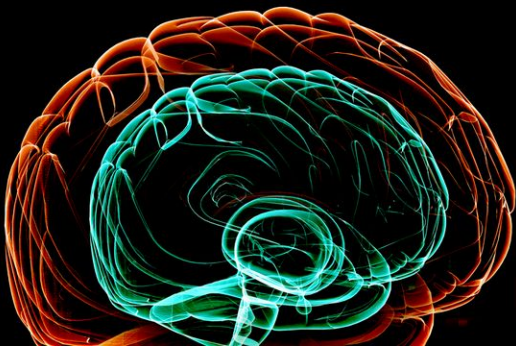
THE MIND WITHIN THE BRAIN

HOW WE MAKE DECISIONS AND HOW THOSE DECISIONS GO WRONG

A. DAVID REDISH



The Mind within the Brain

This page intentionally left blank

The Mind within the Brain

How We Make Decisions and How Those Decisions Go Wrong

A. DAVID REDISH

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trademark of Oxford University Press
in the UK and certain other countries.

Published in the United States of America by
Oxford University Press
198 Madison Avenue, New York, NY 10016

Library of Congress Cataloging-in-Publication Data
Redish, A. David.
The mind within the brain : how we make decisions and how those decisions go wrong / A. David Redish.
pages cm
Includes bibliographical references.
Summary: In *The Mind within the Brain,* A. David Redish brings together cutting-edge research in psychology, robotics, economics, neuroscience, and the new fields of neuroeconomics and computational psychiatry, to offer a unified theory of human decision-making. Most importantly, Redish shows how vulnerabilities, or "failure modes," in the decision-making system can lead to serious dysfunctions, such as irrational behavior, addictions, problem gambling, and post-traumatic stress disorder. Ranging widely from the surprising roles of emotion, habit, and narrative in decision-making to the larger philosophical questions of how mind and brain are related, what makes us human, the nature of morality, free will, and the conundrum of robotics and consciousness, *The Mind within the Brain* offers fresh insight into some of the most complex aspects of human behavior.—Provided by publisher.
ISBN 978–0–19–989188–7 (hardback)
1. Decision making. I. Title.
BF448.R43 2013
153.8′3—dc23* 2012046214

1 3 5 7 9 8 6 4 2
Printed in the United States of America
on acid-free paper

For Laura, Jay, Danny, and Sylvia,
the best decisions I ever made.

This page intentionally left blank

CONTENTS

PREFACE

If I could, I would reach beyond the cage of bone,
to touch the mind within the brain,
to reach the frightened nerves that wrap the heart;
I would speak your name there.

Our decisions make us who we are. Although we would like to think that our decisions are made rationally, deliberatively, many decisions are not. We all know that some of our decisions are made emotionally, and some are made reactively. Some have their intended consequences, and some have consequences we never imagined possible.

We are physical beings. The human brain is a complex network of neurons and other cells that takes information in from the world through its sensory systems and acts on the world through its motor systems. But how does that network of cells, in constant dynamic flux, become the person you are? How does the mind fit into that small place in the cage of bone that is our skull? How does it process information? How does it perceive the world, determine the best option, select an action, and take that action? How does it fall in love? Laugh at the overwhelming emotion of holding an infant? How does it create great art or great music? How does it feel the triumphant emotion of Beethoven's *Ode to Joy* or the devastating pathos of Bob Dylan's *Knock Knock Knocking on Heaven's Door*? Just how does the lady sing the blues? How does it get addicted and how does it break that addiction? How does it have a personality? What makes you you and me me?

Fundamentally, all of these questions are about how the being that you recognize as yourself fits into this physical brain nestled in your skull. Fundamentally, these questions are about how that brain makes decisions. This book is an attempt to answer that question.

Where this book came from

A few years ago, John Gessner, who runs a local program for people with gambling problems and their families, asked me if I would be willing to give a talk to his clients on

decision-making. I had been giving talks to Jan Dubinsky's *BrainU* program for high school teachers interested in neuroscience and had recently given a talk to frontline medical professionals (doctors, nurses, and addiction social workers) on my laboratory's work identifying vulnerabilities in decision-making systems. John had heard of this talk and wanted me to present this work to his clients.

I protested that I was not a medical professional and that I could not tell them how to fix what was broken. He said that they had lots of people to tell them that, what they wanted was someone to tell them *why—Why do we make the choices we do?* He said they wanted to know why those decisions get made wrong, especially when they knew what the right choices were.

The lecture itself went very well. There were several dozen people in the audience, and they were involved and asking questions throughout. And then, afterwards, they had so many questions that I stayed there answering questions for hours. There was a hunger there that I had not appreciated until I met those gamblers and their families, an almost desperate desire to understand how the brain works. They had seen how things can go wrong and needed an explanation, particularly one that could explain how they could both be conscious beings making decisions and yet still feel trapped. Somehow, that science lecture on how the multiple decision-making systems interact reached them. I realized then that there was a book I had to write.

Over the past three years, this book has morphed and grown. My goal, however, remains to explain the science of how we make decisions. As such, an important part of this book will be to identify what questions remain.[1] My goal is not to provide a self-help book to help you make better decisions. I am not going to tell you what you should do. Nor are the answers to curing addiction or poor decisions herein. You should check with your own medical professionals for treatment. Every individual is unique, and your needs should be addressed by someone directly familiar with them. Nevertheless, I hope that you find the book illuminating. I hope you enjoy reading it. It has been tremendously fun to write.

The structure of the book

One of the remarkable things that has occurred over the past several decades is the convergence of different fields on the mechanisms of decision-making. Scientific fields as diverse as psychology, robotics, economics, neuroscience, and the new fields of neuroeconomics and computational psychiatry have all been converging on the recognition that decision-making arises from a complex interaction of multiple subsystems. In fact, these fields have converged on a similar categorization of the differences between the subsystems. In this book, we will explore how this convergence explains the decision-making that we (as humans) do.

I have divided this book into four sections. The first sections (*Decisions and the Brain* and *The Decision-Making System*) will lay out the work that has been done on the basic mechanisms—*What is a decision? How does the brain's decision-making system work? What are the components that make up that decision-making system?* And then, the third and fourth sections will explore the consequences of that system.

The first section consists of five chapters, two chapters to set the stage (1: *What Is a Decision?* and 2: *The Tale of the Thermostat*) and three chapters to introduce the basic neuroeconomics of decision-making (3: *The Definition of Value*, 4: *Value, Euphoria, and the Do-It-Again Signal*, and 5: *Risk and Reward*). In the second section, we will start with the results that the decision-making system is made up of multiple modules or subsystems (Chapter 6), and then spend a chapter each on the component systems (Chapters 7 through 15).

In the third section (*The Brain With a Mind of Its Own*), we will explore the consequences of the physical nature of the brain, how mind and brain are related, and how vulnerabilities in the decision-making system can lead to dysfunction, such as addiction (Chapter 18), problem gambling (Chapter 19), and post-traumatic stress disorder (Chapter 20).

Finally, in the fourth section (*The Human Condition*), we will explore the philosophical questions of what makes us human (Chapter 22), of morality (Chapter 23), and of free will and consciousness (24: *The Conundrum of Robotics*) in the light of the new work on decision-making systems discussed in the previous sections.

I've tried to write the book so that it can be read straight through from start to finish by a reader with only a basic knowledge of neuroscience and computers; however, some readers may feel that they want a more detailed introduction to the concepts that are being discussed in this book. For those readers, I've included three chapters in an appendix, including *What is information processing* and *How neurons process information* (Appendix A), *How we can read that information from neural signals* (Appendix B), and *How memories are stored* (by content, not by index, Appendix C).

