-and-terror.

Locke, John. 1689. *An Essay Concerning Human Understanding*. Edited by Peter Nidditch. Oxford: Oxford UP 1979.

Loewenstein, George, and Thaler, Richard. 1989. "Anomalies: Intertemporal choice," *Journal of Economic Perspectives* 3: 181–193.

Loftus, Elizabeth. 1986. "Ten years in the life of an expert witness," *Law and Human Behavior* 10(3): 241–263.

Loftus, Elizabeth and Hoffman, Hunter. 1989. "Misinformation and memory: The creation of new memories," *Journal of Experimental Psychology* 118(1): 100–104.

Logan, Corina, Brigit Harvey, Barney Schlinger, and Michelle Rensel. 2016. "Western scrub-jays do not appear to attend to functionality in Aesop's Fable experiments," *PeerJ* 4:e1707.

Logothetis, Nikos and Pauls, Jon. 1995. "Psychophysical and physiological evidence for viewer-centered object representations in the primate," *Cerebral Cortex* 5: 270–288.

Logothetis, Nikos, Jon Pauls, Heinrich Bülthoff, and Tomaso Poggio. 1994. "View-dependent object recognition by monkeys," *Current Biology* 4: 401–414.

Lord, Charles, Lee Ross and Mark Lepper. 1979. "Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence," *Journal of Personality and Social Psychology* 37(11): 2098–2109.

Lorenz, Hendrik. 2006. *The Brute Within: Appetitive Desire in Plato and Aristotle*. New York: Oxford UP.

Lorenz, Konrad. 1981. *The Foundations of Ethology*. Revised Edition. Translated by Konrad Lorenz and Robert Kickert. New York: Springer.

Lowe, Edward Jonathan. 2002. *A Survey of Metaphysics*. New York: Oxford UP.

McNamara, Robert. 1962. "No cities." Speech, July 9. Atomic Archive. http://www.atomicarchive.com/Docs/Deterrence/Nocities.shtml.

McPhee, Peter. 2012. *Robespierre: A Revolutionary Life*. New Haven, CT: Yale UP.

MacIntyre, Alasdair. 2004. *The Unconscious: A Conceptual Analysis*. New York: Routledge.

Maier, Norman. 1931. "Reasoning in humans. II. The solution of a problem and its appearance in consciousness," *Journal of Comparative Psychology* 12(2): 181–194.

Mantel, Hilary. 2000. "What a man this is, with his crowd of women around him!," *London Review of Books* 22(7): 3–8. https://www.lrb.co.uk/v22/n07/hilary-mantel/what-a-man-this-is-with-his-crowd-of-women-around-him.

Marshall, Lorna. 1961. "Sharing, talking, and giving: Relief of social tensions among Kung Bushmen," *Africa* 31: 231–249.

Mechner, Francis. 1958. "Probability relations within response sequences under ratio reinforcement," *Journal of the Experimental Analysis of Behavior* 1: 109–122.

Mele, Alfred. 1987. *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control*. New York: Oxford UP.

Mele, Alfred and Moser, Paul. 2003. "Intentional action," in *The Philosophy of Action*. Edited by Alfred Mele. New York: Oxford UP, 223–255.

Menn, Stephen. 2002. *Descartes and Augustine*. Cambridge: Cambridge UP.

Mercier, Hugo, Guy Politzer, and Dan Sperber. 2017. "What causes failure to apply the Pigeonhole Principle in simple reasoning problems?," *Thinking & Reasoning* 23(2): 184–189.

Mercier, Hugo and Sperber, Dan. 2017. *The Enigma of Reason*. Cambridge: Harvard UP.

Miller, Rachael, Sarah Jelbert, Alex Taylor, Lucy Cheke, Russell Gray, Elsa Loissel, and Nicola Clayton. 2016. "Performance in object-choice Aesop's Fable tasks are influenced by object biases in new Caledonian crows but not in human children," *PLoS ONE* 11(12): e0168056.

Milner, David, David Perrett, Rhona Johnston, Philip Benson, Timothy Jordan, David Heeley, Diego Bettucci, Franco Mortara, Roberto Mutani, Emanuela Terazzi, and Duncan Davidson. 1991. "Perception and action in 'visual form agnosia'," *Brain* 114(1): 405–428.

Minsky, Marvin. 2006. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. New York: Simon & Schuster.

Mobbs, Dean, Cindy Hagan, Tim Dalgleish, Brian Silston, and Charlotte Prévost. 2015. "The ecology of human fear: Survival optimization and the nervous system," Frontiers in Neuroscience 9(55): 1–22.

