© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

CONTENTS

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

1

Player One

Are we the authors of our own stories? Or is our apparent freedom of choice really an illusion? These questions were brought home to me recently as I was watching my son play a video game—one where you wander around an open world, meeting interesting denizens of one type or another (and killing quite a few of them). As I watched, his character entered a tavern and approached the bartender, who offered a generic greeting. The game then threw up some options for things you could say in reply to get information about the prospects for fortune and glory in the surrounding territory.

In this exchange, my son's possibilities for action were limited by the game, but he *was* really making choices among them, and these choices then affected how the conversation went and what would subsequently unfold. His decisions were based on his overall goal in the game, the tension between his goals of taking some immediate action or to keep exploring, his need to have enough information to make a decision with confidence, the risk of biting off more than he could chew and losing his hard-won stuff: all these considerations fed into the decisions he made. He had his reasons and he acted on them, just like you or I do every day, all day long.

The bartender, in contrast, was not making choices. He was a classic "non-player character," an NPC. His responses were completely determined by his programming: he had no degrees of freedom. His actions were merely the inevitable outcome of a flow of electrons through the circuits of the game console, constrained by the rules encoded in the

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

software. Even the more sophisticated NPCs in the game, including the monster that eventually caramelized my son's avatar, were similarly constrained. The monster's actions—even in the fast-moving melee—were determined by the software programming and mediated by the electronic components in the console.

Thus the NPCs only *appear* to be making choices. They're not autonomous entities like us: they're just a manifestation of lots of lines of code, implemented in the physical structure of the computer chips. Their behavior is entirely determined by the inputs they get and their preprogrammed responses. We, in contrast, are causes of things in our own right. We have *agency*: we make our own choices and are in charge of our own actions.

At least it seems that way. It certainly feels like we have "free will," like we make choices, like we are in control of our actions. That's pretty much what we do all day—go around making decisions about what to do. Some are trivial, like what to have for breakfast; some are more meaningful, like what to say or do in social or professional situations; and some are momentous, like whether to accept a job offer or a marriage proposal. Some we deliberate on consciously, and others we perform on autopilot—but *we* still perform them. Of course, our options may be more or less constrained (or informed) by all kinds of factors at any given moment, but generally we feel like the authors of our own actions.

And we interpret other people's behavior in terms of their reasons for selecting different actions—their intentions, beliefs, and desires that make up the content of their mental states. We constantly analyze each other's motives and habits and character, looking for explanations and predictors of their behavior and the decisions they make. Why people act the way they do is ultimately the theme of most entertainment, from Dostoyevsky to *Big Brother*. All this rests on the view that we are not just acted on—we are actors. Things don't just happen to us, in the way they happen to rocks or spoons or electrons: *we do things*.

The problem is that, if you think about this view for too long, it becomes difficult to escape a discomfiting thought. After all, like the NPCs, our decisions, however complex they may be, are mediated by the flow of electrical ions through the circuits of our brains and thus are constrained by our own "programming," by how our circuits are configured.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Unless you invoke an immaterial soul or some other ethereal substance or force that is really in charge—call it spirit or simply mind, if you prefer—you cannot escape the fact that our consciousness and our behavior emerge from the purely physical workings of the brain.

There is no shortage of evidence for this from our own experience. If you've ever been drunk, for example, or even just a little tipsy, you've experienced how altering the physical workings of your brain alters your choices and the way you behave. There is a whole industry of recreational drugs—from caffeine to crystal meth—that people take because of the way that physically tweaking the brain's machinery in various ways makes them feel and act. The ultimate consequence in some cases is addiction—perhaps the starkest example of how our actions can sometimes be out of our control.

And, of course, if the machinery of your brain gets physically damaged—as occurs with head injuries, strokes, brain tumors, neurodegenerative disorders, or a host of other kinds of insults—or its function is impaired in other ways, as in conditions such as schizophrenia, depression, or mania, then your ability to choose your actions may also be impaired. In some situations the integrity of your very *self* may be compromised.

We all like to think that we are Player One in this game of life, but perhaps we are just incredibly sophisticated NPCs. Our programming may be complex and subtle enough to make it *seem* as if we are really making decisions and choosing our own actions, but maybe we're just fooling ourselves. Perhaps "we" are just the manifestations of genetic and neural codes, implemented in biological rather than computer hardware. Perhaps we are the victims of a cruel joke, tragic figures in the grip of the Fates. As Gnarls Barkley sang, "Who do you, who do you, who do you think you are? Ha ha ha, bless your soul, you really think you're in control."

Robots with Personality

In my 2018 book *Innate* I described how we all come pre-wired with a set of innate psychological predispositions. At the most basic level, we all share the profile of human nature. Evolution has shaped the behavior

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

of our species just as much as that of any other. Human nature is encoded in our DNA in a genetic program that specifies the building and wiring of our human brains.

