THE GREAT COURSES®

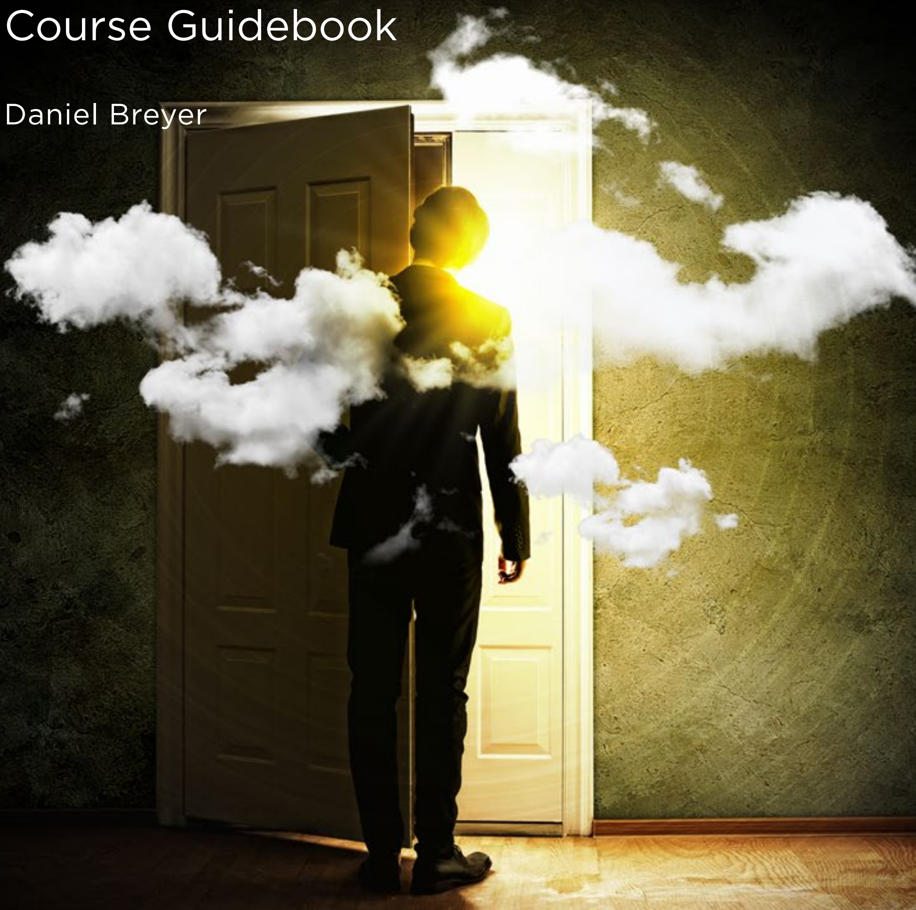
Topic
Philosophy, Religion
& Intellectual History

Subtopic
Understanding
the Mind

The Power of Thought Experiments

Course Guidebook

Daniel Breyer



THE GREAT COURSES®

LEADERSHIP

President & CEO	PAUL SUIJK
Chief Financial Officer	BRUCE G. WILLIS
Chief Marketing Officer	CALE PRITCHETT
SVP, Marketing	JOSEPH PECKL
SVP, Content Development	JASON SMIGEL
VP, Content Production	KEVIN BARNHILL
VP, Marketing	EMILY COOPER
VP, Customer Engagement	KONSTANTINE GELFOND
VP, Technology Services	MARK LEONARD
VP, Content Strategy	KEVIN MANZEL
VP, People	AUDREY WILLIAMS
General Counsel	DEBRA STORMS
Sr. Director, Content Operations	GAIL GLEESON
Director, Talent Acquisition	WILLIAM SCHMIDT
Director, Creative	OCTAVIA VANNALL

PRODUCTION

Studio Operations Manager	JIM M. ALLEN
Video Production Director	ROBERTO DE MORAES
Technical Engineering Manager	SAL RODRIGUEZ
Quality Assurance Supervisor	JAMIE MCCOMBER
Sr. Postproduction Manager	PETER DWYER
Sr. Manager, Production	RIMA KHALEK
Executive Producer	JAY TATE
Producer	KATY MERRY HANNAH
Sr. Content Developer	WILLIAM WOJTACH
Assistant Content Developer	MARK HARDY
Image Rights Analyst	KATE MANKOWSKI
Postproduction Manager	OWEN YOUNG
Sr. Editor	MILES MCNAMEE
Sr. Audio Engineer	ED SALTZMAN
Audio Engineer	GORDON HALL IV
Director	JOHN NAPOLITANO
Camera Operators	GEORGE BOLDEN, RICK FLOWE
Production Assistants	LAKE MANNIKKO, PAUL SHEEHAN, KELLY TAGLIAFERRI, VALERIE WELCH

EDITORIAL & DESIGN

Director	FARHAD HOSSAIN
Sr. Managing Editor	BLAKELY SWAIN
Writer/Editor	JENNIFER ROSENBERG
Editorial Associates	MOLLY LEVY, MARGI WILHELM
Research Associate	L. VIOLA KOZAK
Graphics Manager	JAMES NIDEL
Sr. Graphic Artist	BRIAN SCHUMACHER
Graphic Artist	DANIEL RODRIGUEZ
Graphics Coordinator	KATE STEINBAUER
Graphic Designer	TIM OLABI

Daniel Breyer

Daniel Breyer is a Professor of Philosophy at Illinois State University, where he also directs the Religious Studies program. He earned a PhD in Philosophy from Fordham University. He has received many accolades, including the Outstanding University Teacher Award, the highest instructional honor at Illinois State University. His research often explores thought experiments, and his articles have appeared in journals such as *Philosophy and Phenomenological Research*, *Pacific Philosophical Quarterly*, and the *Journal of Buddhist Ethics*.

Table of Contents

The Power of Thought Experiments

Scope

Imagine yourself running alongside a beam of light. What would that be like? What would it look like? And what could you learn about the world by imagining it?

Or imagine that you have the chance to plug into a machine that would allow you to experience anything you've ever dreamed of. The only catch is that you can never unplug. Would you do it? And what would your choice tell you about yourself—and what you value?

The physicist Albert Einstein thought that imagining what it would be like to chase a beam of light revealed the "germ" of special relativity, and the philosopher Robert Nozick thought that imagining his "experience machine" demonstrated that we care about more than experiences; we care about actually doing things. Were they right?

These products of the imagination are thought experiments. At their core, thought experiments are what-if situations, or hypothetical scenarios, that help us think about ourselves and the world, what we care about and what's important, how things are, how things might be, and how things must be. Thought experiments are potent intellectual tools, and in this course, we'll explore what makes them so powerful.

Our exploration will take us around the world, through history, and across disciplines. We'll consider landmark thought experiments by scientists like Galileo Galilei, Erwin Schrödinger, and, yes, Albert Einstein. And we'll wonder how imaginary cases can tell us something about how the world really is. What's the relationship between a thought experiment and an actual experiment? And when is a thought experiment better?

Thought experiments don't just help scientists think about the world. Philosophers, economists, theologians, mathematicians, and other thinkers, past and present, have used thought experiments to explore issues as wide-ranging as rational choice, social cooperation, moral responsibility, the nature of knowledge, the good life, identity, and the importance of subjective experience, to name just a few of the topics we'll consider in the course.

As we explore thought experiments from around the world, we'll put thinkers from different traditions, both past and present, in a dynamic conversation, not only with each other, but also with us. We won't just learn about thought experiments in this course; we'll question them, challenge them, and see whether they have the power to change our minds and inform our lives. Sometimes, we'll even imagine our own.

So, are you ready to explore the power of thought experiments along with great thinkers like the Islamic polymath Ibn Sīnā, the British epistemologist Miranda Fricker, and the Confucian philosopher Xunzi? The journey will be fascinating and fun. And don't worry. To appreciate the power of thought experiments, you won't need any specialized training. All you'll need is a good dose of intellectual curiosity and a sincere desire to discover the truth. In the laboratory of the mind, after all, there are no lab coats required.

1

HOW THOUGHT EXPERIMENTS WORK

This course will cover some of the most important thought experiments that have been developed. By the end, you will be able to see for yourself the power they have to expand our minds. You'll also explore how they are used in different intellectual traditions, where they have been crucial in advancing the boundaries of knowledge. In this lecture, you'll learn some general features of thought experiments and consider how they work.

Transformative Experiences

Imagine that you're on vacation in Romania's central region of Transylvania. One night, someone offers you the chance to become a vampire. The process is quick and painless, but it's not without consequences. It's irreversible, and as a vampire you will have a very different life from the one you have now. You'll have amazing powers and abilities along with intense sensory experiences that will reveal a new world to you.

Now, suppose you've been interested in vampires since you can remember, and this opportunity is really attractive to you. Moreover, many of your friends have become vampires themselves. Those friends love it and have no regrets. They tell you how meaningful their lives are now, and how much more deeply they understand the world—and themselves.

But they also say that they can't really explain what it's like to be a vampire to someone who's still just a human. You'd have to be a vampire to really get it. If you pass up this opportunity, you'll probably never get the chance again. It's now or never.

The philosopher L. A. Paul uses this thought experiment about vampires to explore what she calls transformative experience. There are two important features of this kind of experience. They are epistemically transformative in the sense that they give us a new knowledge of what something is like. They also change who we are—our personal way of experiencing the world—in deep and far-reaching ways. They alter our priorities, preferences, and even self-conception. Being a vampire is so different from being human that it would give you a new perspective on the world—one unavailable to mere humans. It would also change what you care about, and even who you are, in fundamental ways.