Molden, Daniel and Higgins, Edward. 2005. "Motivated thinking," in *The Cambridge Handbook of Thinking and Reasoning*. Edited by Keith Holyoak and Robert Morrison. Cambridge: Cambridge UP, 295–317.

Morrot, Gil, Brochet, Frédéric, and Dubourdieu, Denis. 2001. "The color of odors," *Brain and Language* 79(2): 309–320.

Myers, David and Bach, Paul. 1974. "Discussion effects on militarism-pacifism: A test of the group polarization hypothesis," *Journal of Personality and Social Psychology* 30(6): 741–747.

Myers, David and Bishop, George. 1970. "Discussion effects on racial attitudes," *Science* 169(3947): 778–779.

Nagel, Thomas. 1974. “What is it like to be a bat?,” *Philosophical Review* 83(4): 435–450.

Nisbett, Richard and Wilson, Timothy. 1977. “Telling more than we can know: Verbal reports on mental processes,” *Psychological Review* 84: 231–259.

Nordenfelt, Lennart. 2007. *Rationality and Compulsion: Applying Action Theory to Psychiatry*. New York: Oxford UP.

Nørretranders, Tor. 1998. *The User Illusion: Cutting Consciousness Down to Size*. Translated by Jonathan Sydenham. New York: Viking Penguin.

Nuss, Philippe. 2015. “Anxiety disorders and GABA neurotranmission: a disturbance of modulation,” *Neuropsychiatric Disease and Treatment* 11: 165–175.

Nussbaum, Martha. 2001. *Upheavals of Thought: The Intelligence of Emotions*. Cambridge, Cambridge UP.

O’Shaughnessy, Brian. 1982. “The id and the thinking process,” in *Philosophical Essays on Freud*. Edited by Richard Wollheim and James Hopkins. Cambridge: Cambridge UP, 106–123.

Olson, James, and Zanna, Mark. 1979. “A new look at selective exposure,” *Journal of Experimental Social Psychology* 15: 1–15.

Orians, Gordon. 2015. *Snakes, Sunrises, and Shakespeare: How Evolution Shapes Our Loves and Fears*. Chicago: University of Chicago Press.

Paulhus, Delroy. 1998. "Interpersonal and intrapsychic adaptiveness of trait self-enhancement: A mixed blessing?," *Journal of Personality and Social Psychology* 74: 1197–208.

Pears, David. 1982. "Motivated irrationality, Freudian theory and cognitive dissonance," in *Philosophical Essays on Freud*. Edited by Richard Wollheim and James Hopkins. Cambridge: Cambridge UP, 264–288.

——— 1986. *Motivated Irrationality*. New York: Oxford UP.

Penner, Terrence. 1990. "Plato and Davidson: Parts of the soul and weakness of will," *Canadian Journal of Philosophy* 20:sup1, 35–74.

Perry, Clint and Barron, Andrew. 2013. “Neural mechanisms of reward in insects,” *Annual Review of Entomology* 58: 543–562.

Pinker, Steven. 2002. *The Blank Slate: The Modern Denial of Human Nature*. New York: Penguin.

——— 2011. “Representations and decision rules in the theory of self-deception,” *Behavioral and Brain Sciences* 34(1): 35–36.

——— 2018. *Enlightenment Now: The Case of Reason, Science, Humanism, and Progress*. New York: Viking Penguin.

Plato. 1994. *Phaedo*. Translated by Eva Brann, Peter Kalkavage, and Eric Salem. Newburyport, MA: Focus Pub.

——— 1998. *Plato’s Symposium*. Translated by Avi Sharon. Newburyport, MA: Focus Pub.

——— 2001. *Plato’s Timaeus*. Translated by Peter Kalkavage. Newburyport, MA: Focus Pub.

——— 2003. *Plato’s Phaedrus*. Translated by Stephen Scully. Newburyport, MA: Focus Pub.

——— 2004a. *Plato’s Meno*. Translated by George Anastaplo and Laurence Berns. Newburyport, MA: Focus Pub.

——— 2004b. *Plato Republic*. Translated by CDC Reeve. Indianapolis, IN: Hackett Pub.

——— 2007. *Republic*. Translated by Joe Sachs. Newburyport, MA: Focus Pub.

——— 2011a. *The Protagoras*, in *Socrates and The Sophists*. Translated by Joe Sachs. Newburyport, MA: Focus Pub, 41–89.