However, the details of that genetic program inevitably vary among individuals. I use the word "inevitably" because there is no way that this variation could not exist. Every time DNA is copied in a cell, including when sperm or egg cells are made, some small number of copying errors or mutations arise. New variations in the DNA sequence thus enter the gene pool in every generation, and—if their effects are tolerated—they can spread through the population over time, leading to the accumulation of genetic variation that we observe.

This leads to the differences that we observe in people's physical traits, such as how tall they are or the shape of their faces or various aspects of their physiology. This variation occurs just the same in the physical structure of their brains and the way they function. The fact that all these traits are affected by genetic variation explains why people who are related to each other resemble each other more than do unrelated people, both physically and psychologically. So, even though the "canonical" human genome (which doesn't really exist anywhere) encodes a program to build a canonical human brain, your particular genome encodes a program to build *a brain like yours*.

But not exactly like yours. The program in your genome does not encode one particular outcome, specified down to the level of individual nerve cells or synaptic connections between them. It does not encode the outcome at all, in fact: it just encodes a set of biochemical rules that, when played out over the complicated processes of development, will tend to result in an outcome within a certain range. Exactly how these processes played out in your specific case was also affected by all kinds of random events during development that added considerable variation. If you ran the program again, you would not get exactly the same outcome. Even the brains of identical twins who share the same genetic program are quite distinct from each other already at birth.

All this means that the way your brain is wired is affected by millions of years of common evolution, by the specific genetic variations that you carry, and by the unique trajectory of development that occurred

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

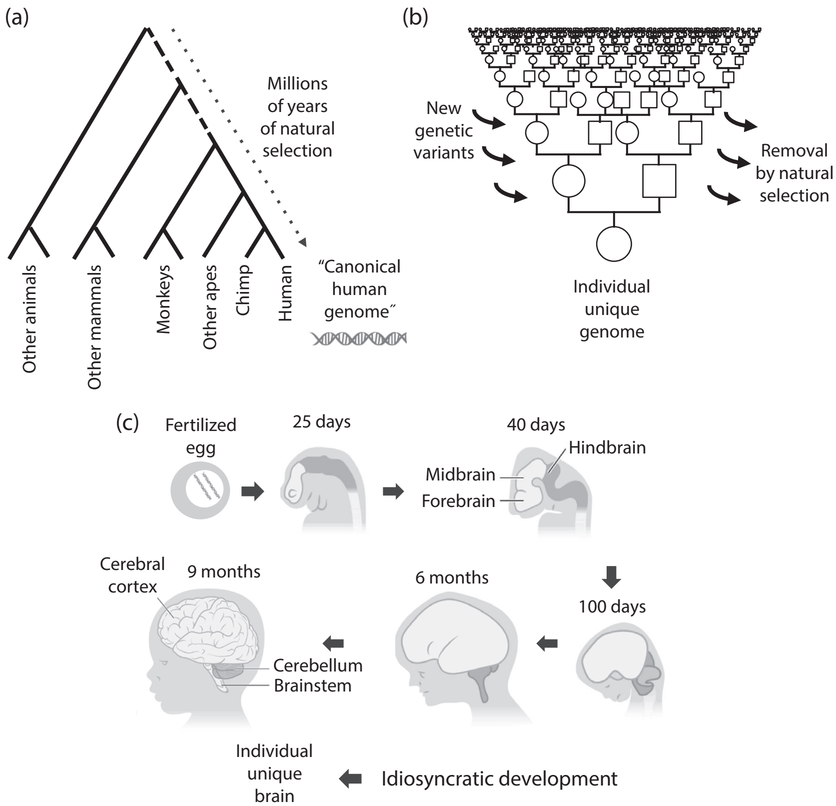

FIGURE 1.1. The making of you. **(a)** Like all species, the genome of Homo sapiens has been shaped by millions of years of evolution, selecting for all the traits that comprise "human nature" generally. **(b)** Your individual genome is a unique version of the canonical human genome, reflecting the processes of mutation and selection in your specific ancestors. **(c)** The outcome of brain development in any individual is idiosyncratic, shaped by genetic variation and the unique trajectories of development itself. Our individual natures (or innate predispositions) are thus variations on the theme of human nature generally.

while your brain was developing (see Figure 1.1). And the way your brain is wired affects how it works and how you will tend to behave.

We can think of this variation like that in the internal tuning of the behavioral controls of a robot. Imagine you and I were asked to build an autonomous robot that has to make its way in the world—finding

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

fuel, avoiding threats, interpreting sensory information, assessing situations, and deciding among possible actions.

There are all kinds of things we would have to build in our robot for it to accomplish its tasks. It would need some sensors, of course, to detect things in the environment, and it would require motors so it could move around and perform various actions. It would need to be programmed to move toward fuel and away from threats, but it might also require some fancy circuitry for it to recognize which is which. And what would happen if fuel supplies and something threatening happen to be in the same place? It would have to weigh the opportunity versus the risk and make a decision accordingly about where to move. And it would be good if that decision were informed by how much fuel it had left at the time. So some way to monitor its internal states and use them to inform decisions would certainly be beneficial.