> A good thought experiment presents a transformative experience that is vivid, gripping, and powerful.

So, would you become a vampire? Should you? How can you make a truly informed decision? If you can't know what it's like to be a vampire without becoming a vampire first, then you can't weigh the options; you can't rationally make that choice because you don't have the information you need.

L. A. Paul's thought experiment is vivid because we can imagine ourselves in the scenario. It's gripping because the scenario dares us to consider the choice of a lifetime. It's powerful because it presents us with interesting problems for rational choice. As such, it's the perfect introduction to how thought experiments work. The fact that it's fanciful, because vampires don't actually exist, is beside the point.

As Paul notes, some of our most significant life choices are transformative. Think about the choice to have a child and become a parent. That choice is epistemically transformative. Becoming a parent involves coming to know what it's like to be attached to your child and have a relationship with them, but you can't know what that's like until you do it. It is also personally transformative. Once people become parents, their preferences, priorities, and self-conceptions change—sometimes dramatically—along with other, more mundane aspects of their lives.

Thought experiments like Paul's are fascinating and fun, but they also do serious work. They can open us to new ideas, demonstrate that old ideas are wrong, and help us investigate the world and ourselves in ways other intellectual tools simply can't. In short, thought experiments are powerful tools to stretch our imagination.

Learning through Scenarios

If you've studied the philosopher Immanuel Kant, you may be familiar with this thought experiment:

Imagine that you're at home alone one evening, and you hear an urgent knock at the door. You hesitate at first, but the knocking turns into frantic pounding, and you hear your friend's voice shouting to let them in. You rush to the door, and your friend comes in, clearly in distress. They tell you someone has been chasing them through the streets, and they don't know why. They're clearly frightened, and you know they wouldn't lie about something like this. You tell them to hide in your basement. Then you hear another knock at the door. The stranger standing there doesn't look right to you. They ask if your friend has come by recently, and they claim to have an urgent message they need to deliver to them.

> Thought experiments allow us to do things in what the philosopher James Robert Brown calls "the laboratory of the mind" that we couldn't possibly do otherwise.

Would you tell the stranger the truth, or would you lie and tell them that you haven't seen your friend? What should you do?

Kant's view is that you should not lie to the stranger, but his judgment here hasn't fit well with most people's intuitions about the case. Most people say that they wouldn't tell the stranger the truth, and moreover that they shouldn't tell the stranger the truth. What do you think?

This thought experiment is a potential counterexample to Kant's view that lying is always wrong, no matter what, and it puts pressure on Kant's general view that there are moral absolutes—moral rules that we can never, under any circumstances, violate.

Whether that's right or wrong is up for debate. For our purposes, what's important is the power of this thought experiment to consider what's truly important, and what we have reason to think is morally upright in a situation like this. It's an extreme scenario, but that's part of its power. The choice we have to make is stark. The stakes are high. And we learn something about ourselves and the world by thinking through these cases—that's their real power.

Let's note some general features of thought experiments and consider how they work. We use them to make judgments not only about what would be the case if an imaginary scenario were real, but also about what ought to be the case. Should you choose to become a vampire? Should you lie to the person at the door?

Thought experiments also challenge us to see problems that we might not have noticed otherwise. For someone who thinks it's always wrong to lie, the stranger at the door presents an unappreciated problem to solve. For someone who's never really thought about transformative experiences and the challenges we face making choices about them, Paul's vampire scenario does much the same thing. How can we make rational choices when transformative experiences are involved?

Philosophers Helen De Cruz and Johan De Smedt suggest that thought experiments have a lot in common with scenario building. We imagine all kinds of scenarios, big and small, all the time. When we do this, we don't just think about what we should do; we learn things about ourselves and the world. The hope is that they provide us with new knowledge, new insights, new ways of understanding things. What we gain insight into depends on the thought experiment. We might discover something about what's necessary, what must be the case, or what's possible.

Intuition Pumps

Thought experiments can lead us astray if we're not careful. At least some thought experiments are what the philosopher Daniel Dennett calls intuition pumps. According to Dennett:

> Intuition pumps are cunningly designed to focus the reader's attention on "the important" features, and to deflect the reader from bogging down in hard-to-follow details. There is nothing wrong with this in principle. Indeed, one of philosophy's highest callings is finding ways of helping people see the forest and not just the trees. But intuition pumps are often abused, though seldom deliberately.

Daniel Dennett

The tricky thing about thought experiments is that their success often (though not always) depends on our intuitive judgments about cases—our intuitions. What's an intuition? We'll be using that word a lot, so we need to say something about what it means. We might characterize intuitions in a number of ways, but in general, they are judgments about how things seem to us and about what seems to us to be true. The hope is that it's more than a mere psychological gut reaction, but sometimes that's all it is.

Our initial judgments about things often go wrong due to confusion, counterintuitive facts about the world and ourselves, and other irrelevant factors. Our intuitions about thought experiments can also vary depending on how the thought experiments are described. It might be that small changes in wording have big effects on our intuitive judgments, or that adding or subtracting details can change what we think about a case. While we want to be precise, we also should watch out for it.

Another problem with thought experiments as intuition pumps is that the context in which we consider them can alter our judgments. For example, if a case is presented in a different order, that can influence our opinions.

This doesn't mean that our intuitive responses to thought experiments are inherently unreliable, but it does highlight that our judgments are fallible and that we have to be careful when we draw concrete conclusions from thought experiments.

Shaun Nichols

With this in mind, it's worth making a distinction between two kinds of intuitions. Following the philosophers Shaun Nichols and Tamler Sommers, we can distinguish between starting intuitions and considered intuitions.

Starting intuitions represent our initial judgments or reactions to thought experiments. They just feel right intellectually, and we might not be able to provide any further reasons or justification for them. In contrast, considered intuitions represent our initial judgments that have withstood critical scrutiny; they have further reasons backing them up. The power of good thought experiments isn't primarily that they elicit starting intuitions; it's that they help us secure considered intuitions.

Philosopher of science John Norton thinks that thought experiments are rearrangements of information we already have—a way of reframing or repackaging an argument that's already implicitly there. Norton thinks this is the case for scientific thought experiments in particular, because they can't possibly give us new empirical data about the world, since we conduct thought experiments in the laboratory of the mind. And yet Norton acknowledges that thought experiments have played an important role in the development and history of science—especially physics.

Often, the very best thought experiments will function as both arguments and invitations, drawing us into them. How could you rationally choose to become a vampire? Should you lie to the stranger at the door? To appreciate the power of thought experiments, you don't need specialized training. All you need is intellectual curiosity and an interest in pursuing the truth.

Reading

- Breyer, Daniel. *World Philosophy: 50 Puzzles, Paradoxes, and Thought Experiments*. Routledge, 2023.
- De Cruz, Helen. *Philosophy Illustrated: Forty-Two Thought Experiments to Broaden Your Mind*. Oxford University Press, 2022.
- Dennett, Daniel. *Intuition Pumps and Other Tools for Thinking*. W. W. Norton & Company, 2013.
- Knobe, Joshua. "Intentional Action and Side Effects in Ordinary Language." *Analysis* 63 (2003): 190–194.
- Paul, L. A. *Transformative Experience*. Oxford University Press, 2014.
- Sorensen, Roy. *Thought Experiments*. Reprint ed. Oxford University Press, 1998.
- Tittle, Peg. *What If . . .: Collected Thought Experiments in Philosophy*. Routledge, 2004.

2

SAVING OTHERS OR LETTING THEM DIE

Thought experiments in the realm of ethics can be very powerful, not so much because they give us definitive answers, but because they invite us to think in new and often challenging ways about fundamental issues we care a lot about. This lecture looks at several scenarios that will allow you to work through questions about our obligations to others, the moral difference between killing and letting die, and the factors that make an action wrong.

The Prevention Principle

Imagine you're walking through a park, and you see a child struggling in a pond. You know that the pond is shallow and that you could easily wade in and rescue the child. You'd ruin your clothes and other property, like your watch, phone, and wallet. Saving the child seems like the right thing to do. It also seems like it would be wrong not to save the child—like you're morally obligated to do it. If we agree that not saving the child would be morally wrong, what does that tell us?

Philosopher Peter Singer, in his influential 1972 article "Famine, Affluence, and Morality," argues that our intuition that we should save the drowning child is driven by an underlying moral principle: "If it is in our power to prevent something very bad from happening, without thereby sacrificing anything else morally significant, we ought, morally, to do it." Let's call this the prevention principle.

Keeping this in mind, we all know that there are many people, in our communities and around the world, who are suffering from lack of food, shelter, and medical care. They're in distress, just like the drowning child. And we wouldn't have to sacrifice anything morally significant to donate our time, money, or talents to effective charities that could save the lives of people who would, without that help, continue to suffer and even die.

Therefore, it seems like we don't just have an obligation to save the drowning child; we have an obligation to help anyone in serious need if it's in our power to help them without sacrificing anything morally significant.

If this is right, then morality is much more demanding than most of us think. We might think that it would be generous to donate money or time to a good cause, but we also think that it would be morally permissible for us not to donate. Singer is suggesting that we're wrong about this. His view is that we're morally obligated to help those in serious need, as long as doing so wouldn't force us to sacrifice anything morally significant.

This is why Singer's thought experiment is so powerful and important. It's supposed to illustrate a clear and intuitive application of the prevention principle. But it doesn't just do that; it seems to provide intuitive support for it as well. We're supposed to recognize that there's no morally relevant difference between saving a drowning child and helping those in serious need, because the underlying moral principle—the prevention principle—is the same.