——— 2011b. *The Cratylus*, in *Socrates and The Sophists*. Translated by Joe Sachs. Newburyport, MA: Focus Pub, 157–222.

Platt, John and Johnson, David. 1971. “Localization of position within a homogeneous behavioral chain: Effects of error contingencies,” *Learning and Motivation* 2(4): 386–414.

Premack, David and Woodruff, Guy. 1978. “Does the chimpanzee have a theory of mind?,” *The Behavioral and Brain Sciences* 4: 515–526.

Preston, Caroline and Harris, Stanley. 1965. "Psychology of drivers in traffic accidents," *Journal of Applied Psychology* 49(4): 284–288.

Pronin, Emily, Thomas Gilovich, and Lee Ross. 2004. “Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others,” *Psychological Review* 3: 781–799.

Putnam, Hilary. 1981. *Reason, Truth, and History*. Cambridge: Cambridge UP.

Pyszczynski, Tom and Greenberg, Jeff. 1987. "Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model," in *Advances in Experimental Social Psychology*. Volume 20. Edited by Leonard Berkowitz. New York: Academic Press, 297–340.

Quine, Willard Van Orman. 1950. "Identity, Ostension, and Hypostasis," in *From a Logical Point of View: Nine Logico-Philosophical Essays*. 2nd Edition. Cambridge, MA: Harvard UP 2003, 65–79.

Ramachandran, Vilaynur. 1992. "Blind spots," *Scientific American* 266: 86–91.

——— 2009. "Self-awareness: The last frontier," in *Edge*. Edited by John Brockman. URL = https://www.edge.org/conversation/self-awareness-the-last-frontier.

Ramachandran, Vilaynur and Blakeslee, Sandra. 1998. *Phantoms in the Brain: Probing the Mysteries of the Human Mind*. New York: William and Morrow Co.

Ramachandran, Vilaynur and Gregory, Richard. 1991. "Perceptual filling in of artificially induced scotomas in human vision," *Nature* 350: 699–702.

Reber, Paul and Kotovsky, Kenneth. 1997. "Implicit learning in problem solving: The role of working memory capacity," *Journal of Experimental Psychology* 126(2): 178–203.

Rees, John. 1998. *The Algebra of Revolution: The Dialectic and the Classical Marxist Tradition*. New York: Routledge.

Rescher, Nicholas. 2009. *Wishful Thinking and Other Philosophical Reflections*. Piscataway, NJ: Transaction Books.

Ricoeur, Paul. 1994. *Oneself as Another*. Trans. By Kathleen Blamey. Chicago, IL: University of Chicago Press.

Robespierre, Maximilien. 1794 / 1997. "On the principles of political morality." Speech, February. *Internet Modern History Sourcebook*. Translated by Paul Halsall. https://sourcebooks.fordham.edu/mod/1794robespierre.asp.

Robins, Richard and John, Oliver. 1997. "The quest for self-insight: Theory and research on the accuracy of self-perceptions," in Handbook of Personality Psychology. Edited by Robert Hogan, John Johnson, and Stephen Briggs. San Diego, CA: Academic Press.

Rorty, Amélie. 1980. "*Akrasia* and Pleasure: *Nicomachean Ethics* Book 7," in *Essays on Aristotle's Ethics*. Edited by Amélie Rorty. Los Angeles, CA: University of California Press, 267–284.

Russell, Daniel. 2007. *Plato on Pleasure and the Good Life*. New York: Oxford UP.

Ryle, Gilbert. 1949 / 2009. *The Concept of Mind*. 60th Anniversary Edition. New York: Routledge.

Sapolsky, Robert. 2017. *Behave: The Biology of Humans at Our Best and Worst*. New York: Penguin.

Sartre, Jean-Paul. 1943 / 1956. "Mauvaise foi and the unconscious," selection from *Being and Nothingness*, reprinted in *Philosophical Essays on Freud*. Edited by Richard Wollheim and James Hopkins. Translated by Hazel E. Barnes. Cambridge: Cambridge UP 1982, 203-11.

Scheff, Thomas and Fearon Jr., David. 2004. "Cognition and emotion? The dead end in self-esteem research," *Journal for the Theory of Social Behaviour* 34(1): 73–90.

Searle, John. 1994. "Animal minds," *Midwest Studies in Philosophy* 29: 206–219.

Sellars, Wilfrid. 2003. *Empiricism and the Philosophy of Mind*. Cambridge, MA: Harvard UP.