A very fancy robot might also be able to learn from experience; for example, it might learn that there tends to be fuel in some particular spot or that some kind of otherwise innocuous stimulus (a rustle in the robot grass perhaps) signals a hidden threat. Now imagine we give our robot another goal: not just to survive but also to find robot love and reproduce. Then it would have to balance the short-term goal of ensuring it has enough fuel with the longer-term goal of finding a mate, all while not getting destroyed by a bigger robot.

All those functions—some means for inferring what is out in the world from the data gathered by its sensors; integrating both external and internal multiple signals to derive a picture of the whole situation; comparing that with the data in its memory bank to help inform its next action; weighing threats versus opportunities, short-term versus long-term goals, and good versus bad outcomes; and eventually picking one action to perform while inhibiting all other possibilities—would have to be configured into its circuitry.

With so many circuits and parameters that could vary, it is inevitable that the way you would tune your robot would differ from how I would tune mine. You might set the threat sensitivity a little higher and the reward sensitivity a little lower. I might tune the circuits in my robot with a different balance between short- and long-term goals. All these

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

settings would manifest as different *patterns of behavior* over time and across contexts. Your robot might appear more cautious than mine. Mine might show more perseverance: it might be willing to work longer for a delayed reward. The robots could differ in how much evidence they need to make a decision (impulsivity), how much they value mating opportunities (sex drive), and even how salient they find novel objects or situations (curiosity). In short, our robots would have personalities, just like you and I do.

And just like you or me, they would not have had any hand in choosing those traits. Even if the robots learn over their lifetime and adapt to the various scenarios they encounter, all this learning is also physically embodied in the configuration of their circuits at the moment they are faced with a decision. The sense of fatalism that this realization engenders is aptly summarized by prominent free will skeptic Sam Harris:

> Take a moment to think about the context in which your next decision will occur: You did not pick your parents or the time and place of your birth. You didn't choose your gender or most of your life experiences. You had no control whatsoever over your genome or the development of your brain. And now your brain is making choices on the basis of preferences and beliefs that have been hammered into it over a lifetime—by your genes, your physical development since the moment you were conceived, and the interactions you have had with other people, events, and ideas. Where is the freedom in this? Yes, you are free to do what you want even now. But where did your desires come from?[1]

The essence of the problem was captured by the famously pessimistic (or some might say realistic) philosopher Arthur Schopenhauer, who said, "A man can do what he will, but not will as he will."[2] Even if we are making choices right now, those choices are not free from all kinds of prior causes or influences, over which we had no control.

1. Sam Harris, *Free will* (New York: Free Press, 2012), 44.
2. Arthur Schopenhauer, *Essay on the freedom of the will* (New York: Dover, 1960), 6.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

The Machine

As a neuroscientist, this kind of existential worry is an occupational hazard. But it gets worse. The more we learn about the mechanisms of perception and cognition and, in particular, of decision making and action selection, the more *mechanistic* it all seems and the less there seems to be for *the mind* to do. How can we even think that *we* are making choices at all, when we can see that the process is the result of just a bunch of gears turning in the machine? What reason is there to think that an entity is in charge?

And, thanks to modern technology, we can actually see the figurative gears turning. Using a variety of neuroimaging tools in humans and animals to track the activity of different neural circuits or brain areas, it is possible to tease out the types of information they carry and the cognitive operations they perform as the organisms or individuals make decisions or select actions. We can, for example, distinguish patterns of neural activity that correlate with (and seem to internally "represent") the accumulation of evidence about something in the world, the degree of certainty attached to some signal, the confidence level in a belief, the adoption of a new goal, the rewards associated with a positive outcome, the learning that happens in response to such rewards, the emotional signals that accompany decision making, the gradual formation of habits, the real-time switch from habitual to goal-directed or exploratory behavior as circumstances change, and on and on. We can see the thinking happening.

We can even, in some circumstances, predict an incipient action before the individual performs it. There are many experimental setups using rodents or monkeys where researchers can track patterns of brain activity, observe a threshold being approached that will result in an action, and even predict (not with complete accuracy but significantly better than chance) what action it will be—whether a rat will turn left or right in a maze, for example.

In humans there is a famous example where an action was not only predicted ahead of time but also before the subject even became consciously aware of having chosen to do it. In these experiments, performed

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

by Benjamin Libet and colleagues in the 1980s, subjects had to randomly decide to move their fingers while watching a clock and while their brainwaves were being recorded by an electroencephalograph. The striking result: the onset of brain activity leading to a movement preceded the reported timing of the conscious awareness of the intention to act by several hundred milliseconds.

Although not relevant to truly deliberative decisions, these findings can still shake your faith in your conscious mind really being in control of your actions. Is the rest of the brain just flattering us, making us feel that we're in charge, like a wily civil servant expertly managing his elected boss?

If pulling back the curtain to expose the neural machinery of decision making at work were not enough of a threat to our egos (in both senses of that word), it is also possible to intervene in the machine—to drive patterns of neural activity from the outside—and *cause* the individual to behave in certain ways.