Throughout the book, every statement is backed up with citations. These citations will be marked with superscript numbers, matching the list in the bibliographic notes, which will then reference the actual list of citations.[2] These numbers are not endnotes and will not be used to refer to any additional text; they are there only to back up the claims in the book. Instead, extra information and discussion that could distract from the flow will be put into footnotes, marked with superscript letters.[A]

Each chapter begins with a short poem and ends with a set of follow-up readings. In my mind, I think of the poems as contemplative moments that can be used to shape one's perspective when reading the chapter. As a friend recovering from cancer in his hospital bed recently told me, "Sometimes you need the poetry to get to the heart of the science." The follow-up readings at the end of each chapter are books, review articles, starting points for those who want to pursue a topic in depth. While the superscript citations will refer to the primary literature, some of which can be quite difficult to understand, the follow-up readings should be understandable by anyone reading this book.

[A] I prefer footnotes to endnotes because footnotes allow you to glance at the text without having to lose your place in the book. I will generally be including three kinds of information in footnotes: (1) parenthetical comments that are too long to be included in parentheses (such as the etymology and location of brain structures); (2) cheeky jokes and stories that I can't resist including but would disrupt the flow of the book if I included them in the main text; and (3) technical details when there are subtle, second-order effects that need to be noted but are too complicated for those without the necessary background. The book should be readable without the footnotes, but I hope the footnotes will add an extra dimension for those who want more depth.

This page intentionally left blank

ACKNOWLEDGMENTS

Any project of this magnitude depends on many people. The results presented in this book include work by a host of scientists working in a number of fields over the past hundred years. Throughout, I have tried my best to acknowledge and cite the papers in which these ideas first appeared. I am sure that I have missed some, and apologize here to my colleagues for any omissions.

This work would never have been possible without the amazing and wonderful students and postdocs who have come through my lab, particularly Adam Johnson, Steve Jensen, Matthijs van der Meer, and Zeb Kurth-Nelson, with whom I have worked directly on these ideas, as well as Neil Schmitzer-Torbert, Jadin Jackson, John Ferguson, Anoopum Gupta, Beth Masimore, Andy Papale, Nate Powell, Paul Regier, Adam Steiner, Jeff Stott, Brandy Schmidt, and Andrew Wikenheiser. I am also indebted to Kelsey Seeland and Chris Boldt, without whom my laboratory would not function as it does and I would not have had the freedom to write this book.

Many of the ideas presented here have been worked out in conversations with colleagues. And so I need to thank those colleagues for the invigorating discussions we have had at conferences and workshops over the years, including Peter Dayan, Read Montague, Larry Amsel, Nathaniel Daw, Yael Niv, Warren Bickel, Jon Grant, Bernard Balleine, Dave Stephens, John O'Doherty, Antonio Rangel, Cliff Kentros, Andre Fenton, and many others too numerous to mention. I want to take the time here to thank my friends and colleagues who read drafts of part or all of this book, particularly Jan Dubinsky, Matthijs van der Meer, and John Gessner, as well as my editor Joan Bossert.

Much of the work presented in this book has been funded by the National Institutes of Health, particularly by grants for the study of decision-making from NIMH and NIDA. My first foray into decision-making was funded by a Fellowship from the Sloan Foundation, and from a McKnight Land-Grant Fellowship from the University of Minnesota. Students were funded by training grants from the National Science Foundation and the National Institutes for Health. This work would have been impossible without their generous support.

Finally, this book is dedicated to my wife Laura, my first reader, who read every chapter, before the first draft, after it, and then again, when the book was done, who has put up with the time and stress this book put on me, who inspires my poetry, and who doesn't let me get away with anything less than the best I can do. This book, all the work I have done, is for her.

This page intentionally left blank

PART ONE

DECISIONS AND THE BRAIN

This page intentionally left blank

1

What Is a Decision?

snow flurries fall from the roof
a squirrel skids to a stop
Leap! into the unknown

In order to be able to scientifically measure decision-making, we define decisions as "taking an action." There are multiple decision-making systems within each of us. The actions we take are a consequence of the interactions of those systems. Our irrationality occurs when those multiple systems disagree with each other.

Because we like to think of ourselves as rational creatures, we like to define decision as the conscious deliberation over multiple choices. But this presumes a mechanism that might or might not be correct. If we are such rational creatures, why do we make such irrational decisions? Whether it be eating that last French fry that's just one more than we wanted or saying something we shouldn't have at the faculty meeting or the things we did that time we were drunk in college, we have all said and done things that we regret. Many a novel is based on a character having to overcome an irrational fear. Many an alcoholic has sworn not to drink, only to be found a few days later, in the bar, drink in hand.

We will see later in this book that the decisions that we make arise from an interaction of multiple decision-making systems. We love because we have emotional reactions borne of intrinsic social needs and evolutionary drives (the *Pavlovian* system, Chapter 8). We decide what job to take through consideration of the pros and cons of the imagined possibilities (*episodic future thinking* and the *Deliberative* system, Chapter 9). We can ride a bike because we have trained up our *Procedural* learning system (Chapter 10), but it can also be hard to break bad habits that we've learned too well (Chapter 15). We will see that these are only a few of the separable systems that we can identify. All of these different decision-making systems make up the person you are.

The idea that our actions arise from multiple, separably identifiable components has a long history in psychology, going back to Freud, or earlier, and has gained recent traction with theories of distributed computing, evolutionary psychology, and behavioral economics.[1] Obviously, in the end, there is a single being that takes an action, but

sometimes it's helpful to understand that being in terms of its subsystems. The analogy that I like, which I will use several times in this book, is that of a car. The car has a drive train, a steering system, brakes. Cars often have multiple systems to accomplish the same goal (the foot-pedal brake and the emergency/parking brake, or the electric and gasoline engines in a hybrid like the Toyota Prius).

The psychologist Jonathan Haidt likes to talk of a rider and an elephant as a metaphor for the conscious and unconscious minds both making decisions,[2] but I find this analogy unsuitable because it separates the "self" from the "other" decision-making systems. As is recognized by Haidt at the end of his book, you are both the rider *and* the elephant. When a football star talks about being "in the zone," he's not talking about being out of his body and letting some other being take over—he feels that he is making the right decisions. (In fact, he's noticing that his procedural/habit system is working perfectly and is making the right decisions quickly and easily.) When you are walking through a dark forest and you start to get afraid and you jump at the slightest sound, that's not some animal reacting—that's you. (It's a classic example of the Pavlovian action-selection system.) Sometimes these systems work together. Sometimes they work at cross purposes. In either case, they are all still *you*. "Do I contradict myself?" asks Walt Whitman in his long poem *Song of Myself*. "Then I contradict myself. I am large. I contain multitudes."

So how do we determine how these multiple systems work? How do we determine when they are working together and when they are at cross-purposes? How do we identify the mechanisms of Whitman's multitudes?

To study something scientifically, we need to define our question in a way that we can measure and quantify it. Thus, we need a measure of decision-making that we can observe, so we can compare the predictions that arise from our hypotheses with actual data. This is the key to the scientific process: there must be a comparison to reality. If the hypothesis doesn't fit that reality, we must reject the hypothesis, no matter how much we like it.

One option is to simply to ask people what they want. But, of course, the idea that we always do what we say we want makes very strong assumptions about how we make decisions, and anyone who has regretted a decision knows that we don't always decide to do what we want. Some readers may take issue with this statement, saying that you wanted that decision when you took the action. Just because you regret that night of binge drinking when you have a hangover in the morning doesn't mean you didn't want all those drinks the night before. Other readers may argue that a part of you wanted that decision, even if another part didn't. We will come to a conclusion very similar to this, that there are multiple decision-making modules, and that the members of Whitman's multitudes do not always agree with each other. Much of this book will be about determining who those multitudes are and what happens when they disagree with each other.

As a first step in this identification of the multitudes, careful scientific studies have revealed that a lot of conscious "decisions" that we think we make are actually rationalizations after the fact.[3] For example, the time at which we think we decided to start an action is often after the action has already begun.