Shafir, Eldar. 1993. "Choosing versus rejecting: Why some options are both better and worse than others," *Memory and Cognition* 2(1): 546–556.

——— 2003. "Context, conflict, weights, and identities: Some psychological aspects of decision making," *Conference Series*; [*Proceedings*], 48, issue Jun. URL = https://EconPapers.repec.org/RePEc:fip:fedbcp:y:2003:i:jun:n:48:x:5.

Shafir, Eldar, Itamar Simonson, and Amos Tversky. 1993. "Reason-based choice," *Cognition* 49: 11–36.

Shafir, Eldar, Tom Waite, and Brian Smith. 2002. "Context-dependent violations of rational choice in honeybees (*Apis mellifera*) and gray jays (*Perisoreus canadensis*." *Behavioral Ecology and Sociobiology* 51: 180–187.

Simler, Kevin and Hanson, Robin. 2018. *The Elephant in the Brain: Hidden Motives in Everyday Life*. New York: Oxford UP.

Simon, Herbert. 1955. "A behavioral model of rational choice," *The Quarterly Journal of Economics* 69(1): 99–118.

Sirigu, Angela and Duhamel, Jean-René. 2016. "Reward and decision processes in the brains of humans and nonhuman primates," *Dialogues in Clinical Neuroscience* 18(1): 45–53.

Slotnick, Scott and White, Rachel. 2013. "The fusiform face area responds equivalently to faces and abstract shapes in the left and central visual fields," *NeuroImage* 83: 408–17.

Smith, David. 2011. "Aiming at self-deception: Deflationism, intentionalism, and biological purpose," *Behavioral and Brain Sciences* 34(1): 37–38.

Söderberg, Patrik and Fry, Douglas. 2017. "Anthropological aspects of ostracism," in *Ostracism, Exclusion, and Rejection*. Edited by Kipling Williams and Steve Nida. New York: Routledge, 258–272.

Spalding, Kirsty, Ratan Bhardwaj, Bruce Bucholz, Henrik David, and Jonas Frisén. 2005. "Retrospective birth dating of cells in humans," *Cell* 122(1): 133–143.

Spalding, Kirsty, Olaf Bergmann, Kanar Alkass, Samuel Bernard, Mehran Salehpour, Hagen Huttner, Emil Boström, Isabelle Westerlund, Celine Vial, Bruce Buchholz, Göran Possnert, Deborah C. Mash, Henrik Druid, and Jonas Frisén. 2013. "Dynamics of hippocampal neurogenesis in adult humans," *Cell* 153(6): 1219–1227.

Sperry, Roger. 1974. "Lateral specialization in the surgically separated hemispheres," in *Neuroscience 3rd Study Program*. Edited by Francis Schmitt and Frederic Worden. Cambridge, MA: MIT Press.

Spurr, John. 2002. "The English 'Post-Reformation'?," *The Journal of Modern History* 74(1): 101–119.

Stanton, Lauren, Emily Davis, Shylo Johnson, Amy Gilbert, and Sarah Benson-Amram. 2017. "Adaptation of the Aesop's Fable paradigm for use with raccoons (Procyon lotor): considerations for future application in non-avian and non-primate species," *Animal Cognition* 20(6): 1147–1152.

Surian, Luca, Caldi Stefana and Dan Sperber. 2007. "Attribution of beliefs by 13-month-old infants," *Psychological Science* 18(7): 580–586.

Svenson, Ola. 1981. "Are we all less risky and more skillful than our fellow driver?," *Acta Psychologica* 47: 143–148.

Swindal, James. 2012. *Action and Existence: A Case for Agent Causation*. New York: Palgrave Macmillan.

Szabados, Béla. 1973. "Wishful thinking and self-deception," *Analysis* 33(6): 201–205.

Taylor, Alex, Douglas Elliffe, Gavin Hunt, Nathan Emery, Nicola Clayton, and Russell Gray. 2011. “New Caledonian crows learn the functional properties of novel tool types,” *PloS ONE* 6(12): e26887.

Tennenbaum, Sergio. 2013. “Akrasia and irrationality,” in *A Companion to the Philosophy of Action*. Edited by Timothy O’Connor and Constantine Sandis. Malden, MA: Blackwell, 274–281.

Thaler, Richard. 1981. "Some empirical evidence on dynamic inconsistency," *Economics Letters* 8: 201–207.

——— 2015. *Misbehaving: The Making of Behavioral Economics*. New York: W.W. Norton.