Famous experiments carried out in the 1940s by neurologist Wilder Penfield and his colleagues in human subjects undergoing brain surgery (who were awake and aware throughout the procedure) showed that stimulating different parts of the cerebral cortex with electrodes could produce all kinds of sensations, emotions, urges, memories, or movements of various parts of the body (see Figure 1.2). This work contributed greatly to the mapping of functions across the brain and reinforced the view of a complex electrical machine *producing* the contents of the mind, rather than being controlled by that mental content.

Similar experiments are possible in animals, but, as in humans, they're a bit crude. Just poking an electrode into a part of the brain and zapping it activates all the neurons in that area in a nonspecific fashion. The brain then attempts to make sense of that mini-explosion of activation, but this process is very different from how neural signaling normally happens. Indeed, within any little chunk of brain, there are hundreds of different types of nerve cells connected in intricate microcircuits designed to carry out diverse sorts of computations. Just blasting them all at once is thus not hugely informative about how these computations mediate cognitive operations.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

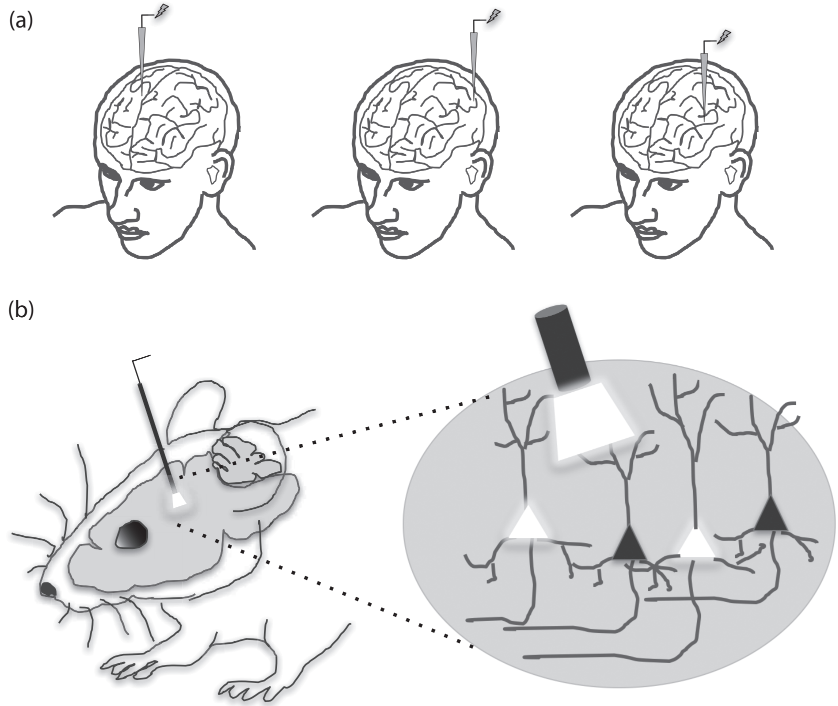

FIGURE 1.2. Brain stimulation. **(a)** Direct stimulation of brain areas in awake subjects can lead to involuntary movements (motor cortex, left), sensory percepts (visual cortex, middle), or even activation of memories (temporal lobe, right), depending on the area stimulated. **(b)** Optogenetic techniques in rodents allow much more specific activation of genetically targeted subsets of neurons, providing a powerful platform to dissect the functions of neural circuits in awake, behaving animals.

Seventy years after Penfield's experiments, the study of the neural systems that control behavior in animals was revolutionized by the invention of molecular tools that allow researchers to drive the activity of very specific subsets of neurons in an animal's brain while observing its behavior in real time. Like most techniques in molecular biology, this one—called optogenetics—borrows from nature. It uses a protein made by blue-green algae that sits in the membrane of the cell and responds to light by opening up a channel through which electrically charged atoms (or ions) can pass. That protein is related to ones that we

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

use in our eyes to detect light, and it is exactly the opening of ion channels in the membrane that drives nerve cells to "fire" or send a sharp electrical signal.

Researchers including Karl Deisseroth, Edward Boyden, and others realized that if they cloned the algal gene that encodes this protein and transferred it to mammalian neurons, they could effectively turn "on" the neurons with exquisite temporal precision by shining a blue light on them. Hooking the piece of DNA that codes for this light-responsive channel protein (called channel-Rhodopsin) to the DNA codes that regulate the expression of all kinds of different genes in the mouse brain enabled them to generate lines of transgenic mice expressing channel-Rhodopsin in extremely specific subsets of neurons in different brain regions.

Shining a light on the relevant bit of the brain—accomplished by threading a minute fiber optic cable through the skull—allows researchers to activate just that specific subset of neurons within the circuit and study the effects on behavior of the animal. Using this technique, specific sets of neurons were identified that, when activated, drive all kinds of behaviors—from general locomotion to more subtle motor actions like reaching or grasping, from aggression to mating, from freezing in fright to lunging attacks on prey that are not present, from eating to sleeping to looking after pups, and on and on.