In what are now considered classic studies of consciousness, Benjamin Libet asked people to perform an action (such as tapping a finger whenever they wanted to) while

watching a dot move around a circle.[4] The people were asked to report the position of the dot when they decided to act. Meanwhile, Libet and his colleagues recorded electrical signals from the brain and the muscles. Libet found that these signals preceded the conscious decision to act by several hundred milliseconds. Both the brain and muscles work by manipulating electricity, which we can measure with appropriate technologies. With the appropriate mathematics, we can decode those signals and determine what is happening within the brain. Libet decoded when the action could be determined from signals in the motor areas of the brain and compared it to when consciousness thought the action had occurred. Libet found that the conscious decision to take an action was delayed, almost as if consciousness was perceiving the action, rather than instigating it. Several researchers have suggested that much of consciousness is a monitoring process, allowing it to keep track of things and step in if there are problems.[5]

Much of the brain's machinery is doing this sort of filling-in, of making good guesses about the world. Our eyes can focus on only a small area of our visual field at a time. Our best visual sensors are an area of our retina that has a concentration of cells tuned to color and specificity in a location called the *fovea*. This concentration of detectors at the fovea means that we're much better at seeing something if we focus our eyes on it. Our vision focuses on a new area of the visual world every third of a second or so. The journeys between visual focusings are called *saccades*. Even while we think we are focusing our attention on a small location, our eyes are making very small shifts called *microsaccades*. If the visual world is aligned to our microsaccades so that it shifts when we do,[A] the cells adapt to the constant picture and the visual image "grays out" and vanishes. This means that most of our understanding of the visual world is made from combining and interpreting short-term visual memories. We are inferring the shape of the world, not observing it.

In a simple case familiar to many people, there's a small location on our retina where the axons from the output cells have to pass through to form the optic nerve sending the visual signals to the rest of our brain. This leaves us with a "blind spot" that must be filled in by the retina and visual processing system.[B] Our brain normally fills in the "blind spot" from memories and expectations from the surrounding patterns.[7]

In a similar way, we don't always notice what actually drives our decision-making process. We rationalize it, filling in our reasons from logical perspectives. Some of my favorite examples of this come from the work of V. S. Ramachandran,[8] who has studied patients with damage to parts of the brain that represent the body. A patient who is

---

[A] The visual world can be aligned to our microsaccades by tracking the movement of the eyes and shifting a video display very quickly. This is feasible with modern computer technology.

[B] The wires (axons) of our optic nerve have to pass through the retina because the retina is built backwards, with the processing cells on top and the light-detecting cells on the bottom. The processing cells are generally transparent, so light passes through them and the light-detecting cells can still see, but at some point the axons have to leave the processing cells to travel to the brain. There are no light-detecting cells at the point where the axons pass through, leaving a blind spot. This is a relic of the evolutionary past of the human species. Eyes do not have to be built this way—the octopus eye, with a different evolutionary history, is oriented correctly, with the light-detecting cells on top and the processing cells below. Octopi therefore do not have a blind spot that needs to be filled in.[6]

physically unable to lift her arm denies that she has a problem and merely states that she does not want to. When her arm was made to rise by stimulating her muscles directly, she claimed that she had changed her mind and raised her arm because she wanted to, even though she had no direct control of the arm. In a wonderful story (told by Oliver Sacks in *A Leg to Stand On*), a patient claims that his right hand is not his own. When confronted with the fact that there are four hands on the table (two of his and the two of the doctor's), the patient says that three of the hands belong to the doctor. "How can I have three hands?" asks the doctor. "Why not? You have three arms." replies the patient.

In his book *Surely You're Joking, Mr. Feynman*, the famous physicist Richard Feynman described being hypnotized. He wrote about how he was sure he could take the action (in this case opening his eyes) even though he had been hypnotized not to, but he decided not to in order to see what would happen. So he didn't, which was what he had been asked to do under hypnosis. Notice that he has rationalized his decision. As Feynman himself recognized, even if he said "I could have opened my eyes," he didn't. So what made the decision? Was it some effect of the hypnosis on his brain or was it that he didn't want to? How could we tell? Can we tell?

Many of the experiments we're going to talk about in this book are drawn from animals making decisions. If we're going to say that animals make decisions, we need to have a way of operationalizing that decision—it's hard to ask animals what they think. There are methods that allow us to decode information represented within specific parts of the brain, which could be interpreted as a means of asking an animal what it thinks (see Appendix B). However, unlike asking humans linguistically, where one is asking the overall being what it thinks, decoding is asking a specific brain structure what it is representing. Of course, one could argue that assuming what people say is what they think assumes that humans are unified beings. As we will see as we delve deeper into how decisions are made, humans (like other mammals) are mixtures of many decision-making systems, not all of which always agree with each other.

Just as the Libet experiments suggest that parts of the brain can act without consciousness, there are representations in the brain that are unable to affect behavior. In a remarkable experiment, Pearl Chiu, Terry Lohrenz, and Read Montague found signals in both smokers' and nonsmokers' brains that represented not only the success of decisions made but also what they could have done if they had made a better choice.[9] This recognition of what they could have done is called a *counterfactual* (an imagination of what might have been) and enables enhanced learning. (It is now known that both rats and monkeys can represent counterfactual reward information as well. These signals appear to use the same brain structures that Chiu, Lohrenz, and Montague were studying, and to be the same structures involved when humans express regret.[10]) Counterfactuals enhance learning by allowing one to learn from imagined possibilities. For example, by watching someone else make a mistake. Or (in the example used by Chiu, Lohrenz, and Montague) "if I had taken my money out of the stock market last week, I wouldn't have lost all that money when it crashed." While signals in both groups' brains reflected this counterfactual information, only nonsmokers' behavior took that information into account. If we want to understand how decisions are made and how they go wrong, we are going to need a way to

determine not only the actions taken by the subject but also the information processing happening in his or her brain.

Certainly, most nonhuman animals don't have language, although there may be some exceptions.[C] Nevertheless, it would be hard to ask rats, pigeons, or monkeys (all of which we will see making decisions later in the book) what they want linguistically. Given that it is also hard to ask humans what they really want, we will avoid this language problem altogether and operationalize making a decision as *taking an action*, because taking an action is an observable response.

This is very similar to what is known in behavioral economics as *revealed preferences*.[13] Economic theory (and the concomitant new field of neuroeconomics) generally assumes that those revealed preferences maintain a rational ordering such that if you prefer one thing (say chocolate ice cream) to another (say vanilla ice cream), then you will always prefer chocolate ice cream to vanilla if you are ever in the same situation. We will not make that assumption.

Similarly, we do not want to assume that a person telling you what he or she wants actually reflects the choices a person will make when faced with the actual decision.[14] Given the data that our conscious observations of the world are inferred and the data that our spoken explanations are rationalizations,[15] some researchers have suggested that our linguistic explanations of our desires are better thought of as the speech of a "press secretary" than the actions of an executive.[16] Thus, rather than asking what someone wants, we should measure decisions by giving people explicit choices and asking them to actually choose. We will encounter some of the strangenesses discovered by these experiments in subsequent chapters.

Often, experimentalists will offer people hypothetical choices. It is particularly difficult to get funding to provide people a real choice between $10,000 and $100,000, or to give them a real choice whether or not to kill one person to save five. Instead, subjects are asked to imagine a situation and pretend it was real. In practice, in the few cases where they have been directly compared, hypothetical and real decision-making choices tend to match closely.[17] But there are some examples where hypothetical and real decisions diverge.[18] These tend to be with rewards or punishments that are sensory, immediate, and what we will recognize later as *Pavlovian* (Chapter 8).

A classic example of this I call *the parable of the jellybeans*. Sarah Boysen and Gary Berntson tried to train chimpanzees to choose the humble portion.[19] They offered the subject two trays of jellybeans. If he reached for the larger tray, he got the smaller one and the larger one was given to another chimpanzee; however, if the deciding chimpanzee reached for the smaller tray, he got the larger tray and the smaller one was given to another chimpanzee. When presented with symbols (Arabic numerals that they had previously been trained to associate with numbers of jellybeans), subjects were able

[C] The extent to which animals can acquire human-level languages is not without its controversies. Since we're not going to trust human language either,[11] this issue is actually irrelevant to our question of decision-making. For those interested in the question of animals learning language, I recommend one of the excellent books written by the various researchers who have tried to teach language to nonhuman animals, such as Daniel Fouts teaching Washoe the chimpanzee, Penny Patterson teaching Koko the gorilla, Sue Savage-Rumbaugh teaching Kanzi the bonobo, or Irene Pepperberg teaching Alex the parrot.[12]

to choose the smaller group, but when the choices were physical jellybeans, they were unable to prevent themselves from reaching for the larger group of jellybeans. Other experiments have found that more linguistically capable animals are more able to perform these self-control behaviors.[20] This may be akin to our ability to talk ourselves out of doing things that we feel we really want: "I know I'm craving that cigarette. But I don't want it. I really don't."