It looks like endorsing Singer's prevention principle puts pressure on our traditional ways of thinking about moral obligation. Because, if he's right, then it seems like we'd be morally wrong to buy extravagant gifts, wear expensive clothes, go on luxury vacations, and eat at fancy restaurants. Many people agree with Singer, at least to some degree, but many more want to resist what he has to say. After all, if we agree with him, then we also probably have to agree that we're moral failures!

Saving Many Drowning Children

In his book *Why It's Okay to Want to Be Rich*, Jason Brennan acknowledges that Singer's drowning-child case is powerful, and he agrees that we should save the child. He questions whether the case provides support for the prevention principle and whether that principle is what's driving our intuitions about the case. To make his point, Brennan asks us to imagine a slightly different case involving many drowning children.

Jason Brennan

Suppose you're walking in a park, and you notice a swimming pool "filled with drowning children." There are no other adults around, and you can't call for help because your phone battery died. You do your best to save them, sacrificing everything you have with you and giving as much of your time as you can, but no matter how many children you save, more keep falling in. Do you think you're morally obligated to keep saving these children?

Brennan thinks that your intuitive response to this thought experiment is especially telling because this case is "more analogous to the real world" than Singer's original case. After all, no matter what you do, there's always going to be someone else who needs help.

If our response to Singer's original case is driven by the prevention principle, our view about Brennan's scenario should be that we're morally obligated to keep saving children until we have to sacrifice something morally significant. But Brennan bets that "you probably think that you have to save the first child, but at some point, you can move on and live your life, even though children will die."

If that was your reaction, then when and why would it be morally permissible to stop saving children? It seems like it would be morally permissible to stop saving children to attend to your own child's needs, to rest, to eat, or to seek medical care. Could you stop forever, knowing that there's a pool nearby that draws children who can't swim into it? Doubtful. You'd probably have to keep returning to the pool each day to help, for at least some time, or work with the community to find a solution.

Brennan thinks his alternate scenario undermines Singer's prevention principle, but it actually highlights what's so powerful about it. What's so intuitively powerful is the idea that we have serious moral obligations to others, but those obligations aren't unbounded—they're limited by other morally relevant concerns.

Singer is asking us to consider helping a child in need over, for instance, having a fancy dinner. This provides some compelling intuitive support for the prevention principle while also nicely illustrating its practical application. But if that's right, then most of us aren't doing as much as we should to help those in serious need.

Killing Someone versus Letting Someone Die

Perhaps you're thinking that there's a morally important difference between not saving the child and actually killing the child yourself. Singer himself rejects the distinction that our moral obligation to prevent someone's death isn't as strong as our moral obligation not to cause it.

Imagine that you're on a crowded city street, and someone suddenly stabs a man next to you. Unable to move from the shock of it all, you watch as the injured man crumples to the ground and dies. It seems like you let the man die—that you had a moral obligation to save him and didn't. But it also seems like letting the man die is not nearly as wrong as stabbing him. The person who stabbed him did something morally wrong. You didn't do anything. But if you didn't really do anything wrong, then did you really have a moral obligation to save him?

In a 1975 article, the philosopher James Rachels explores the distinction between killing and letting die while examining euthanasia. Passive euthanasia is a matter of allowing a patient to die by withholding treatment. Active euthanasia, by contrast, is a matter of doing something to bring about a patient's death—perhaps giving them a life-ending substance. Rachels argues that the distinction between killing and letting die has, in itself, no moral significance. To make his argument, he has us consider a thought experiment involving two cases.

In the first case, we have Smith. If his six-year-old cousin dies, he's going to inherit a lot of money. One night, while Smith's cousin is taking a bath, Smith sneaks into the bathroom and drowns the child, making it look like it was an accident.

In the second case, we have Jones. He will also get a big inheritance if his young cousin dies, and Jones has the same plan as Smith. But right as Jones goes into the bathroom, his cousin slips, hits his head on the tub, and falls face down in the water. Jones waits in the bathroom, ready to push the child's head back under if necessary. But his cousin dies all on his own, accidentally, as Jones watches and does nothing.

Is the difference between what Smith and Jones did morally significant? Rachels suggests that it's not. The badness and wrongness of what they did seem the same, and they both seem equally responsible for the death of the child.

But the philosopher Scott Hill argues that the cases aren't actually identical, because Smith and Jones have different abilities. Jones has the ability to let his cousin die and the ability to kill him, whereas Smith only has the ability to kill his cousin—he can't just let him die. That difference might not seem like a big deal, but a third case shows why it matters.

Now we have Adams. Like Smith and Jones, he will also get a big inheritance if his cousin dies. He has the same plan. But right as Adams goes into the bathroom, his cousin slips, hits his head on the tub, and falls face down in the water. But instead of just standing by, Adams insists on killing him.

Hill argues that we shouldn't compare the cases of Smith and Jones, because they don't have the same opportunities and abilities. We should be comparing Jones and Adams because both could let the child die or kill him. Hill's view is that what Adams does is morally worse than what Jones does. If Hill is right, then we have reason to think that killing someone is worse than letting them die, because the moral difference between what Adams does and what Jones does "is very easily explained by the hypothesis that killing is worse than letting die."

Hill thinks that our intuitions about the Smith and Jones cases are unreliable. Why? What Smith and Jones do is awful, and Hill thinks "it feels disturbingly cold and pedantic to try to measure out and quantify the precise difference between the two acts." It's also hard for us to keep track of what exactly we're evaluating when we consider the cases. Are we focusing on the wrongness or badness of what they do? Or are we making judgments about the moral status of the people involved? Are we focusing on their moral character, motivations, or intentions? It's easy to run all these judgments together when we consider the cases. Both Smith and Jones seem like they're nasty people who are responsible for the death of a child, and so it's tempting "to gloss over the [morally relevant] difference between" their respective acts—of actively killing and passively letting die.

This is part of what's tricky about using thought experiments as arguments and relying on our intuitions about them as evidence. We have to make sure we're responding to the same salient features of the cases we're considering, appealing to the same underlying principles, ensuring that any cases we compare are relevantly equivalent, and in general doing our best to discover the truth, or at least uncover what we think is true.

Reading

- Brennan, Jason. *Why It's Okay to Want to Be Rich*. Routledge, 2020.
- Hill, Scott. "Murdering an Accident Victim: A New Objection to the Bare-Difference Argument." *Australasian Journal of Philosophy* 96, no. 4 (2018): 767–778.
- Rachels, James. "Active and Passive Euthanasia." *New England Journal of Medicine* 292 (1975): 78–86.
- Singer, Peter. "Famine, Affluence, and Morality." *Philosophy & Public Affairs* 1, no. 3 (1972): 229–243.

Table of Contents

3

WHAT THE TROLLEY PROBLEM REVEALS ABOUT YOU

You've probably encountered the trolley scenario at some point in your lifetime—in a book, in a movie, or even in a class discussion. In this lecture, you're going to explore it in some detail. As you'll see, it's a thorny problem, driven by a series of powerful thought experiments that will force you to question what you're doing when you make moral judgments.

Trolley and Transplant Scenarios

Imagine that you're driving a trolley. As you drive around a bend, you see five people working on the track up ahead, and they have no way of getting off the track. You attempt to stop the trolley, but the brakes fail. You notice a spur of track leading off to the right. If you take the turn, you'll avoid hitting the five workers, but you'll hit one other worker. You understand that anyone you hit will almost certainly die.

For now, we'll focus on whether it would be morally permissible for you to turn the trolley. You probably think it would be, and a lot of people have the intuition that it's also the right thing to do in this scenario. But we haven't quite seen the problem yet. This thought experiment is only part of the story.

Consider a different scenario: You're a transplant surgeon. You have five patients who need transplants by the end of the day, or they'll die. Two of them need a lung, two need a kidney, and one needs a heart. If you can find organs for them, they'll all live; if you can't, they'll all die. Time is running out, but then a young man, in perfect health, comes in for his annual checkup, and you realize that he's the perfect donor. You ask him if he'd be willing to donate his organs, but he's not willing to volunteer. Would it be morally permissible for you to operate on him anyway?

In the trolley driver case, it seemed like it was morally permissible to kill one person to save five, but it doesn't seem so in the transplant case. So, what's the difference between the cases?

The trolley problem was first introduced by Philippa Foot, and it was carefully developed by Judith Jarvis Thomson. According to Thomson, Foot has a simple solution to the problem: In the transplant case, your choice, as a surgeon, is whether to operate on the young man or not. If you operate, you kill one person to save five. If you don't operate, you don't kill anybody, but you do let five people die. And killing someone is much worse than letting someone die. In fact, it's plausibly so much worse that killing even one person is still worse than letting five people die. And this is why it's not morally permissible for you to operate in the transplant case.

In the trolley driver case, your choice is whether to turn the trolley or not. If you turn, you don't let someone die; you kill someone. And if you don't turn, you don't let five people die; you kill five people. Of course, killing five people is much worse than killing one person. And this is why it's morally permissible—maybe even morally obligatory—for you to turn the trolley to the right.

The Bystander at the Switch

Thomson thinks that solution is too easy. To show why, she has us consider another scenario, where you're not driving the trolley. Instead, you're on a walk near the trolley tracks, and you see things unfold. You see that the trolley driver is slumped over, unconscious. It's clear that if the trolley continues on, the five workers will die. But you happen to know that just behind you, within your reach, is a switch that can send the trolley off the track to the right, where you can see there's another worker who will get hit by the trolley and surely die. What should you do?