But this research reaches far beyond directly activating particular actions from the animal's repertoire of behavior. It has made it possible to dissect the cognitive machinery involved in choosing among actions, weighing options, signaling rewards and punishments, judging the reliability of sensory information, assigning a level of certainty or confidence to a decision, using past memories to guide current actions, and selecting one option while inhibiting every other possibility. It is even possible, as my colleague Tomás Ryan and others have done, to implant false memories in an animal's brain that will influence its future behavior. This is not just remote control of what the animal is doing: it is control of what the animal *is thinking*.

It's hard not to look at this growing body of work and see only the machine at work. Driving this circuit or that one either directly causes an action or influences the cognitive operations that the animal—mouse

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

or human or anything else—uses to decide between actions. If we were dissecting a robot in this way we would apply engineering approaches to understand the kinds of information being processed, the control mechanisms configured into the different circuits, and the computations that lead to one output or another. There does not seem to be any need for something like *a mind* in that discussion. There is no real need for *life*, for that matter.

If the circuits just work on physical principles, then who cares what the patterns of activity *mean*? Why does it matter what the mental content associated with a particular pattern of neural activity is, if it is solely the physical configuration of the circuitry that is going to determine what happens next? We may have set out, as neuroscientists, to explain how the workings of the brain generate or realize psychological phenomena, but we are in danger of explaining those phenomena away.

It's All Just Physics at the End of the Day

If the neuroscientists have it bad, pity the poor physicists, whose existential angst must run much deeper. Where neuroscientists can at least hold onto the view that the circuits in the brain are doing things (whether "*you*" are or not), some physicists claim that even that functionality is an illusion. After all, the brain is made of molecules and atoms that must obey the laws of physics, just like the molecules and atoms in any other bit of matter.

These small bits of matter are pushed and pulled by all the forces acting on them—gravity, electromagnetism, the so-called strong and weak nuclear forces that hold atoms together—and where each atom goes is fully determined by the way those interactions play out. These processes are no doubt complicated, as they would be in any system with so many atoms simultaneously acting on each other, and in practice how the system will evolve is unpredictable—but it is still all driven by the physics. Even at the lower levels of subatomic particles, how the system evolves is captured by the equations of quantum mechanics in a way that many would argue theoretically leaves no room for any other causes to be at play.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

So, then, what does it matter what you are thinking? You cannot push the atoms in your brain around with a thought. You cannot override the fundamental laws of physics or exert some ghostly control over the basic constituents of matter. According to this view, the very idea of mental causation—of the content of your thoughts and beliefs and desires mattering in some way—is a naive superstition, a conceptual hangover inherited from philosophers like the famous dualist Rene Descartes.

Here is the late Stephen Hawking on the subject: "Biological processes are governed by the laws of physics and chemistry and therefore are as determined as the orbits of the planets. Recent experiments in neuroscience support the view that it is our physical brain, following the known laws of science, that determines our actions and not some agency that exists outside those laws . . . so it seems that we are no more than biological machines and that free will is just an illusion."[3] Brian Greene, another well-known physicist and author, agrees: "Free will is the sensation of making a choice. The sensation is real, but the choice seems illusory. Laws of physics determine the future."[4]

There are two main flavors of this kind of physical determinism. In the first, the low-level laws of physics rule completely: every aspect of the way the universe and everything in it evolves is fully determined by how these interactions play out. There is no room for any other force and, in particular, no role for any kind of randomness or indeterminacy. This model can be summed up as follows:

$$\text{current state} + \text{laws of physics} \rightarrow \text{next state}$$

The consequences of this view are stark. If you keep on working through from one state to the next, you quickly realize that the current state predicts not just one step ahead but also two, or three, or actually an infinite number. And you can work backward just as easily as forward. If this is really the whole picture, then the entire history of the

3. Stephen Hawking and Leonard Mlodinow, *The grand design* (New York: Bantam Books, 2010), 32.

4. Brian Greene (@bgreene) on Twitter, June 5, 2013, https://twitter.com/bgreene/status/342376183519916033?lang=en.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

universe up until now and for the rest of time was *pre*determined from shortly after the Big Bang. Indeed, our conception of time as having a direction goes out the window. The whole universe, over all time, is simply given, as a block: there is no real difference between the past and the future. There are no possibilities—only what has happened and what will happen. This view is known as *hard determinism*.

The implied softer version differs in allowing some randomness or indeterminacy to exist. It holds that the future is *not* fully predetermined by the current state (and certainly not by the initial state of the universe). Here, the past and the future are very different: the past is fixed while the future is a branching web of possibilities, only one line of which will be realized at any choice point.

However, even though the branch that is taken is not predetermined in this model, it is still decided by the low-level interactions of all the atoms and molecules. It is just that some of those interactions are a bit random. You might sum up this view like this:

current state + laws of physics + randomness → next state

The debate over whether there really is any true randomness in physical events has been raging since the days of Einstein and Bohr. When you get down to the quantum level of subatomic particles, weird things happen, and even though the weirdness can be fully accounted for in the equations that physicists use, allowing them to make exquisitely precise predictions, there is no consensus at all about what these equations imply about the fundamental nature of reality.