Thibodeau, David and Whiteson, Leon. 2018. *Waco: A Survivor’s Story*. New York: Hachette Books.

Thompson, Christopher. 2012. “Augustine’s Confessions and the source of Christian character,” in *The Confessions: Saint Augustine of Hippo*. Ignatius Critical Edition. Edited by David Meconi. San Francisco: Ignatius Press, 505–522.

Tinbergen, Nikolaas. 1961. *The Herring Gull’s World: A Study of the Social Behaviour of Birds*. Revised Edition. New York: Basic Books.

Tomaka, Joe, Jim Blascovich, and Robert Kelsey. 1992. "Effects of self-deception, social desirability, and repressive coping on psychophysiological reactivity to stress," *Personality and Social Psychology Bulletin* 18: 616–24.

Tuominen, Miira. 2009. *The Ancient Commentators on Plato and Aristotle*. Berkeley, CA: University of California Press.

Turner, John. 1985. "Social categorization and the self-concept: A social cognitive theory of group behavior," in *Advances in Group Processes*. Vol. 2. Edited by Edward Lawler. Greenwich, CT: JAI Press, 77–121.

Tversky, Amos, Paul Slovic, and Daniel Kahneman. 1990. “The causes of preference reversal,” *The American Economic Review* 80(1): 204–217.

van Boven, Leaf and Loewenstein, George. 2003. "Social projection of transient drive states," *Personality and Social Psychology Bulletin* 29(9): 1159–1168.

van Inwagen, Peter and Sullivan, Meghan. 2014. “Metaphysics,” in *The Stanford Encyclopedia of Philosophy* (Fall 2017 Edition). Edited by Edward N. Zalta. https://plato.stanford.edu/archives/fall2017/entries/metaphysics/.

van der Velde, Frank, Joop van der Pligt, and Christa Hooykaas. 1994. "Perceiving AIDS-related risk: Accuracy as a function of differences in actual risk," *Health Psychology* 13(1): 25.

Vinauger, Clément, Chloé Lahondère, Gabriella Wolff, Lauren Locke, Jessica Liaw, Jay Parrish, Omar Akbari, Michael Dickinson, and Jeffrey Riffell. 2018. "Modulation of host learning in *Aedes aegypti* mosquitoes," *Current Biology* 28(3): 333–344.

von Hippel, William and Trivers, Robert. 2011. "The evolution and psychology of self-deception," *Behavioral and Brain Sciences* 34 (1): 1–16.

Wallace, William. 1972. *Causality and Scientific Explanation*. Vol. 1. Ann Arbor, MI: University of Michigan Press.

Wason, Peter. 1968. "Reasoning about a rule," *Quarterly Journal of Experimental Psychology* 20(3): 273–281.

Wason, Peter and Evans, Jonathan. 1975. "Dual process in reasoning?," *Cognition* 3: 141–154.

Wehner, Rüdiger. 2003. "Desert ant navigation: How miniature brains solve complex tasks," *Journal of Comparative Physiology A* 189(8): 579–588.

Weinstein, Neil. 1980. "Unrealistic optimism about future life events," *Journal of Personality and Social Psychology* 39: 806–820.

——— 1982. "Unrealistic optimism about susceptibility to health problems," *Journal of Behavioral Medicine* 5: 441–460.

Wiggins, David. 1980a. "Deliberation and practical reason," in *Essays on Aristotle's Ethics*. Edited by Amélie Rorty. Los Angeles, CA: University of California Press, 221–240.

——— 1980b. "Weakness of will, commensurability, and the objects of deliberation and desire," in *Essays on Aristotle's Ethics*. Edited by Amélie Rorty. Los Angeles, CA: University of California Press, 241–265.

Wilson, Catherine. 2010. *Epicureanism at the Origins of Modernity*. New York: Oxford UP.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Edited by G.E.M. Anscombe. Oxford: Blackwell.

——— 1972. *On Certainty*. New York: Harper and Row.

Wollheim, Richard. 1982. “The bodily ego,” in *Philosophical Essays on Freud*. Edited by Richard Wollheim and James Hopkins. Cambridge: Cambridge UP, 124–138.

Wood, Allen. 2005. *Kant*. Malden, MA: Blackwell.

Woodward, Llewellyn. 1938. *The Age of Reform, 1850–1870*. London: Oxford UP.

Yamagishi, Toshio, Hirofumi Hashimoto, and Joanna Schug. 2008. “Preferences versus strategies as explanations for culture-specific behavior,” *Psychological Science* 19(6): 579–584.