Thomson thinks that it would be morally permissible for you to throw the switch. If you agree, then Philippa Foot's solution to the trolley problem won't work. That's because your choice is now exactly like the surgeon's choice. Why is it morally permissible to throw the switch but morally wrong to operate on the perfectly healthy man?

Maybe the cases seem different because a surgeon should care for all their patients, and it seems like this one is disregarding the well-being of one patient in favor of the well-being of five others. Maybe people, as patients, should not be treated in certain ways.

The Bystander on the Bridge

Let's consider another of Thomson's cases. In this scenario, you're on a walk near the trolley tracks, but this time you're walking over the tracks on a footbridge. To your left, you see the trolley barreling down the tracks, and to your right, you can see the five workers. You realize that if you could drop something down on the tracks, you could derail the trolley and save the five people. You notice a very large man leaning over the rails, gawking at the trolley. Would it be morally permissible for you to push the man over the rail onto the tracks to derail the train?

This case is analogous to the transplant case and should elicit the same intuitions. If we think it's morally wrong to operate on the perfectly healthy man, then it should also seem morally wrong to push a hapless bystander off the footbridge to derail the trolley.

Now we have a more precise way to frame the trolley problem: Why is it morally permissible to divert the trolley away from five people into just one person by throwing a switch, but morally wrong to divert the trolley away from five people by pushing someone onto the trolley's path?

Perhaps the issue is that you intend to kill the man on the footbridge, whereas you don't intend to kill the person standing on the track. But while you can, of course, foresee that pushing the man onto the tracks will likely kill him, that doesn't have to be something you intend. If the man were to survive, you could be happy about that. So, we can't insist that the morally relevant difference between the two scenarios switch lies in your intentions.

Using People as the Means to an End

In the bridge scenario, what matters isn't so much whether you foresee or intend a certain outcome; what matters is how you treat the large man. You can't just use people as mere means to your own ends, can you? The idea of respecting people as ends in themselves is important to Kantian ethics, which traces its roots back to the 18th-century German philosopher Immanuel Kant. In the bridge scenario, it seems like you're not respecting the man as a person but rather you're treating him merely as means. Let's explore this idea further.

You don't use someone merely as a means whenever you make use of them. For instance, you can make use of the employees at the store to buy groceries; that's morally permissible. They have consented to work there and to help you as a customer. What's not morally permissible is using them in ways that circumvent their agency. We have to treat people in ways that respect their autonomy.

Immanuel Kant

With this in mind, it's clear that the man in the bridge scenario hasn't consented to being pushed onto the trolley tracks. Without that consent, he's been used merely as means. But the philosopher Pauline Kleingeld points out that whether someone uses someone else depends on how they reason, not necessarily on whether the other person consents. According to her, this insight helps us solve the trolley problem, because we've identified a morally relevant difference between the switch and the bridge scenarios: You necessarily consider using the man as a means when you push him off the bridge, and that's what makes that action morally wrong, but you don't use the person on the track merely as a means because what you consider using is the switch—and a switch is not a person.

Pauline Kleingeld

This might seem like a tortured and insincere way of reasoning about the situation. If that's the case, then we're back where we started. We might conclude that there's no moral difference between the two scenarios after all. But if they are indeed morally equivalent, then we have to revise our initial judgments, so that our judgment about throwing the switch aligns with our judgment about pushing the man. If one is morally permissible, then so is the other. And the same goes if one is morally wrong. Which one is it?

If this is where we land, then the power of this thought experiment, at least in part, is that it forces us to reconsider our moral judgments and clarify the principles and values that drive those judgments. This in turn helps us work through our own views about what's right and wrong and what matters to us, morally speaking.

Pushing Our “Moral Buttons”

There’s another problem that these thought experiments raise. As moral psychologist Joshua Greene puts it, this problem is to explain “why, as a matter of psychological fact, people tend to approve of trading one life to save several lives in some cases but not others.”

What Greene suggests, based on various studies, is that several factors play a big role in driving our intuitions. The first of these is personal harm. For instance, people tend to say that it’s wrong to push someone and cause them harm, and people tend to disapprove of personal force more than impersonal force, like throwing a switch. Is this a moral intuition or something else?

Greene also notes that we tend to disapprove of intentionally harming someone as a means to an end. In fact, we tend not to worry so much about personal force unless we think it’s intentional. What also matters to us is the negative emotional response we have to intentional personal harm. When all of this combines, Greene suggests that it pushes our “moral buttons.” So, when our moral buttons get pushed, do our moral judgments get clouded?

We might think that the fundamental tension in the trolley problem is between consequentialist and deontological considerations. From the consequentialist perspective, it seems like it’s always best to save more lives, and that seems to be how we think about the bystander at the switch. From the deontological perspective, however, it seems like there are just some things we should not do no matter what the consequences are, and that seems to be how we think about the bystander on the bridge.

The psychologists Fiery Cushman and Molly Crockett suggest that our tendency to go back and forth between these two perspectives is grounded in two different “learning systems” in our cognitive architecture. Model-free learning mechanisms emphasize past experience and derive value from that. They tell us that pushing the man off the bridge is associated with negative outcomes like bad feelings and social disapproval. Model-based learning mechanisms emphasize causal relations in the world. They tell us that saving five lives would have great real-world outcomes. What’s clear is the power of the trolley problem and the thought experiments that make it up challenge us to think carefully, not only about what we value, but also about why.

Reading

- Foot, Philippa. “The Problem of Abortion and the Doctrine of the Double Effect.” In *Virtues and Vices and Other Essays in Moral Philosophy*. Blackwell Publishers, 1978.
- Greene, Joshua. *Moral Tribes: Emotion, Reason, and the Gap between Us and Them*. New York: Penguin Press, 2013.
- Kleingeld, Pauline. “A Kantian Solution to the Trolley Problem.” *Oxford Studies in Normative Ethics* 10 (2020): 204–228.
- Thomson, Judith Jarvis. “The Trolley Problem.” *The Yale Law Journal* 94, no. 6 (1985): 1395–1415.

4

SUPPOSE YOU'RE IMPARTIAL; SUPPOSE YOU CARE

Would you want to live in a world void of close relationships if it meant everyone was cared for equally and was slightly happier than people in this world? This lecture looks at this question through the lenses of Confucian, Mohist, and Buddhist traditions. It offers several thought experiments to challenge your intuitions around what it means to be partial and impartial.

Privileged Relationships

Imagine that you live in a world called Equim. Everyone there is naturally impartial, "equally fond of everyone." Because of this, they don't have friends as we think of them, since friendships are privileged relationships. In fact, if someone had to choose between saving their own child and saving someone else's child, they would make every attempt to ensure an impartial decision, perhaps even by flipping a coin to determine which child they'd save. The overall happiness in Equim is slightly greater than in our world. And this is the case for the whole society as well as each individual—including you.

What if you could take a pill that would make your desires and interests just like those of the people of Equim, and you'd be slightly happier than you are now? Would you take the pill? The philosopher Thomas Donaldson uses this thought experiment to explore morally privileged relationships. We normally think that favoritism of certain kinds is morally justified. Donaldson thinks that most of us would choose not to take the pill.

If you agree, then that raises another question: Is taking the pill the right choice? Wouldn't caring for everyone equally be morally preferable to caring for some more than others? How could the fact that we just happen to know certain people—our friends and family—provide a moral reason for caring about them more than people we don't know? And isn't the fact that a world like Equim is happier morally significant? Should we take that into account?

Imagine that two people are drowning in front of you, but you can save only one. One of these people is a stranger, and the other is someone you love. You'd probably save the person you love. But why?

The philosopher Charles Fried thinks it would be absurd to require you to treat both of these people equally. And one explanation for why this would be absurd is that, as long as you aren't the captain of a ship or a lifeguard or someone who holds some official position like that, you're permitted, in such situations, to prefer your loved one over the complete stranger. Fried's suggestion might strike you as reasonable, but the philosopher Bernard Williams thinks the fact that this person is a loved one is the only explanation we need. What Fried suggests provides too much of an explanation.

Charles Fried

Confucian Tradition

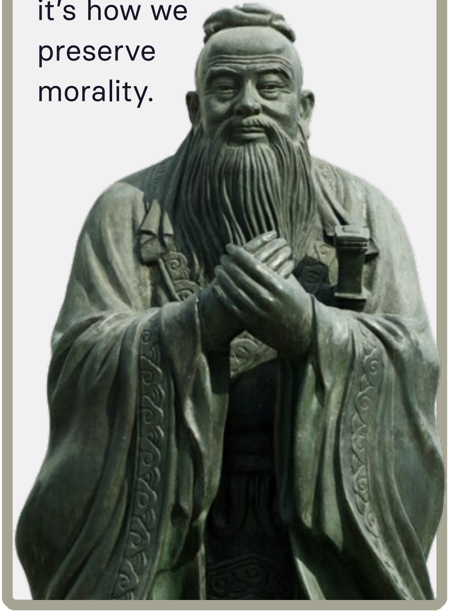
With that in mind, consider another thought experiment that we find in Kongzi's *Analects*. (Kongzi might be better known to you as Confucius.) One day, the duke of She told Kongzi about a man who was known as Upright Gong. Gong was "upright" because he had turned in his own father for stealing sheep. Kongzi told the duke that among his people, fathers covered for their sons, and vice versa. "Uprightness," Kongzi told the duke, "is to be found in this."

This scenario isn't just about doing the right thing. It's about the Confucian virtue of filial piety, which is the loyalty, respect, deference, and care we owe to our parents, to senior family members, and to no one else. So, a good Confucian son wouldn't privilege someone else over his father.