We'll return to this topic of randomness later. For now, what are the implications of this softer version of determinism? It is often summed up with the pithy line: "every event has a cause." This doesn't seem to align with the idea of random events happening, which would seem not to have a cause, by definition. What this statement really seems to imply is that everything that happens—at a system level—is caused by the interactions of particles at the lowest level, even if some randomness is at play there.

Yet, that view seems to be just as problematic for the idea of organisms like us being in charge of anything that happens. The future may

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

not be written, but if what happens is still decided by how the physical forces play out at the minutest scale of matter, there doesn't seem to be much scope for us to be in control. Even neuroscientist Patrick Haggard, a leader in the study of volition, agrees: "As a neuroscientist, you've got to be a determinist. There are physical laws, which the electrical and chemical events in the brain obey. Under identical circumstances, you couldn't have done otherwise; there's no 'I' which can say 'I want to do otherwise.'"[5]

In hard determinism, there are no causes. The universe just inexorably unfolds according to the laws of physics. If nothing could ever be or have been different, then you cannot point to one thing being a certain way and say it caused something else. The concept just doesn't apply. In *soft determinism*, there are causes—some things could be different, depending on how that little bit of randomness plays out—but all the causes are located at the lowest levels. That lowest level is deemed to be the bedrock of reality.

Some physicists, like Sean Carroll or Sabine Hossenfelder, may be magnanimous enough to allow that descriptions at higher levels of organization are "useful ways of talking about" complicated systems. We can productively do chemistry or biology or psychology with theories and methods that remain at those higher levels. But Carroll maintains that the real truth—the whole truth—resides at the lowest level, with the fundamental physical interactions of the smallest particles. If you had a complete accounting of what is going on down there, then you would not need any other information to fully predict what the system will do: everything happening at the higher levels simply derives or emerges from the low-level dynamics. Every other description is just a kind of coarse-grained picture, a *simplification* or statistical averaging that allows our puny minds to grasp how various systems—like cells or brains or minds—behave, despite all the underlying complexity.

5. Patrick Haggard, Neuroscience, free will and determinism: "I'm just a machine," interview by Tom Chivers, *The Telegraph,* October 12, 2010, https://www.telegraph.co.uk/news/science/8058541/Neuroscience-free-will-and-determinism-Im-just-a-machine.html.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Given the phenomenal successes of modern physics in confirming the predictions of quantum mechanics with eye-watering precision, it is not surprising that the focus has been on continuing to develop and test such theories while not worrying too much about what they mean for the nature of reality. The admonition to "Shut up and calculate!" by quantum physicist David Mermin is effectively the motto of the field. Let the philosophers worry about what it all means, especially for metaphysical concepts like free will.

The Blame Game

Philosophers, for their part, have been debating the implications of deterministic theories of the physical universe for free will for thousands of years, at least as long ago as Democritus and Epicurus in ancient Greece. That these debates continue today with unabated fervor tells you that they have not yet resolved the issue.

In fairness, free will is a uniquely vexing problem. The phenomenon we are trying to explain—our own experience of having the power to make choices—seems inherently at odds with what we know about how everything else works in the universe. The scientific rejection of the idea of an immaterial soul or spirit that is somehow pulling the strings has left us scrambling to explain instead how the machine could pull its own strings. And the progress of physics into the wonderful weirdness of the quantum realm has only deepened the mysteries of what the machine and the world around it are made of in the first place.

But if philosophy can be excused for not having provided an answer, one might at least have hoped for some consensus on what is the right question. The popular framing, "Do we have free will?" is undermined in an obvious way by a lack of agreed-on definitions. If you define the capacity of free will as being able to make decisions in a way that is necessarily *free from every prior cause*, then you have set an unattainable standard, one that could only be met by supernatural means. Alternatively, if your criterion is merely that a person is doing things *based on causes internal to his or her physical self*, then you have not met the chal-

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

lenge of physical determinism but merely sidestepped it with appeals to complexity and unpredictability.

Less obviously, the question "Do we have free will?" is more deeply undermined by a lack of clarity of the terms "we" and "have." We cannot profitably approach the question of whether *you* have free will until we have answered the much more fundamental question, "What kind of thing are you?" The contrasting criteria cited earlier are founded on differing conceptions of the nature of the self, where the philosophical footing is equally treacherous. Without a shared understanding of what everyone is talking about, it's not surprising that the debate seems to go round and round interminably.

Another barrier to a clear explication of the arguments around whether free will exists is that they are often approached from the direction of their *consequences* for our positions on moral responsibility. If people are not really in control of their actions—if we are nothing more than physical automata, mounting a wonderfully sophisticated but ultimately empty simulacrum of free will—then how can we be worthy of praise or blame? How can we defend judgment or punishment? The stakes here could not be higher. The idea of moral responsibility is the foundation not only of our legal systems but also of all our social interactions. We are constantly thinking about what we should or shouldn't do in any given circumstance and probably spend even more time thinking about what other people should or shouldn't do (or should or shouldn't have done).