A similar experiment is known colloquially as *the marshmallow experiment*.[21] Put a single marshmallow in front of a child sitting down at the kitchen table. Tell the child that if the marshmallow is still sitting there in five minutes, you'll add a second marshmallow to it. Then leave the room. It is very, very difficult for children not to reach for the marshmallow. It is much easier for children to wait for two pennies or two tokens than for two marshmallows. We will discuss the marshmallow experiment in detail in the chapter on self-control (Chapter 15).

Studies of decision-making in psychology (such as the marshmallow experiment) as well as studies in behavioral economics and the new field of neuroeconomics tend to measure choices within the limitation of discrete options. In the real world, we are rarely faced with a discrete set of options. Whether it be deciding when to swing the bat to hit a baseball or deciding where to run to on a playground, we are always continuously interacting with the world.[22] As we will see later, some of the mechanisms that select the actions we take are not always deliberative, and do not always entail discrete choices.

So where does that leave us in our search for a definition of decision-making? We will not assume that all decision-making is rational. We will not assume that all decision-making is deliberative. We will not assume that decision-making requires language. Instead, we define decision-making as *the selection of an action*. Obviously, many of the decisions we take (such as pushing a lever or button—say on a soda machine) are actual actions. But note that even complex decisions always end in taking an action. For example, buying a house entails signing a form. Getting married entails making a physical statement (saying "I do"). We are going to be less concerned about the physical muscle movements of the action than about the selection process that decided on which action to take. Nevertheless, defining "decision-making" as "action-selection" will force us to directly observe the decisions made. It will allow us to ask *why we take the actual actions we do.* Why don't those actions always match our stated intentions? How do we choose those actions over other potential actions? What are the systems that select actions, and how do those systems interact with the world? How do they break down? What are their vulnerabilities and failure-modes? That is what the rest of this book is about.

2

The Tale of the Thermostat

jet exhaust shimmers above the tarmac

I remember the strength of swans
on the lake up North, their wings
stretching forward, beating back…

acceleration pushes us into our seats
and we lift into the sky

Your brain is a decision-making machine, a complex but physical thing. Like any physical process, there are multiple ways in which decisions can go wrong. Being a physical being does not diminish who you are, but it can explain some of the irrational choices you make.

The decision-making that you do arises from the physical processes that occur in your brain. Because your brain is a physical thing, it has *vulnerabilities*, what one might call "failure-modes" in the engineering world. We see these vulnerabilities in our susceptibility to bad choices (*Do you really need a new sports car?*), in our susceptibility to addictions (*Why can't you just quit smoking those cigarettes?*), in our inability to control our emotions or our habits (*Why do you get so angry about things you can't change?*). We would like to believe that we are rational creatures, capable of logic, always choosing what's best for us. But anyone who has observed their own decisions (or those of their friends) will recognize that this is simply not correct. We are very complex decision-making machines, and sometimes those decision-making processes perplex us. Understanding how the brain makes decisions will help us understand ourselves. To understand those vulnerabilities, we need to understand the mechanism of decision-making in our brains.

Today, we are all familiar with complex machines, even complex machines that make decisions. The simplest machine that makes a decision is the thermostat—when the house is too hot, the thermostat turns on the air conditioning to cool it down, and when the house is too cold, the thermostat turns on the furnace to heat it up. This process is called *negative feedback*—the thermostat's actions are inversely related to the

difference between the temperature of the room and the temperature you'd like it to be (the *set-point*). But is the process really so simple? Taking the decision-making process of the thermostat apart suggests that even the simple decision-making process of the thermostat is not so simple.

The thermostat has three key components of decision-making that we will come back to again and again in this book. First, it *perceives the world*—the thermostat has a sensor that detects the temperature. Second, it *determines what needs to be done*—it compares that sensor to the set-point and needs to increase the temperature because it is too cold, needs to decrease the temperature because it is too hot, or doesn't need to do anything because the temperature is just right. Finally, it *takes an action*—it turns on either the furnace or the air conditioning.

In the artificial intelligence literature, there was an argument through much of the 1980s about whether a thermostat could have a "belief."[1] Fundamentally, a belief is a (potentially incorrect) representation about the world. Clearly, a working thermostat requires a representation of the target temperature in order to take actions reflecting the temperature of the outside world. But can we really say that the thermostat "recognizes" the temperature of the outside world? The key to answering this question is that the thermostat does not take actions based on the temperature of the world, but rather on its internal representation of the temperature of the world. Notice that the internal representation might differ from the real temperature of the room. If the sensor is wrong, the thermostat could believe that the room is warmer or cooler than it really is and take the wrong action.

One of the key points in this book is that knowing how the brain works allows us a better perception of what happens when something goes wrong. I live in Minnesota. In the middle of winter, it can get very cold outside. Imagine you wake up one morning to find your bedroom is cold. Something is wrong. But simply saying that the thermostat is broken won't help you fix the problem. We need to identify the problem, to *diagnose* it, if you will.

Maybe something is wrong with the thermostat's perception. Perhaps the sensor is broken and is perceiving the wrong temperature. In this case, the thermostat could think that the house is fine even though it is too cold. (Notice the importance of belief here—the difference between the thermostat's internal representation and the actual temperature can have a big impact on how well the thermostat makes its decision!) Maybe the set-point is set to the wrong temperature. This means that the thermostat is working properly—it has correctly moved the temperature of your house to the set-point, but that's not what you wanted. Or, maybe there's something wrong with the actions available to the thermostat. If the furnace is broken, the thermostat may be sending the signal saying "heat the house" but the house would not be heating correctly. Each of these problems requires a different solution. Just as there are many potential reasons why your bedroom is too cold and knowing how a thermostat works is critical to understanding how to fix it, when smokers say that they really want to quit smoking, but can't, we need to know where each individual's decision-making process has gone wrong or we won't be able to help. Before we can identify where the decision-making process has broken down, we're going to need to understand how the different parts of the brain work together to make decisions.

Many readers will object at this point that people are much more complicated than thermostats. (And we are.) Many readers will then conclude that people are not

machines. Back when negative feedback like the thermostat was the most complicated machine mechanism that made decisions, it was easy to dismiss negative feedback as too simple a model for understanding people. However, as we will see later in the book, we now know of much more complicated mechanisms that can make decisions. Are these more complicated mechanisms capable of explaining human decision-making? (I will argue that they are.) This leaves open some difficult questions: Can we be machines and still be conscious? Can we be machines and still be human?

The concept of conscious machines making decisions pervades modern science fiction, including the droids C3P0 and R2D2 in *Star Wars*, the android Data of *Star Trek: The Next Generation*, the desperate Replicants of Ridley Scott's *Blade Runner*, and the emotionally troubled Cylons of *Battlestar Galactica*. *Star Trek: The Next Generation* spent an entire episode (*The Measure of a Man*) on the question of how the fact that Data was a machine affected his ability to decide for himself whether or not to allow himself to be disassembled. In the episode, the judge concludes the trial with a speech that directly addresses this question—"Is Data a machine? Yes. Is he the property of Starfleet? No. We have all been dancing around the basic issue: does Data have a soul? I don't know that he has. I don't know that I have. But I have got to give him the freedom to explore that question himself. It is the ruling of this court that Lieutenant Commander Data has the freedom to choose."[2] We will examine the complex questions of self and consciousness in detail at the end of the book (*The Conundrum of Robotics*, Chapter 24), after we have discussed the mechanisms of decision-making. In the interim, I aim to convince you that we can understand the mechanisms of our decision-making process without losing the freedom to choose.