But it's not so simple. Kongzi, like you, thinks that it's morally important to be honest and that stealing from others is wrong. The thought experiment highlights a tension in what it means for someone to be "upright." As the philosophers Yong Huang and Liang Tao point out, the duke of She seems to think of it in broadly social terms, whereas Kongzi seems to emphasize the relationship between uprightness and filial piety.

But it's more than that. Kongzi also says that filial piety is the root of humanity. Our relationship with our parents provides us with our moral foundation—like the roots of a tree are its foundation. Against an impartial ethic, the Confucian tradition defends the view that morality is partial. Instead of caring for everyone equally, we should practice "love with distinctions." If this line of thinking is right, it gives us a way to respond to Donaldson's Equim challenge. We shouldn't take the pill because we'd destroy the very roots of morality!

In Confucian tradition, failing to turn a loved one in for stealing isn't privileging a personal relationship over morality; it's how we preserve morality.



The philosopher Yong Huang suggests that we interpret Kongzi's thought experiment along these lines. Huang suggests that filial piety, and love with distinctions more generally, isn't about unthinking loyalty to those closest to us; it's about caring for their well-being. Kongzi isn't saying that covering up for our family and friends is what's right. It's that we find uprightness in covering up for them because that provides the opportunity for moral improvement.

Think again about whether you should take that pill and become completely impartial like the inhabitants of Equim. Why shouldn't you take it? The Confucian answer is that becoming completely impartial would uproot morality. But you might not like that lesson. What about justice for the victim? Even if the victim is a stranger, shouldn't we also care about them?

Mohist Tradition

Mozi

In the Chinese philosophical tradition, Confucian philosophers defend partiality and love with distinctions, but Mohist philosophers defend impartiality and universal love (or impartial care).

The Mohist philosophical tradition traces its roots back to Mozi, who flourished in the 5th century BCE, roughly a century after Kongzi. Mozi would take Donaldson's pill, because Equim is just the sort of utopia he hoped to establish in the world. And Mozi, along with later Mohists, argued that you value impartiality and universal love, too.

To make his case, Mozi gave this thought experiment: Imagine that you're going away on a long and dangerous journey. You're not sure when you'll be home or even if you'll make it back. While you're gone, who would you trust to care for your family, someone impartial or someone partial? Mozi thinks that partial people are selfish and that they don't care for anyone but themselves. He insists that impartial people aren't selfish.

Now, you surely agree with Mozi that you'd want to entrust your family to someone who would actually care for them, rather than someone who would ignore their needs altogether. But the way Mozi characterizes partiality is odd. It's so closely linked with selfishness that the two are indistinguishable, but of course this is problematic, because partiality, especially under the Confucian tradition, isn't the same as selfishness or radical egoism.

Although it's true that a Confucian friend would care more for her own family than for yours, that's very different from not caring about your family at all. The philosopher Bryan Van Norden characterizes the impartial Mohist as someone "who cares for your family as much as he cares for everyone else" but the partial Confucian as someone "who cares more for your family than he does for strangers."

Now it seems like it would be better to entrust your family's welfare to your Confucian friend, rather than to your Mohist friend. If your Mohist friend is going to care for your family and for a stranger in the same way, maybe that's not good enough. Your Confucian friend will extend their concern for their own family to yours.

Can you see how malleable intuitions are? This tells us that we have to be very careful about how we characterize both the partial Confucian and the impartial Mohist. If impartiality requires that we care about strangers as much as we care about your families, we might be thinking that we'd have to level everyone down to the status of strangers. But we don't have to think of impartiality like that.

Buddhist Tradition

Patrul Rinpoche

The philosopher Emily McRae suggests that we can think of impartiality as leveling everyone up, rather than down: We bring a stranger up to the level of a loved one, making everyone equal. McRae points out that this is how the Tibetan Buddhist thinkers like Patrul Rinpoche characterize impartiality. They believe in reincarnation, and according to them, we should see others as having been our mothers at one time.

But we don't need to believe that, because the point, as McRae puts it, "is to elicit easy and natural feelings of love and gratitude." We can do that by engaging in a contemplative thought experiment. First, we see everyone as though they are our mothers. Next, we imagine just how much we love our mothers. Then, we recognize that we want to repay our mothers with kindness. (If your mother didn't care for you like this, for whatever reason, imagine instead the person who did.) And what we get is a new kind of impartiality.

As McRae notes, this kind of impartiality is grounded in partiality, but it also uproots it. We need that close mother-child relationship to start leveling everyone up, but once we do that, we no longer have partiality—we have impartial care, not a watered-down and toothless version of universal love, but a full-blooded concern for others that's identical to the concern we have for those we love most deeply.

Now go back to Mozi's caretaker thought experiment. Who would you entrust your loved ones with? The Confucian who remains partial but who has extended their concern for their family to yours, or the impartial Mohist, or perhaps we should now say Buddhist, who has leveled everyone up to the status of the person they care for most in the world?

However you answer that question, we still need to address our original question: Would you choose to become like the inhabitants of Equim? Should you? The Confucian answer is that you shouldn't do it, because they are impartial, and partial relationships lie at the very root of morality. The Mohist answer is that you should do it. We should eliminate selfish partiality and care for everyone equally. There's no moral justification for privileging some relationships over others.

But there's also a kind of in-between response. We can understand what it means to be impartial in more than one way. It might mean that we treat everyone equally in a way that levels everyone down to the same generic moral status. Or it might mean that we treat everyone equally in a way that levels everyone up to the same special moral status. The first way seems cold and weird. The second way seems far more attractive. Sure, it would be demanding to care about everyone as much as we care about those we love the most, but Mozi doesn't think it would be impossible.

If we go with the second option, leveling up, then impartial care requires that we start with partiality. With that in mind, we might view the people of Equim as lacking something crucial. Can they level anyone up if they don't first have partial relationships, like the relationship between a mother and a child, that teach them what it means to care about others deeply? The people of Equim don't seem to care about anyone at all. They seem to care about impartiality, not people. If two of your children were drowning, could you flip a coin to decide which one to save, or would you dive in as a desperate attempt to do the impossible? Your answer tells you whether you'd take that pill.

Reading

- Donaldson, Thomas. "Morally Privileged Relationships." *Journal of Value Inquiry* 24 (1990): 1–15.
- Fraser, Chris. *The Essential Mozi*. Oxford: Oxford University Press, 2020.
- Huang, Yong. "Why an Upright Son Does Not Disclose His Father Stealing a Sheep: A Neglected Aspect of the Confucian Conception of Filial Piety." *Asian Studies* 5, no. 1 (2017): 15–45.
- McRae, Emily. "Equanimity and Intimacy: A Buddhist-Feminist Approach to the Elimination of Bias." *Sophia* 52 (2013): 447–462.
- Williams, Bernard. "Persons, Character, and Morality." In *Moral Luck: Philosophical Papers 1973–1980*. Cambridge University Press, 1981.

Table of Contents

5

UNMASKING THE HIDDEN PITFALLS OF TESTIMONY

Epistemology is the philosophical study of knowledge. This lecture considers several thought experiments in social epistemology, which focuses on knowledge in a social setting rather than just on individual knowers. Topics discussed include different views people might take in a disagreement, testimonial knowledge, and epistemic injustices.

Equal-Weight versus Steadfast Views

Imagine that you and your friend Jack are out to dinner, and it's time to pay the bill. Over the years, you've had frequent dinners together, and when the bill comes, you always calculate what you each owe in your heads and then compare answers. Whenever you've disagreed about what you owe, you've each been wrong about half the time.

Tonight, you and Jack disagree on how to split the check. Should you remain steadfast in your belief, staying confident in your calculations, or should you reduce your confidence and maybe even do the math again? In this case, you and Jack are epistemic peers. He's no better or worse at math than you. Otherwise, it might seem reasonable to defer to his judgment in this case. You and he have exactly the same evidence, and you are equally capable, at least with respect to math. What should you and Jack do now that you're faced with this disagreement?

Someone is your epistemic superior when they're in a better epistemic position than you. That could be because they, for example, have a better memory than you, are better at performing a task, or have better evidence than you. Experts often count as our epistemic superiors, at least in their area of expertise.

This sort of thought experiment about peer disagreement comes from the philosopher David Christensen. We'll call it the split-check case. Let's assume that there are no calculators or phones around and that you'll only muddy things by asking the waiter to help.

The power of this thought experiment is that it's supposed to provide intuitive support for what we might call the equal-weight view, which holds that what you should do is give each claim equal weight and either suspend judgment or revise your beliefs in a way that makes you significantly less confident in them.

The split-check case provides a scenario where neither you nor Jack can tell whether you've made a mistake, and so neither of you can justifiably claim that the other has made an error. It's equally likely that you're both wrong. One of you might be right, but even so, that person isn't justified in being confident that they're right—at least, not until you can settle the matter.

If you think the mere fact that Jack disagrees is not enough to lower your confidence, then you endorse the steadfast view. This view says that your evidence that the check should be split a certain way has stronger grounds for you than Jack's evidence. This is because you have access to your line of reasoning and the evidence you used to reach your conclusion but not for how Jack reached his conclusion. The philosopher Linda Zagzebski suggests that this means you have grounds to trust yourself that you don't have for trusting Jack. So, she suggests you can remain steadfast in your belief about how the check should be split.

The Justificationist View

Consider a variation on the split-check thought experiment, based on cases the philosopher Jennifer Lackey first introduced. Suppose you and Jack have met for coffee to discuss how many people you'll be inviting to a dinner party. You count yourself and your partner as two. You count Jack, his partner, and Alice as three. So, five people will be there.