But tying the discussion of free will to the issue of moral responsibility muddies the waters. Questions of moral responsibility are crucially important, of course, but they are confounded by all kinds of additional issues: the nature and origins of our moral sensibilities, the evolution of moral norms, the legal philosophies underpinning our justice systems, and the complex and innumerable pragmatic decisions that societies and individuals have to make to keep our collective existence stable. Asking what kind of free will *we want* that will let us maintain our positions on moral responsibility can become almost a theological exercise in motivated reasoning. It means we are looking for a palatable answer

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

instead of trying to understand what really *is*. It is coming at the question from the wrong end, picking an answer we like and seeing what edifice of arguments we need to build to support it. Instead, I would like to know what kind of free will *we actually have*.

Back to the Start

It's fashionable these days to claim that "free will is an illusion!": either it does not exist at all, or it is really not what we think it is. I am not willing to give up on it so easily. In this book I argue that we really *are* agents. We make decisions, we choose, we act—we are causal forces in the universe. These are the fundamental truths of our existence and absolutely the most basic phenomenology of our lives. If science seems to be suggesting otherwise, the correct response is not to throw our hands up and say, "Well, I guess everything we thought about our own existence is a laughable delusion." It is to accept instead that there is a deep mystery to be solved and to realize that we may need to question the philosophical bedrock of our scientific approach if we are to reconcile the clear existence of choice with the apparent determinism of the physical universe.

But if we want to solve this mystery, humans are the absolute worst place to start. It is a truism in biology to say that nothing makes sense except in the light of evolution—and this is surely true of agency. Instead of trying to understand it in its most complex form, I go back to its beginnings and ask how it emerged, what the earliest building blocks were, and what the basic concepts should be. How can we think about things like purpose and value and meaning without sinking into mysticism or vague metaphor? I argue that we can do so by locating these concepts in simpler creatures and then following how they were elaborated over the course of evolution, increasing in complexity and sophistication as certain branches of life developed ever-greater autonomy and self-directedness.

Indeed, before tackling the question of free will in humans, we have a much more fundamental problem to solve. How can any organism be said to *do anything*? Most things in the universe don't make choices.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Most things—like rocks or atoms or planets—don't do anything at all, in fact. Things happen to them, or near them, or in them, but they are not capable of action. But *you* are. You are the type of thing that can take action, that can make decisions, that can be a causal force in the world: you are an agent. And humans are not unique in this capacity. All living things have some degree of agency. That is their defining characteristic, what sets them apart from the mostly lifeless, passive universe. Living beings are autonomous entities, imbued with purpose and able to act on their own terms, not yoked to every cause in their environment but causes in their own right.

To understand how this could be, we have to go right back to the beginning, to the very origins of life itself (see Figure 1.3). This is the trajectory that I sketch out in this book.

From the chemistry of rocks and hydrothermal vents—the chemistry of the evolving planet itself—life emerged as systems of interacting molecules, interlocked in dynamic patterns that became self-sustaining. The ones that most robustly maintained their own dynamic organization persisted, replicated, evolved. They became enclosed in a membrane—a tiny subworld unto themselves—exchanging matter and energy with their environment while protecting an internal economy and reconfiguring their own metabolism to adapt to changing conditions. They became autonomous entities, causally sheltered from the thermodynamic storm outside and selected to persist.

A new trick was invented: action, the ability to move or affect things out in the environment. Information became a valuable commodity, and mechanisms evolved to gather it from the environment. With that came the crude beginnings of value and meaning. Movement toward or away from various things out in the world became good or bad for the persistence of the organism. These responses were selected for and became wired into the biochemical circuitry of simple creatures.

As multicellular creatures evolved, a class of cells—neurons—emerged that specialized in transmitting and processing information. Initial circuits acted as internal control systems, designed to coordinate the various muscles or other moving parts of the multicellular animal, defining a repertoire of useful actions. At the same time, neurons coupled

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

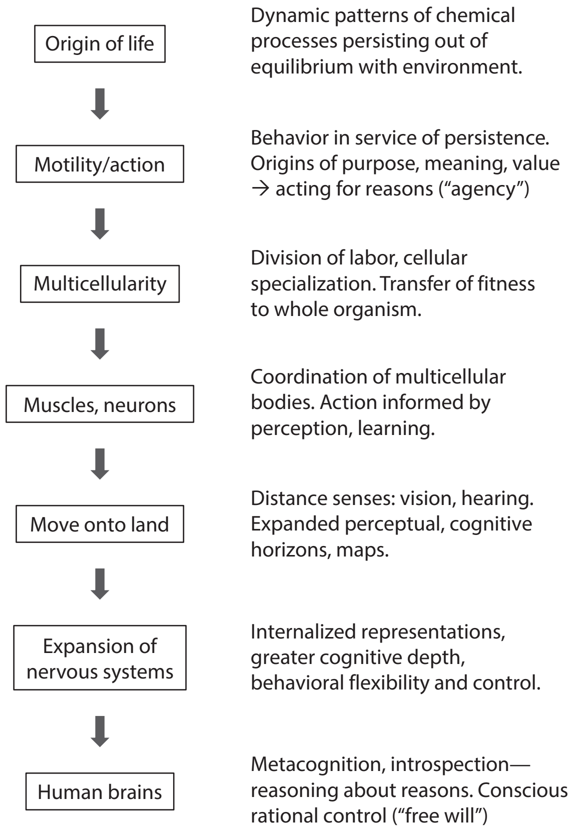

FIGURE 1.3. The evolution of agency and free will. The major stages of evolution of perception, cognition, and behavioral control.