I don't want you to take the actual process of the thermostat as the key to the story here, anymore than we would take jet planes as good models of how swans fly. And yet, both planes and swans fly through physical forces generated by the flow of air over their specially shaped wings. Even though bird wings are bone, muscle, and feathers, while airplane wings are metal, for both birds and planes, lift is generated by airflow over the wings, and airflow is generated by speed through the air. If we can understand what enables a 30-ton airplane to fly, we will have a better understanding of how a 30-pound swan can fly. Even though the forward push through the air is generated differently, both birds and planes fly through physical interactions with the air. We will use analogous methods of understanding decision-making processes to identify and fix problems in thermostats and in ourselves, because both use identifiable computational decision-making processes.

A good way to identify where a system (like a thermostat) has broken down is a process called "differential diagnosis," as in *What are the questions that will differentiate the possible diagnoses?* My favorite example of this is the show *CarTalk* on National Public Radio, in which a pair of MIT-trained auto mechanics (Tom and Ray Magliozzi) diagnose car troubles. When a caller calls in with a problem, the first things they discuss are the basics of the problem. (Anyone who has actually listened to *CarTalk* will know that the first things Tom and Ray discuss are the person's name, where the caller is from, and some completely unrelated jokes. But once they get down to discussing cars, they follow a very clear process of differential diagnosis.) A typical call might start with the caller providing a description of the problem—"I hear a nasty sound when I'm driving." And then Tom and Ray will get down to business—they'll ask questions about

the sound: "What is the sound? Where is it coming from? Does it get faster as you go faster? Does it still happen when the engine is on but the car is not moving?" Each question limits the set of possible problems. By asking the right series of questions, one can progressively work one's way to identifying what's wrong with the car. If we could organize these questions into a series of rules, then we could write a computer program to solve our car problems. (Of course, then we wouldn't get to hear all the *CarTalk* jokes. Whether this is good or bad is a question of taste.)

In the 1980s, the field of artificial intelligence developed "expert systems" that codified how to arrange these sorts of question-and-answer rules to perform differential diagnosis.[3] At the time, expert systems were hailed as the answer to intelligence—they could make decisions as well as (or better than) experts. But it turns out that most humans don't make decisions using differential diagnoses. In a lot of fields (including, for example, medicine), a large part of the training entails trying to teach people to make decisions by these highly rational rules.[4] However, just because it is hard for people to make decisions by rule-based differential diagnoses does not mean that humans don't have a mechanism for making decisions. In fact, critics complained that the expert systems developed by artificial intelligence were not getting at the real question of what it means to be "intelligent" long before it was known that humans didn't work this way.[5] People felt that we understood how expert systems work and thus they could not be intelligent. A classmate in college once said to me that "we would never develop an artificial intelligence. Instead, we will recognize that humans are not intelligent." One goal of this book is to argue that we can recognize the mechanisms of human decision-making without losing our sense of wonder at the marvel that is human intelligence.

Some readers will complain that people are not machines; they have goals, they have plans, they have personalities. Because we are social creatures and much of our intelligence is dedicated to understanding each other, we have a tendency to attribute agency to any object that behaves in a complex manner.[6] Many of my friends name their cars and talk about the personality of their cars. My son named our new GPS navigation system "Dot." When asked why he named the GPS (the voice is definitely a woman's), he said, "So we can complain to her when she gets lost—'Darn you, Dot!'"

A GPS navigator has goals. (These are goals we've programmed in, but they are goals nonetheless.) Dot's internal computer uses her knowledge of maps and her calculation of the current location to make plans to achieve those goals. You can even tell Dot whether you prefer the plans to include more highways or more back-country scenic roads. If Dot were prewired to prefer highways or back-country scenic roads, we would say she has a clear personality. In fact, I wish I had some of Dot's personality—when we miss an exit, Dot doesn't complain or curse, she just says "recalculating" and plans a new route to her goal.

The claim that computers can have goals and that differences in how they reach those goals reflects their personality suggests that goals and plans are simple to construct and that personality is simply a difference in underlying behavioral variables and preferences. We explain complex machines by thinking they're like people. Is it fair to turn that on its head and explain people as complex machines?

In this book, I'm going to argue that the answer to this question is "yes"—*your brain is a decision-making machine*, albeit a very complex one. You are that decision-making machine. This doesn't mean you're not conscious. This doesn't mean you're not *you*. But it can explain some of the irrational things that you do.

Understanding how the human decision-making system works has enormous implications for understanding who we are, what we do, and why we do what we do. Scientists study brains, they study decision-making, and they study machines. By bringing these three things together, we will begin to get a sense of ourselves. In this book, I will discuss what we know about how brains work, what we know about how we make decisions, and what we know about how that decision-making machine can break down under certain conditions to explain irrationality, impulsivity, and even addiction.

This page intentionally left blank

3

The Definition of Value

diamond stars in a circle of gold
declare your love
with two months salary

Value measures how much one is willing to pay, trade, or work for a reward, or to work to avoid a punishment. Value is not intrinsic to an object, but must be calculated anew each time. This produces inconsistencies, in which different ways of measuring value can produce different orderings of one's preferences.

What is "value"? What does it mean to value something? In psychological theories, value is thought of as that which reduces needs or which alleviates negative prospects[1]—when you are hungry, food has value; when you are thirsty, water has value. Pain signals a negative situation that must be alleviated for survival. In economics, the concept of *revealed preferences* says that we value the things we choose, and make decisions to maximize value.[2] Although this is a circular definition, we will use it later to help us understand how the decision-making process operates. In the robotics and computer simulation worlds where many of the mechanisms that we will look at later have been worked out, value is simply a number r for reward or $-r$ for punishment.[3] The robotics and computer science models simply try to maximize r (or minimize $-r$).

In many experiments that study decision-making, value is measured in terms of money. But of course, money has true value only in terms of what it can buy.[4] Economists will often tell us that money is an agreed-upon fiction—I am willing to sell you this apple for a green piece of paper that we both agree is worth $1 because I am confident that I can then use the paper later to buy something that I want. The statement that money is valuable only in terms of what it can buy is not entirely true either: money can also have value as a symbol of what it implies for our place in the social network.[5] This is one reason why extremely well-paid professional athletes fight for an extra dollar on their multimillion dollar salaries—they want to be known as the "best-paid wide receiver" because that identification carries social value. Money and value are complex concepts that interact in difficult ways.

Part of the problem is that you cannot just ask people *How much do you value this thing?* We don't have units to measure it in. Instead, value is usually measured in what people will trade for something and in the decisions they make.[6] This is the concept called *revealed preferences*—by examining your decisions, we can decide what you value. Theoretically, this is very simple: one asks *How much will you pay for this coffee mug?* The price you'll pay is the value of the coffee mug. If we measure things at the same time, in the same experiment, we shouldn't have to worry about inflation or changes in the value of the money itself.

The problem is that people are inconsistent in their decisions. In a famous experiment, Daniel Kahneman and his colleagues Jack Knetsch and Richard Thaler divided their subjects into two groups: one group was asked how much money they would pay for a $6 (in 1990) Cornell University coffee mug and the other group was first given the coffee mug and then asked how much money they would accept for the mug. The first group was willing to pay much less on average ($3) than the second group was willing to accept for it ($7). This is called *the endowment effect* and is a preference for things you already own.[7]

Simply phrasing decisions in terms of wins or losses changes what people choose.[8] The interesting thing about this is that even once one is shown the irrationality, it still feels right. This effect was first found by Daniel Kahneman and Amos Tversky, who made this discovery sitting in a coffee shop in Israel, asking themselves what they would do in certain situations. When they found themselves making irrational decisions, they took their questions to students in their college classes and measured, quantitatively, what proportion made each decision in each condition.

In the classic case from their 1981 paper published in the journal *Science*:[9]

> Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:
>
> Problem 1: If Program A is adopted, 200 people will be saved. If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved. Which of the two programs would you favor?
>
> Problem 2: If Program C is adopted, 400 people will die. If Program D is adopted, there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die. Which of the two programs would you favor?