When you say that, though, Jack disagrees with you. You confirm that you haven't missed anyone and restate the total of five. Jack again disagrees. You think he must be kidding, but he isn't. He says, "Two plus three doesn't equal five."

Here, you and Jack disagree about how many people are coming to the party and how to do some basic math. If you thought that the split-check case supported the view that you should at least lower your confidence about how the check should be split, do you have similar intuitions about this bad-at-math case?

Probably not. In the first case, it seemed equally likely that you or Jack made a mistake, but in this case, it's obvious who's made the mistake. Jack seems to demonstrate that he's bad at math. Now, in the split-check case, it seemed like you and Jack would have had to find some independent source of evidence to figure out which of you had gone wrong. But in this case, you have evidence about who's gone wrong baked into the case. You have reason to think you're right because Jack reasons in such a bad way. The very fact that he disagrees with you like this is evidence that he's the one who's made a mistake, not you.

In small cases of disagreement, the equal-weight view seems right to Lackey. She thinks the bad-at-math case supports a view that lies somewhere between the equal-weight view and the steadfast view. The view she endorses is the justificationist view, which says that when you're highly justified in believing something, like two plus three equals five, before discovering that someone disagrees with you, then there's no reason for you to reduce your confidence in what you believe. In those cases, we can use the disagreement itself as a reason to dismiss our peers.

The real power of these thought experiments, at their core, is that they help us identify various positions we might take up about disagreement, and figuring out how to handle it rationally is surely one of the biggest challenges we all face every day.

Testimonial Knowledge

Testimony is one of our most important sources of knowledge. Our beliefs about history, geography, science, philosophy, medicine—and even ourselves—are grounded in testimony. But it isn't always reliable. Sometimes you hear or read things that are not true. And it seems like one reason testimony isn't always reliable is that sometimes people tell you things they themselves don't really know. You can't learn something from someone else unless they have that knowledge to pass on to you. This is the idea behind the standard transmission model of testimonial knowledge.

But consider this thought experiment: Stella is a fourth-grade science teacher. She thinks it's her duty to teach her students the best science, based on what the scientific community accepts. And so she prepares all her lessons with the best science available in mind—not what her own personal beliefs are. So, when she teaches her students a unit on evolutionary biology, she tells them that present-day human beings evolved over hundreds of thousands of years from *Homo erectus*.

The thing is, she doesn't believe this. She is a young-Earth creationist. She believes God created the world 6,000 years ago, along with all species—human beings included—in their current form at that time. Stella doesn't tell her students about her personal beliefs. But since knowing something requires believing it, Stella doesn't know the facts about evolution that she's teaching her students.

You might wonder why knowledge requires belief. In some ways, this is a dogma of contemporary philosophy, and so you're not supposed to question it, but the idea here is supposed to be straightforward and obvious. With that in mind, can Stella's student learn facts about evolutionary biology from her? This is another thought experiment from Jennifer Lackey. The transmission model says that they can't. Lackey thinks Stella's students learn certain facts about the world—facts that Stella herself doesn't believe but that she is reliably capable of communicating. The students learn the scientific consensus because that's what Stella transmits. But if that's the case, then Stella does know what she tells her students, and it seems like the transmission model remains unscathed.

Of course, fourth graders also learn facts about the world from their teachers, not just scientific consensus, but think about students who learned that Pluto was the ninth planet. After 2006, the scientific community's understanding shifted, and Pluto is no longer counted as a planet. What those earlier students learned wasn't so much a fact about the world as it was the scientific consensus at the time.

Testimonial Injustice

What we want is credible testimony. But we want to be careful. When we make judgments about whether someone is credible, we should assign them the right amount of credibility that they deserve. If we give someone too much credibility, they get a credibility excess. If we give someone less credibility than they deserve, they get a credibility deficit. The philosopher Miranda Fricker argues that credibility deficits are especially worrisome, particularly when they're associated with prejudice.

Testimonial injustice happens when someone receives less credibility than they deserve due to identity prejudice.

Consider what happens in the trial of Tom Robinson in Harper Lee's novel *To Kill a Mockingbird*. Tom is a Black man who has been accused of sexually assaulting a White woman, Mayella Ewell. Evidence shows that Tom could not have been the person who committed the crime, but the all-White jury simply won't believe Tom's story. Because of their prejudice, they think Tom is lying, and they refuse to give his testimony the credence it deserves.

Fricker uses this trial to demonstrate what she calls testimonial injustice. Tom is treated harshly and unjustly by having his testimony dismissed based on negative stereotypes about him. He's harmed because his ability to communicate information, impart knowledge, and participate in the flow of information is undermined in a way that degrades him as a human being.

Philosopher José Medina points out that one of the reasons Tom receives a credibility deficit is because Mayella Ewell enjoys a credibility excess. As a young White woman in a racist, anti-Black society, her false testimony gets more credence than Tom's true testimony, and so her credibility excess crowds out what little credibility Tom might have otherwise had for the all-White jury. José Medina's idea here is that credibility isn't something that individuals have in isolation; credibility judgments are inherently relational.

Hermeneutical Injustice

Imagine that it's 1955, and Susan is working as a receptionist at an insurance company where all the agents are men. At work, they often leer at her, make lewd comments about how she looks, and sometimes even touch her. One time, her boss followed her home. She is scared and knows this is wrong, but she's not sure what to say about it or who to tell.

We know that Susan is being sexually harassed at her job. She's even being assaulted and stalked. But the legal concept of sexual harassment only started to emerge in the US in the 1970s. Susan doesn't know about this concept, and so she's having trouble identifying and communicating what's happening.

Susan is suffering an epistemic injustice that Miranda Fricker calls hermeneutical injustice. What that means is that Susan doesn't have access to the resources or concepts that would help her communicate what she knows. The reason society lacks these resources is due to identity prejudice against women.

The philosopher Rachel McKinnon notes that Fricker is far from the first thinker to highlight these sorts of issues. Many feminist thinkers of color raised similar issues much earlier, including bell hooks, Audre Lorde, and Linda Martín Alcoff. McKinnon's view is that, although Fricker's work is important and has now established the framework for thinking about epistemic injustice, it's important to note that who secures uptake of ideas is also a matter of epistemic justice. When feminist women of color argue for issues we'd clearly describe as epistemic injustice (in Fricker's terms), but that work only secures wide uptake when a white woman (like Fricker) articulates the concepts, then this is itself an instance of epistemic injustice.

What's especially powerful about their work is that they highlight aspects of social experience, knowledge, and the flow of information that otherwise remain hidden and elusive. They validate the experience of people who've suffered such injustices, and they provide a window into those experiences for those who haven't. They also provide the starting point for developing a framework for recognizing and dealing with epistemic injustice—a framework that provides us with new concepts to talk about an under-recognized aspect of disagreement and testimony in social life.

Reading

- Fricker, Miranda. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press, 2007.
- Lackey, Jennifer. *Learning from Words: Testimony as a Source of Knowledge*. Oxford University Press, 2008.
- ———. "What Should We Do When We Disagree?" *Oxford Studies in Epistemology* 3 (2008): 274–293.
- McCain, Kevin. *Epistemology: 50 Puzzles, Paradoxes, and Thought Experiments*. Routledge, 2021.
- McKinnon, Rachel. "Epistemic Injustice." *Philosophical Compass* 11, no. 8 (2016): 437–446.

Table of Contents

6

CAN YOU TIME-TRAVEL AND CHANGE THE PAST?

The grandfather paradox is a fascinating and alluring thought experiment, but it's also powerful, because working through it helps us think carefully about what time travel is, what it would mean to change the past, and—perhaps most surprisingly—what it means to be able to do something. This lecture examines the paradox and various ways thinkers have tried to solve it.

The Grandfather Paradox

Tim has greatly benefited from his paternal grandfather's wealth, which was accumulated mainly during World War II and its aftermath. But Tim also hates his grandfather and would like to kill him for what he did to make the family so wealthy. But Tim can't do that because his grandfather died before Tim was born.

Tim has been using his grandfather's wealth to build a time machine. During his first trip, he travels back to 1920, 10 years before his own father was born. He realizes that he can kill his grandfather and starts working on a plan. One day in 1921, when the conditions are right and he has the skills he needs, Tim's ready to kill his grandfather. Can he do it?

As philosopher David Lewis points out, Tim seems to be in the same position as another person named Tom. Tom has all the same skills as Tim. Tom, however, wants to kill Tim's grandfather's business partner. Can he do it?

We have no reason to doubt that Tom can kill the business partner. But, as Lewis points out, by any ordinary standards of ability, it seems, Tim can kill his grandfather, too. But the problem is that Tim can't kill his grandfather. Why not?

Tim's grandfather lived beyond 1921, and so to kill him would be to change the past, which is impossible. What's worse, if Tim kills his grandfather before he has children, then Tim is never born. If Tim is never born, then, he never travels back to 1920, and he never kills his grandfather. Tim can't kill his grandfather because it would be a self-defeating, logical impossibility for him to do so. It looks like we have a paradox—one commonly known as the grandfather paradox.

This is, at least potentially, a very powerful thought experiment. It might demonstrate that backward time travel isn't just something that's impossible given the laws of physics; it might show that backward time travel is logically and metaphysically impossible!

The Time-Discrepancy Paradox

Perhaps we can avoid the whole paradox by saying that when Tim travels back in time, he enters 1920 on an alternate timeline. Tim's existence on the original timeline doesn't depend on his existence in the new timeline.