various sensory signals to specific actions in this repertoire, hardwiring adaptive instincts for approach or avoidance.

With the elaboration of the nervous system, this kind of pragmatic meaning eventually led to semantic representations. Perception and action were decoupled by layers of intervening cells. Instead of being acted on singly and immediately like a reflex, multiple sensory signals could be simultaneously conveyed to central processing regions and operated on in a common space. Circuits were built that integrated, amplified, compared, filtered, and otherwise processed those signals

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

to extract information about what was out in the world and what that meant for the organism. More and more abstract concepts were extracted—not just about things but also types of things and types of relations between them. Creatures capable of understanding emerged.

Meaning became the driving force behind the choice of action by the organism. That choice is real: the fundamental indeterminacy in the universe means the future is not written. The low-level forces of physics by themselves do not determine the next state of a complex system. In most instances, even the details of the patterns of neural activity do not actually matter and are filtered out in transmission. What matters is *what they mean*—how they are interpreted by the criteria established in the physical configuration of the system. Animals were now doing things for reasons.

That causal power does not come for free: it is packed into the organism through evolution, through development, and through learning. It is encoded in the genome by the actions of natural selection. And it is embodied in the physical structure of the nervous system in the strength of neuronal connections that express functional criteria in relation to a hierarchy of aims of the organism. There is nothing here that violates the laws of physics; it just demands a wider concept of causation over longer timeframes and an understanding that the dynamic organization of a system, which encodes meaning, can constrain and direct the dynamics of its component parts.

And yes, your actions are at any given moment constrained by all those prior causes. Yet you could just as well say, more positively, that they are *informed by* prior experience. That is precisely the property that sets life apart from other types of matter: living things literally *incorporate* their history into their own physical structure to inform future action. For those who would argue this impinges on the freedom of the self to decide at any moment, I counter that it is this very process that enables the self to exist at all. There is no self in a given moment: the self is defined by persistence over time.

And though you are configured in a certain way that reflects all this history, you are not hardwired. We humans have the remarkable capacity for introspection and metacognition. We can inspect our own

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

programming, treating goals and beliefs and desires as cognitive objects that can be recognized and manipulated. We can think about our own thoughts, reason about our own reasons, and communicate with each other through a shared language. We can access the machine code running in our brains by translating high-level abstract concepts into causally efficacious patterns of neural activity. This gives a physical basis for how decisions are made in real time, not just as the outcome of complex physical interactions but also *for consciously accessible reasons*, and it provides a firm footing for the otherwise troublesome concept of mental causation.

So, if you want to know what kind of thing you are, you are the kind of thing that can decide. Not just a collection of atoms pushed around by the laws of physics. Not a complex automaton whose movements are determined by the patterns of electrical activity zipping through its circuits. And not an NPC, unknowingly driven by its programming. You are a new type of thing in the universe—a self, a causal agent. In the game of your life, you are Player One.

What follows is thus a full-throated defense of the idea of free will. Despite many claims to the contrary, the latest science—whether physics, genetics, neuroscience, or psychology—does not in fact imply that we have no choice or control over our actions. It's true that we are learning more and more about the mechanisms underlying our cognition and behavior—from neural systems and circuits down to the level of cells and molecules or even atomic physics. But even though our cognitive systems have a physical instantiation, their workings cannot be *reduced to* this level. We are not a collection of *mere mechanisms*. As we will see, the nervous system runs on meaning.

The fact that our capacities for cognitive control are grounded in definable biological systems does, however, have important implications for issues of moral and legal responsibility, though these are notably more subtle than the typical absolutist framing. I return to consider these and related issues in the final chapter.

Along the way, I offer a perspective on life that centers agency as its defining characteristic. What distinguishes living organisms is that *they do things, for reasons.* They behave in a truly purposeful manner. This is

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

not an illusion or just a convenient way of talking or thinking about them: it's the right way of thinking about them. Causation does not all bubble up from the bottom, nor is it all instantaneous. The way things are organized can and does govern the way complex systems behave. Living organisms accumulate causal power by coming to embody aspects of their history in their own structure, either through evolution or over the course of their individual lifetimes. The story of agency is thus the story of life itself, and that is where we begin, in chapter two.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

INDEX

Note: page numbers in italics refer to figures.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

For general queries, contact webmaster@press.princeton.edu