A careful reading of the two conditions shows that Program *C* is identical to Program *A*, while Program *D* is identical to Program *B*. Yet, 72% of the first group chose program *A* and 28% chose program *B*, while only 22% of the second group chose program *C* and 78% chose program *D*! Kahneman and Tversky interpret this as implying that we are more sensitive to losses than to gains—we would rather risk not gaining something than we would risk losing something. This is clearly a part of the story, but I will argue that there is a deeper issue here. I will argue that "value" is something that we calculate each time, and that the calculation process doesn't always come up with the same "rational" answer.

Illogical value calculations

The most interesting thing about these results (of which there are many) is that even when they are illogical (like the disease example above), they often still sound right. There are other examples where we can see that decisions are irrational, and don't sound right when they are pointed out to us, but which humans definitely show when tested. The best examples of most of these come from the advertising and marketing worlds.

A person goes into an electronics store to buy a television and finds three televisions on sale, one for $100, one for $200, and one for $300. Imagine that the three televisions have different features, such that these are reasonable prices for these three TVs. People are much more likely to buy the $200 TV than either of the other two. If, in contrast, the person goes into the store and finds three televisions on sale, the same $200 TV, the same $300 TV, and now a fancier $400 TV, the person is more likely to buy the $300 TV. In the first case, the person is saying that the $200 TV is a better deal than the $300 one, but in the second case, the person is saying that the $300 TV is the better deal. Even though they've been offered the same televisions for the same prices, the decision changed depending on whether there is a $100 TV or a $400 TV in the mix. This is called *extremeness aversion* and is a component of a more general process called *framing*. In short, the set of available options changes your valuation of the options. This is completely irrational and (unlike the other examples above) seems unreasonable (at least to me). Yet it is one of the most reliable results in marketing and has probably been used since the first markets in ancient times through to the modern digital age.[10]

Products in advertisements used to compare themselves to each other. Tylenol would say it was better than aspirin and Coke would say it was better than Pepsi. I remember an RC Cola commercial with two opposing teams fighting about which drink they preferred, Coke or Pepsi, while a third person sits on the sidelines drinking an RC Cola, out of the fray, smiling. Advertisements today rarely mention the competition. This is because one of the heuristics we use is simply whether we recognize the name or not.[11] So although RC Cola was trying to communicate that Coke and Pepsi were the same, while RC Cola was different, what people took from the advertisement was a reminder that everyone drank Coke and Pepsi. Just mentioning the name reminds people that it exists and reinforces the decision to choose it. Familiarity with a brand name increases the likelihood that one will select it.

It is election season as I write this. All around my neighborhood, people have put out signs for their candidates. The signs don't say anything about the candidates. Often they don't even have the party the candidates belong to. Usually, it's just the name of the candidate, and sometimes, an appeal to "vote for" the candidate. What information do I get from seeing a sign that says nothing other than "Vote for X"? I suppose that one may need to be reminded to vote at all, but then why does the sign include "for X" on it? (One of my neighbors does have a large handmade sign she puts out every election season that just says "VOTE!" on it.) It is true that one can get a sense of the grouping of candidates from people who put out multiple signs: given that I like person X for state representative, seeing that all the people with person X also have person Y for county prosecutor, while all the people who have person X's opponent have person Z for county prosecutor, might suggest to me that I would like person Y over person Z for

county prosecutor. But lots of people have just one sign out. All that tells me is that lots of people like person X. Why would knowing that lots of people like something suggest that I would too?

There are actually three things that these single-sign houses are doing. First, they are increasing my familiarity with that name. As with the products, just knowing the name increases one's likelihood of voting for someone. Second, if lots of people like a movie, it's more likely to be a good movie than one everybody hated. Using the same heuristic, if everyone likes a candidate, isn't that candidate more likely to be a good choice? And third, we like to be on the winning team. If everyone else is going to vote for someone, he or she is likely to win. Of course, we're supposed to be voting based on who we think is a better choice to govern, not who is most popular. But it's pretty clear that a lot of people don't vote that way.

These effects occur because we don't actually calculate the true value of things. Instead, we use rules of thumb, called *heuristics*, that allow us to make pretty good guesses at how we value things.[12] If you like all the choices, picking the middle one is a pretty good guess at good value for your money. If you're making a decision, familiarity is a pretty good guess. (Something you're familiar with is likely to be something you've seen before. If you remember it, but don't remember it being bad, how bad could it have been?)

Some economists have argued that evolutionarily, heuristics are better than actually trying to calculate the true value of things because calculating value takes time and heuristics are good enough to get us by.[13] But a lot of these nonoptimal decisions that we're making are taking time. Knowing that Programs *A* and *C* are the same and that programs *B* and *D* are the same in the Kahneman and Tversky flu example above doesn't change our minds. This makes me suspect that something else is going on. It's not that heuristics are faster and we are making do with "good enough." I suspect that these effects have to do with how we calculate value. We cannot determine how humans calculate value unless we can measure it. So, again, we come back to the question of how we measure value.

Measuring value

With animals, we can't ask them how much they value something; we can only offer them something and determine how much they'll pay for it. Usually, this is measured in terms of the amount of effort an animal will expend to get the reward.[14] *How many lever presses is the animal willing to make for each reward?*

Alternatively, we can give the animal a direct choice between two options:[15] *Would the animal rather have two apple-flavored food pellets or one orange-flavored food pellet? Would it rather have juice or water?* We can also titrate how much of one choice needs to be offered to make the animal switch—if the animal likes apple flavor better than orange flavor, would it still prefer half as much apple to the same amount of orange?

Finally, we can also measure the negative consequences an animal will put up with to get to a reward.[16] Usually, this is measured by putting a negative effect (a small shock or a hot plate) in between the animal and the reward, which it has to cross to get to the reward. (It is important to note that crossing the shock or hot plate is entirely up to the animal. It can choose not to cross the punishment if it feels the reward is not valuable enough.) A common experiment is to balance a less-appealing reward with

a more-appealing reward that is given only after a delay.[17] Because animals are usually hungry (or thirsty) when running these experiments, they don't want to wait for reward. Thus delay becomes a very simple and measurable punishment to use—*how long will the animal wait for the reward?*

Of course, humans are animals as well and all of these experiments work well in humans: How much effort would a smoker spend to get a puff of a cigarette?[18] How much money will you trade for that coffee mug?[19] What negative consequences will you put up with to achieve your reward?[20] How long will you wait for that larger reward?[21] Will you wait for a lesser punishment, knowing that it's coming?[22]

These four ways of measuring value all come down to quantifiable observations, which makes them experimentally (and scientifically) viable. Economists, who study humans, typically use hypothetical choices ("What would you do if…?") rather than real circumstances, which, as we will see later, may not access the same decision-making systems. It has been hard to test animals with hypothetical choices (because hypothetical choices are difficult to construct without language), but there is some evidence that chimpanzees with linguistic training will wait longer for rewards described by symbols than for immediately presented real rewards.[23] (Remember the chimpanzees and the jellybeans?) This, again, suggests that language changes the decision-making machinery. More general tests of hypothetical rewards in animals have been limited by our ability to train animals to work for tokens. Recent experiments by Daeyeol Lee and his colleagues getting monkeys to work for tokens may open up possibilities, but the critical experiments have not yet been done.[24]

One of the most common ways to measure willingness to pay for something is a procedure called the *progressive ratio*—the subject (whether it be human or not) has to press a lever for a reward, and each time it receives the reward, the number of times it has to press the lever for reward increases, often exponentially. The first reward costs one press, the second two, the third four, the fourth eight, and so on. Pretty soon, the subject has to press the lever a thousand times for one more reward. Eventually, the animal decides that the reward isn't worth that much effort, and the animal stops pressing the lever. This is called the *break point*. Things that seem like they would be more valuable to an animal have higher break points than things that seem like they would be less valuable. Hungry animals will work harder for food than satiated animals.[25] Cocaine-addicted animals will work harder for cocaine than for food.[26] A colleague of mine (Warren Bickel, now at Virginia Tech) told me of an experimental (human) subject in one of his experiments who pulled a little lever back and forth tens of thousands of times over the course of two hours for one puff of nicotine!