Tim can kill "his" grandfather on this new timeline without terminating his own existence, because this new grandfather isn't really his original grandfather. And Tim isn't really changing the past at all. He's making things happen on a different timeline, not changing something that already happened on his original timeline. Let's call this the branching-timeline solution to the grandfather paradox. Interestingly, this gives us a way of making sense of time travel. But we haven't actually defined time travel. What does it mean to travel backward in time?

In the most basic sense, it means that Tim departs from one time and arrives at another. If he departs from the year 2025 and arrives in the year 1920, he travels back in time 105 years. Let's say it takes 10 minutes to travel those 105 years. In that case, it looks like time travel would require a discrepancy between time and time, with two events—in this case, Tim's departure and arrival—separated only by different amounts of time. How could 105 years and 10 minutes elapse in the same amount of time? This problem with making sense of time travel is what the philosopher Dennis Holt calls the time-discrepancy paradox.

But if time travel is about moving from one timeline to another, then backward time travel isn't merely a discrepancy between one time and another; it's a discrepancy between the time in one timeline and the time in another timeline. That's not obviously contradictory. The branching-timeline approach makes time travel a lot like moving from one place to another. And if those two places have different times, that's fine. We see that with the different time zones between, say, Illinois and Montana. The discrepancy isn't really between time and time; it's between the time in one place and the time in another. Illinois and Montana are spatially and temporarily related to each other.

As the philosopher Ryan Wasserman notes, what's distinctive about the branching-timeline approach to time travel, and its solution to the grandfather paradox, is that it denies that this is what it's like in time travel. When we move from one timeline to another, the times are no longer related to each other. This means that we can't say that 1920 on the new timeline is earlier than or later than 2025 on the original timeline; they're just not related that way, because they don't share a common temporal frame like Illinois and Montana do.

In effect, we're moving between parallel or alternative worlds. Tim doesn't kill his grandfather in Tim's original timeline; he kills a sort of counterpart to his grandfather, and so he doesn't undermine his own existence. He makes it so that his new-world counterpart—who might have also been named Tim—doesn't exist in this alternative timeline.

Wasserman's view is that the branching-timeline approach "turns out to be a theory of parallel universe travel, rather than time travel." He concludes the approach is irrelevant. It doesn't solve the grandfather paradox; it changes the subject. If Wasserman is right, then we're not only left without a solution to the grandfather paradox; we're also left without a clear understanding of time travel and a response to the time-discrepancy paradox.

Context-Sensitive Abilities

David Lewis thinks of time travel as real travel but not between parallel universes. To make sense of time travel, he distinguishes between external time, which is real time, and personal time. As Lewis puts it, personal time isn't really time, but it plays the role in a time traveler's life that time plays in the life of an ordinary person.

When we travel in time, then, there's a discrepancy between personal time and external time. Tim travels 105 years into the past in 10 minutes. It's paradoxical to say that 10 minutes takes 105 years to pass, or that 105 years goes by in 10 minutes—but it's not paradoxical to say that Tim's personal time, as measured by his watch and the growth of his hair, elapses in only 10 minutes as he travels back in real time 105 years.

Lewis's way of thinking about time travel, which is now the standard view, helps us avoid the time-discrepancy paradox—but we still have to deal with the grandfather paradox. Lewis has an influential response to that, too. He notes something about what it means to have an ability, or to be able to do something. When we say that Tim can or can't kill his grandfather, we have to be careful. Abilities are context-sensitive. As Lewis puts it, "What I can do, relative to one set of facts, I cannot do, relative to another."

If we're thinking about Tim's abilities relative to a certain narrow set of facts, he can obviously kill his grandfather. He has the skills and the opportunity. But that excludes the fact that Tim's grandfather was not killed in 1921, that he fathered a child later in life, a child who would later become Tim's father. If we consider Tim's abilities relative to this broader set of facts, then it's obvious that Tim can't kill his grandfather.

Lewis thinks that we can talk about Tim's abilities relative either to the narrow set or to the broader set of facts, but we can't waver and say in the same breath that Tim both can and can't kill his grandfather and then claim that this contradiction proves that time travel is impossible. If we're clearheaded and consistent, the grandfather paradox dissolves. It's worth noting that what Lewis says isn't just about Tim's abilities; it's a general point about what it means to say that anyone can or can't do something in the ordinary sense. It's also intuitive. It seems right to think about abilities relative to certain facts. It seems right to say that abilities are context-sensitive. Lewis's solution to the grandfather paradox also doesn't commit him to any specific view about the nature of time or the mechanics of time travel.

Failure as Proof of Inability

Kadri Vihvelin argues that Lewis gets things wrong. To make her case, she considers Suzy, also a time traveler. Suzy hates her own life so much that she decides that it would have been better never to have lived it at all. She intends to kill her infant self. When she gets the opportunity, she fires her gun but misses. She fires again, and the gun jams.

Vihvelin doesn't think we need to fill in the details, because we all know how it has to end: All of Suzy's efforts will fail, because Suzy can't actually kill baby Suzy without changing the past and eliminating her own existence. A paradox would result: Suzy both would and wouldn't kill baby Suzy.

Vihvelin thinks Lewis's solution to the grandfather paradox is untenable. She points out that no matter how hard Suzy tries, and no matter how hard Tim tries, neither of them can succeed. Vihvelin asks: Shouldn't we agree that if someone would fail to do something, no matter how hard or how many times they tried, then they cannot do it?

In the narrow sense, an ability is something you're able to do given facts about you. In the broad sense, an ability is something you're able to do given facts about you and facts about your environment. What matters here, Vihvelin argues, is the broad sense of ability. This is the sense of ability that matters for practical deliberation. If you're thinking about what to do, you make decisions based on what you believe are "live" options—things you could actually do. Otherwise, you're wasting your time.

According to Vihvelin, in no ordinary sense can either Suzy or Tim succeed. When we consider the normal facts relevant for determining whether someone has a wide ability or whether something is a live option for someone, it's clear that killing baby Suzy isn't a live option for Suzy and that Tim doesn't have the broad ability to kill his grandfather. Where does this leave us?

Vihvelin doesn't think it shows us that the grandfather paradox is sound. Her view is that time travel is possible, but traveling back in time doesn't give us the ability to undercut our existence. In Tim's case, the difference between him and Tom is environmental. Tom has the broad ability to kill Tim's grandfather's business partner. It was a live option for him. He could have succeeded. But Tim doesn't have the broad ability to kill his grandfather. It's not a live option for him. No matter how hard he might have tried, he couldn't have succeeded, because it's just as impossible for him to kill his target as it is for either him or Tom to, say, run faster than the speed of light.

Reading

- Lewis, David. "The Paradoxes of Time Travel." *American Philosophical Quarterly* 13 (1976): 145–52.
- Rea, Michael. *Metaphysics: The Basics*. 2nd ed. Routledge, 2020.
- Vihvelin, Kadri. "What Time Travelers Cannot Do." *Philosophical Studies* 81 (1996): 315–30.
- Wasserman, Ryan. *Paradoxes of Time Travel*. Oxford University Press, 2020.

Table of Contents

7

PARADOXES AS MENTAL WORKOUTS

Throughout the course, you'll return to thought experiments that drive you into paradoxical conclusions and leave your head spinning, so it's worth understanding what makes for a true paradox. Paradoxes can be incredible teaching tools and serve as some of the most powerful thought experiments. In this lecture, you'll work through Hilbert's hotel, the surprise-quiz paradox, and the paradox of the stone.

What Is a Paradox?

Michael Huemer

The philosopher R. M. Sainsbury notes that paradoxes are fun, but they're also bafflingly serious, because they also raise serious problems. According to Sainsbury, a paradox provides "an apparently unacceptable conclusion derived by apparently acceptable reasoning from apparently acceptable premises." Viewed this way, paradoxes are fundamentally arguments. However, we don't want to commit ourselves to the view that all thought experiments are reducible to arguments.

The philosopher Michael Huemer understands a paradox as "a situation in which we have seemingly compelling reasoning for a contradictory or otherwise absurd conclusion." Huemer's definition allows us to think of thought experiments that can't be reduced to arguments as paradoxes. He also notes that paradoxes need to have a kind of intuitive or widespread appeal. They have staying power as intellectual problems.

Let's first consider a thought experiment that's not quite paradoxical, despite having the air of paradox.

A truly paradoxical thought experiment will provide us with intuitively compelling reasons, or maybe just bare intuitions, for accepting seemingly absurd conclusions about things that matter.

Hilbert's Grand Hotel

David Hilbert

Imagine a grand hotel. It's Hilbert's hotel, named after the German mathematician David Hilbert, who dreamed it up in 1924. The popular hotel is always fully occupied, and yet it always has room for another guest, because it has an infinite number of rooms, numbered from 1 on up. To make room for a new guest, all the guests move over one room from their current room, R, to a new room, $R + 1$. This way, the hotel can accommodate any number of new guests.

But one day, an infinite number of new guests arrive. Rather than follow the same procedure, the staff open up all the odd-numbered rooms for the new guests. Using a doubling procedure, they move each current guest from their room R to a new room, $2R$, which is their room number times 2. Now the hotel has an infinite number of odd rooms available for the infinite number of new guests! They can repeat the process without end.

Hilbert's hotel might seem deeply paradoxical, like a contradiction, but it isn't really a paradox. The fact that an infinite number of guests can register for a hotel that's already full of an infinite number of guests sounds absurd, but that's only because we're thinking about infinite sets of things the same way we think of finite sets. If a real hotel was actually full, then it simply couldn't accommodate more guests without vacating rooms. Both things couldn't be true.

Though not really a paradoxical thought experiment, it illustrates mathematical facts about infinite sets that are otherwise provable and true, even though they're deeply counterintuitive. So, what does count as a truly paradoxical thought experiment?