It's been known for a long time that drugs are not infinitely valuable. When given the choice between drugs and other options, both human addicts and animals self-administering drugs[A] will decrease the amount of drug taken as it gets more expensive relative to the other options.[27] Drugs do show what economists call *elasticity*: the more expensive they get, the less people take. This is why one way to reduce smoking in a population is to increase the tax on cigarettes.

[A] The animal experimental literature is generally unwilling to call animals "addicts" and instead refers to their behavioral actions as "self-administering drugs." As we will see later in the book, animals self-administer the same drugs that humans do, and animals self-administering drugs show most of the same behaviors that humans self-administering drugs do.

Elasticity measures how much the decision to do something decreases in response to increases in cost. Luxuries are highly elastic; nonluxuries are highly inelastic. As the costs of going to a movie or a ballgame increase, the likelihood that people will go decreases quickly. On the other hand, even if the cost of food goes up, people aren't about to stop buying food. Some economists have argued that a good definition of addiction is that things we are addicted to are inelastic. This allows them to say that the United States is "addicted to oil" because we are so highly dependent on driving that our automobile use is generally inelastic to the price of oil. However, again, we face an interesting irrationality. The elasticity of automobile use is not linear[28]—raising the price of gasoline from $1 per gallon to $2 had little to no effect on the number of miles driven, but when a gallon of gasoline crossed the $3 mark, there was a sudden and dramatic drop in the number of miles driven in 2005. People suddenly said, "That's not worth it anymore." Of course, people then got used to seeing $3 per gallon gasoline and the number of miles driven has begun to increase again.

This irrationality is where the attempt to measure value gets interesting. In a recent set of experiments, Serge Ahmed and his colleagues found that even though the break point for cocaine was much higher than for sweetened water,[B] when given the choice between cocaine and sweetened water, almost all of the animals chose the sweetened water.[29] (Lest readers think this is something special about cocaine, the same effects occur when examining animals self-administering heroin.[30]) Measuring value by how much the animal was willing to pay said that cocaine was more valuable than sweetened water. (Cocaine had a higher break point than the sweetened water did.) But measuring value by giving the animals a choice said that the sweetened water was more valuable. (They consistently chose the sweetened water over the cocaine.) The two measures gave completely different results as to which was more valuable.

One of my favorite examples of this irrationality is a treatment for addiction called *Contingency Management*, where addicts are offered vouchers to stay off drugs.[31] If addicts come into the clinic and provide a clean urine or blood sample, showing that they have remained clean for the last few days, they get a voucher for something small—a movie rental or a gift certificate. Even very small vouchers can have dramatic effects. What's interesting about this is that raising the cost of a drug by $3 would have very little effect on the amount of drug an addict takes (because drugs are inelastic to addicts), but providing a $3 voucher option can be enough to keep an addict clean, straight, and sober. Some people have argued that this is one of the reasons that Alcoholics Anonymous and its 12-step cousins work.[C] It provides an option (going to meetings and getting praised for staying sober) that works like a voucher; it's a social reward that can provide an alternative to drug-taking.[32]

So why is it so hard to measure value? Why are we so irrational about value? I'm going to argue that value is hard to measure because value is not something intrinsic to an object.

[B] Ahmed and colleagues used saccharin in the water rather than sugar because saccharin has no direct nutritive value but tastes sweet to both rats and humans.

[C] The Anonymous meetings (AA, Narcotics Anonymous, Gamblers Anonymous, etc.) and their 12-step cousins also include additional elements that seem to access important decision-making components. One example is the presence of a "sponsor" (someone you can call 24/7), who provides a low-cost option to pull one away from the high-value drug-taking option. Suicide hotlines also work this way, providing a very low-cost option (a phone call) that can short-circuit a dangerous path.

First, I'm going to argue in this book that there are multiple decision-making systems, each of which has its own method of action-selection,[33] which means there are multiple values that can be assigned to any given choice. Second, in the deliberative experiments we've been looking at, we have to calculate the value of the available options each time.[34] (Even recognizing that options are available can be a complex process involving memory and calculation.) The value of a thing depends on your needs, desires, and expectations, as well as the situation you are in. By framing the question in different ways, we can guide people's attention to different aspects of a question and change their valuation of it. In part, this dependence on attention is due to the fact that we don't know what's important. A minivan can carry the whole family, but the hybrid Prius gets better gas mileage, and that convertible BMW looks cooler. It's hard to compare them.

This means that determining the true value of something depends on a lot of factors, only some of which relate to the thing itself. We'll see later that value can depend on how tired you are, on your emotional state, and on how willing you are to deliberate over the available options.[35] It can even be manipulated by changing unrelated cues, such as in the *anchor effect*, in which unrelated numbers (like your address!) can make you more likely to converge to a higher or lower value closer to that number.[36]

Economists argue that we should measure value by the reward we expect to get, taking into account the probability that we will actually get the reward, and the expected investment opportunities and risks.[37] If we tried to explore all possibilities and integrate all of this, we could sit forever mulling over possibilities. This leads us to the concept of *bounded rationality*, introduced by Herb Simon in the 1950s, which suggests that the calculation takes time, that sometimes it's better to get to the answer quickly by being less complete about the full calculation, and that sometimes a pretty good job is good enough.[38] Instead, we use heuristics, little algorithms that work most of the time.

The problem with this theory is that even when we are given enough time, we continue to use these heuristics. Economists and psychologists who argue for bounded rationality (such as Gerd Gigerenzer[39]) argue that evolution never gives us time. But if the issue were one of time, one would expect that the longer one was given, the more rational one would be. This isn't what is seen. People don't change their minds about taking the middle-value television if they are given more time; in fact, they become *more* likely to pick the middle television, not less, with more time.[40]

We pick the middle television in the three-television example because one of the algorithms we use when we want to compare the choices immediately available to us is to find the middle one. Another example is that we tend to round numbers off by recognizing the digits.[41] $3.99 looks like a lot less than $4.00. Think about your "willingness to pay" $3.75 per gallon for gasoline relative to $3.50 per gallon. If both options are available, obviously, you'll go to the station selling it for $3.50. But if it goes up from $3.50 to $3.75 from one day to the next, do you really stop buying gasoline? Now, imagine that the price goes up from $3.75 to $4.00. Suddenly, it feels like that gasoline just got too expensive. We saw a similar thing at the soda machine by my office. When the cost went up from 75¢ to $1, people kept buying sodas. But then one day it went up from $1 to $1.25, and people said, "It's not worth it" and stopped. (It's now $1.50 and no one ever buys sodas there anymore.) Sometimes the cost crosses a line that we simply won't put up with anymore.

Part of this is due to the mechanisms by which we categorize things. Whole dollar amounts draw our attention and we are more likely to pick them. In a recent experiment, Kacey Ballard, Sam McClure, and their colleagues asked people which they would prefer, $7 today or $20 in a week.[42] (This is a question called *delay-discounting*, which we will examine in detail later in our chapter on impulsivity [Chapter 5].) But then they asked the subjects to decide between $7.03 today and $20 in a week. People were more likely to pick the $20 over $7.03 than $20 over $7 even. In a wonderful control, they then asked the subjects to decide between $7 today and $20.03 in a week. This time, people were more likely to pick the $7. There is no way that these decisions are rational, but they do make sense when we realize (1) that making the decision requires determining the value of the two options anew each time and comparing them, and (2) that we use heuristics that prefer even dollar amounts to compare them.[D]

Value is an important concept to understanding decision-making, but our brains have to calculate how much we value a choice. That calculation is based on heuristics and simple algorithms that work pretty well most of the time but can be irrational under certain conditions. How do we actually, physically determine value? What are the neurophysiological mechanisms? That brings us to the differences between pleasure, pain, and the do-it-again signal.

Books and papers for further reading

- Daniel Ariely (2008). *Predictably Irrational: The Hidden Forces that Shape Our Decisions*. New York: HarperCollins.
- Daniel Kahneman (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.
- Scott Plous (1993). *The Psychology of Judgment and Decision-Making*. New York: McGraw-Hill.

[D] Why people prefer even dollar amounts is not known. Perhaps even dollar amounts are easier to calculate with. Perhaps people are suspicious of odd amounts because they are concerned that there's a trick being pulled (like the extra $\frac{9}{10}\,¢$ used by American gas stations). Perhaps people can measure the even dollar amounts more easily, which may draw their attention. Whatever the heuristic, people do prefer even dollar amounts.[43]

4

Value, Euphoria, and the Do-It-Again Signal

At the base of *dead man's hill,*
sled toppled in a mound of snow,
exhilaration lights up his face.
Let's do that again!

Pleasure and value are different things. One can have pleasure without value and value without pleasure. They are signaled by different chemicals in the brain. In particular, the brain contains a signal for "better than expected," which can be used to learn to predict value. The lack of expected reward and the lack of expected punishment are separate processes (disappointment, relief) that require additional neural mechanisms.

Euphoria and dysphoria

It is commonly assumed that pleasure is our interpretation of the things we have to do again. "People seek pleasure and avoid pain." While this is often true, a better description is that people seek things that they recognize will have high value. As we will see below, pleasure is dissociable from value.

Although the saying is "pleasure and pain," pain is not the opposite of pleasure. Pain is actually a sensory system, like any of the other sensory systems (visual, auditory, etc.).[1] Pain measures damage to tissues and things that are likely to damage tissues. It includes specific receptors that project up to mid-brain sensory structures that project to cortical interpretation areas. The sensation of pain depends on cortical activation in response to the pain sensors, but this is true of the other sensory systems as well. The retina in your eyes detects photons, but you don't see photons—you see the objects that are interpreted from photons. Pleasure, in contrast, is not a sensation, but an evaluation of a sensation.

Euphoria, from the Greek word φορία (*phoria*, meaning "bearing" or "feeling") and the Greek root εύ- (*eu-*, meaning "good"), is probably a better word than "pleasure" for the brain signal we are discussing. And, of course, euphoria has a clear antonym in *dysphoria*, meaning "discomfort," deriving from the Greek root δυσ- (*dys-*, bad). The terms

euphoria and *dysphoria* represent calculations in the brain of how good or bad something is.

Differences between euphoria and reinforcement

Watching animals move toward a reward, early psychologists were uncomfortable attributing emotions to the animal, saying that it "wanted the reward" or that it felt "pleasure" at getting the reward. Instead they defined something that made the animal more likely to approach it as *reinforcement* (because it *reinforced* the animal's actions).

Experiments in the 1950s began to identify that there was a difference between euphoria and reinforcement in humans as well. In the 1950s, it was found that an electrode placed into a certain area of the brain (called the medial forebrain bundle) would, when stimulated, lead to near-perfect reinforcement.[2] Animals with these stimulations would forego food, sex, and sleep to continue pressing levers to produce this brain stimulation.

From this, the medial forebrain bundle became popularly known as "the pleasure center."[3] But even in these early experiments, the difference between euphoria and reinforcement was becoming clear. Robert Heath and his colleagues implanted several stimulating electrodes into the brain of a patient (for the treatment of epilepsy), and then offered the patient different buttons to stimulate each of the electrodes. On stimulation from pressing one button, the patient reported orgasmic euphoria. On stimulation from the second button, the patient reported mild discomfort. However, the patient continued pressing the second button over and over again, much more than the first. Euphoria and reinforcement are different.[4]

It turns out that the key to the reinforcement from medial forebrain bundle stimulation is a brain neurotransmitter called *dopamine*.[5] Dopamine is chemically constructed out of precursors[A] in a small area of the brain located deep in the base of the midbrain called the *ventral tegmental area* and the *substantia nigra*.[B]

[A] The pharmacology term "precursors" is used to identify molecules that are converted into the molecule being studied.[6] Your cells are, in a sense, chemical factories that convert molecules to other molecules, using specialized molecules called *enzymes*. For example, the chemical levodopa (L-dopa) is a precursor for dopamine, which means that if you had extra levodopa in your system, your cells would have more building blocks with which to make more dopamine, and you would find that your cells had more dopamine to use. This is why levodopa forms a good treatment for Parkinson's disease, since Parkinson's disease is, in fact, a diminishment of dopamine production in certain areas of the brain.[7]

[B] The ventral tegmental area is so called because it is *ventral* (near the base of the brain) in the *tegmentum* (Latin for "covering"), an area that covers the brain stem. *Substantia nigra* is Latin for "black stuff," so named because it contains melatonin, which makes the area appear black on a histological slice.

Melatonin is a chemical used by the body to signal the nighttime component of day/night cycles.[8] This is why it is sometimes used to reset circadian rhythm problems. It interacts with dopamine, particularly the release of dopamine in the brain, which means that manipulations of melatonin (say for jetlag) can affect learning and cognition. This is why pharmacology is so complex:[9] manipulating one thing often produces lots of other effects.

An early theory of dopamine was that it was the representation of pleasure in the brain.[10] Stimulating the medial forebrain bundle makes dopamine cells release their dopamine.[11] (Remember that these stimulating electrodes were [incorrectly] hypothesized to be stimulating the "pleasure center.") In animal studies, blocking dopamine blocked reinforcement.[12] Although one could not definitively show that animals felt "pleasure," the assumption was that reinforcement in animals translated to pleasure in humans. Most drugs of abuse produce a release of dopamine in the brain.[13] And, of course, most drugs of abuse produce euphoria, at least on early use.[C] Since some of the largest pharmacological effects of drugs of abuse are on dopamine, it was assumed that dopamine was the "pleasure" signal. This theory turned out to be wrong.

In a remarkable set of experiments, Kent Berridge at the University of Michigan set out to test this theory that dopamine signaled pleasure.[15] He first developed a way of measuring euphoria and dysphoria in animals—he watched their facial expressions. The idea that animals and humans share facial expressions was first proposed by Charles Darwin in his book *The Expression of the Emotions in Man and Animals*. Berridge and his colleagues used cameras tightly focused on the faces of animals, particularly rats, to identify that sweet tastes (such as sugar or saccharin) were accompanied by a licking of the lips, while bitter tastes (such as quinine) were accompanied by a projection of the tongue, matching the classic "yuck" face all parents have seen in their kids being forced to eat vegetables.

What is interesting is that Berridge and his colleagues (particularly Terry Robinson, also at the University of Michigan) were able to manipulate these expressions, but not with manipulations of dopamine. Manipulations of dopamine changed whether an animal would work for, learn to look for, or approach a reward, but if the reward was placed in the animal's mouth or even right in front of the animal, it would eat the reward and show the same facial expressions. What did change the facial expressions were manipulations of the animal's *opioid* system. The opioid system is another set of neurotransmitters in the brain, which we will see are very important to decision-making. They are mimicked by opiates—opium, morphine, heroin. Berridge and Robinson suggested that there is a distinction between "wanting something" and "liking it." Dopamine affects the wanting, while opioids affect the liking.

Opioids

The endogenous opioid system includes three types of opioid signals and receptors in the nucleus accumbens, hypothalamus, amygdala, and other related areas.[16] Neuroscientists have labeled them by three Greek letters (mu [μ], kappa [κ], and delta [δ]). Each of these receptors has a paired endogenous chemical (called a ligand) that attaches to it (*endorphins* associated with the μ-opioid receptors, *dynorphin* associated with the κ-opioid receptors, and *enkephalins* associated with the δ-opioid receptors). Current work suggests that activation of the μ-opioid receptors signals euphoria, and activation of the κ-opioid receptors signals dysphoria. The functionality of the δ-opioid

[C] Not all abused drugs are euphoric. Nicotine, for example, is often highly dysphoric on initial use, even though it is one of the most reinforcing of all drugs.[14]