The Surprise-Quiz Paradox

Imagine you're in school. It's Friday, and your teacher announces that there will be a surprise quiz next week. Since it's a surprise, you won't be able to predict which day of the week it will be on. You don't like this situation, so you speak up, saying that there can't be a surprise quiz next week. Puzzled, the teacher asks you to explain.

You say that if the quiz happens next Friday, it won't be a surprise, because if the class gets to the end of Thursday without having a quiz, everyone will know it will be on Friday. You then explain that the quiz also can't happen next Thursday, because the class already knows that it can't happen on Friday. So, if the class hasn't had it by the end of Wednesday, everyone will know it's going to be on Thursday, and so it won't be a surprise. Using the same reasoning, you explain why Tuesday and Wednesday are out as well. And if the quiz has to be on Monday, then it can't be a surprise; therefore, there won't be a surprise quiz next week.

This thought experiment has its roots in a 1948 article by D. J. O'Connor and has since taken on a life of its own. What makes it a paradox is that common sense tells us that even if we're told about the surprise quiz ahead of time, then as long as we're not told the exact day of the quiz, there can indeed be a surprise quiz next week—even if your clever line of reasoning seems to show that there cannot. We seem to have contradictory conclusions.

O'Connor's view is that the teacher's announcement is "pragmatically self-refuting," in the sense that it undermines itself. This is because the conditions of the upcoming event, the surprise quiz, are defined in such a way that announcing it entails that it can never be carried out. Viewed this way, the problem is with your teacher, who fails to recognize that the moment they announce a surprise quiz, they undermine it.

Or the problem might be that you made predictions that you weren't justified to make. If your teacher can just randomly select a day to give the quiz, then you can't make a prediction. If your line of reasoning depends on the ability to predict anything about when the quiz will be given, your argument isn't any good. The philosopher Roy Sorensen points out that this would be a risky plan on your teacher's part. If your teacher ends up randomly selecting next Friday, then your original line of reasoning would hold, and you can see the problem.

Another way to look at this paradox is to focus on what you, as the student, really know. The philosopher W. V. O. Quine thinks your argument that there cannot be a surprise quiz next week is a reductio ad absurdum, showing that you do not in fact know that your teacher's announcement is true. This type of argument shows that an argument is flawed because it leads to an absurd conclusion. After all, the conclusion you reach—that there cannot be a surprise quiz next week—is absurd, or at least quite implausible.

To see why, the philosopher Michael Huemer points out that we can extend your line of reasoning well beyond the next week. You could argue that there cannot be a surprise quiz in the next nine months, which is implausible. So, if Quine is right and you don't actually know that there will be a surprise quiz based on your teacher's announcement, then that helps explain why your reasoning goes wrong: It's based on faulty assumptions. This move is attractive, but it has deeply unappealing implications. It seems like a big cost to claim that you can't know that your teacher's announcement is true—or at least be strongly justified in believing that it's true.

What's giving us trouble is the fact that the quiz is supposed to be a surprise. If your teacher just announced that there will be a quiz next week, your clever argument wouldn't work. With that in mind, we probably don't want a radical solution to the paradox, like one that eliminates our ability to know things about the future. The surprise-quiz paradox doesn't have a universally-agreed-on solution; it remains a puzzle.

The Paradox of the Stone

Imagine an omnipotent, or all-powerful, being. Let's call that being God. Can God create a stone so heavy that not even God can lift it? This is a micro–thought experiment, captured by a simple question that raises some serious philosophical questions.

The paradox is an objection to the very notion of omnipotence, because there doesn't seem to be a good answer to the question. If we say yes, God can create a stone so heavy that not even God can lift it, then there's something God can't do, and so God's powers seem limited rather than unlimited. But, if we say no, God can't create such a stone, then, again, there's something God can't do. Either way, there's something this omnipotent God can't do. And so, there must be something fundamentally incoherent in the very notion of omnipotence if asking this simple question leads to such an absurd conclusion.

Your first response might be to deny that the stone in question is in fact a possible object. Maybe the paradox turns on a kind of conceptual mistake. But even if we grant that such a stone is inconceivable, we can just change the example to something more easily conceivable, like an indestructible object, and alter the thought experiment accordingly.

You next response might be to deny that the power to do this sort of thing is possible. The point here isn't that God lacks a power; it's that the request makes no sense. If we ask God to create a stone that's too heavy even for God to lift but then require that God lift it, we're asking God to do something that is incoherent or meaningless and is, in that sense, impossible. But it's not obvious why it's a meaningless request.

How else might we respond to the paradox? As the philosophers of religion Michael Murray and Michael Rea note, we can point out an ambiguity in the paradox itself. If we answer no to the original question, it seems like we'd be saying that either God's creating powers or God's lifting powers are limited.

As they point out, however, we can answer the original question negatively in a completely different way. We can say no, God cannot create a stone so heavy that not even God can lift it, because God's relevant powers—those of lifting and creating—are unlimited, and this means that God can never create a stone that he can't lift, since God could lift any stone, but he can nonetheless create any stone that's possible to create, since God can bring about anything that is possible. If this is right, we can conclude that the paradox of the stone fails to show that there's something fundamentally incoherent with the very notion of omnipotence.

In the middle of the 20th century, philosophers like J. L. Mackie used considerations like the paradox of the stone to argue not only that omnipotence is nonsense but also that the existence of an omnipotent being like God is logically impossible. Over the past 70 years or so, philosophers of religion have responded to the paradox of the stone in ways that have clarified the nature of omnipotence. In this way, the paradox of the stone has begun to serve as a teaching tool, a way to walk the uninitiated through the sometimes paradoxical-sounding mysteries in theology.

Reading

- Huemer, Michael. *Paradox Lost: Logical Solutions to Ten Puzzles of Philosophy*. Palgrave MacMillan, 2018.
- Mavrodes, George. "Some Puzzles Concerning Omnipotence." *Philosophical Review* 72 (1963): 221–223.
- O'Connor, D. J. "Pragmatic Paradoxes." *Mind* 57 (1948): 358–359.
- Sainsbury, R. M. *Paradoxes*. 3rd ed. Cambridge University Press, 2009.

8

WHAT NEWCOMB'S PARADOX SAYS ABOUT DECISIONS

The great 20th-century philosopher Robert Nozick first introduced the problem in this lecture in a 1969 paper. He attributed it to the physicist William Newcomb, which is why it's known as Newcomb's problem, or Newcomb's paradox. The power of the thought experiment is that it not only pulls our intuitions in different directions, but it also pits two principles of rational choice against each other.

Choosing between One Box and Two

You have two boxes in front of you. You can see there's $1,000 in one box. You can't see into the other box, but your friend told you that there's $1 million in it. You can choose to take only the box you can't see into, or you can take both boxes. What would you do?

The obvious choice seems to be to take both boxes. If your friend is telling the truth, then you'll get $1,001,000—but even if your friend is lying, you'll still get $1,000. In contrast, if you take only the box you can't see into, then, if your friend is lying, you won't get anything. But even if your friend is telling the truth, you'll still miss out on that extra $1,000. It's obvious that you should take both boxes.

Imagine there's a twist. You still have two boxes in front of you—a transparent box, which you can see has $1,000 in it, and an opaque box. Your friend tells you that they've been working with a team of neuroscientists on a prediction device called Apollo. Apollo's prophecies are grounded in hard facts and rigorous science, not guessing or magic.

Apollo can predict how people will act, but not without first mapping a person's brain. It combines that information with specific potential scenarios and predicts what that person would do in a given scenario. It has been amazingly accurate so far—it almost always gets things right. Your friend tells you that while you were sleeping last night, Apollo scanned your brain, because they wanted you to participate in their final study. As a reward for participating, you'll be offered a cash prize of up to $1,001,000, which you can win if you make the right choice.

The transparent box has $1,000 in it, and you can see that. Your friend tells you that in the opaque box, there may or may not be $1 million. You have a choice: take only the opaque box or take both boxes. The difference between this choice and the earlier one is that Apollo made a prediction. If it predicted that you would choose only the opaque box, the scientists put $1 million in it, but if Apollo predicted that you would choose both boxes, they didn't put any money in the opaque box. Knowing this information, what should you do?

If you take only the opaque box, you're going to get either no money at all or $1 million. But if you take both boxes, you're going to get either $1,000 or $1 million plus $1,000. Either way, you get more money if you choose both boxes. Right? So, you should choose both boxes, because choosing more money over less money seems like the rational or reasonable thing to do, doesn't it?

But then again, given Apollo's accuracy, you have good reasons for thinking that, if you were to open both boxes, Apollo would have predicted that, and the scientists wouldn't have put any money in the opaque box. If that's the case, you have a great reason for not choosing both boxes. Likewise, you have good reasons for thinking that, if you were to open only the opaque box, Apollo would have predicted that, and the scientists would have put $1 million in the box. If that's the case, it seems like you should take only the opaque box, not both.

We now have a problem. We have a seemingly decisive argument for choosing both boxes, but also a seemingly decisive argument for choosing only the opaque box. The arguments are equally compelling yet lead to incompatible conclusions. The version of the Newcomb's problem in this lecture comes from the philosopher Michael Huemer's book *Paradox Lost*. What's nice about Huemer's retelling is that he provides some details that make it plausible that your behavior can be predicted.

Principles of Rational Choice

If you think you should choose both boxes, then you might think that the rational thing to do, in general, is whatever will bring about the outcome you want. Since there's nothing you can do to change what Apollo predicted or what the scientists did based on that prediction, then what you should do now is choose two boxes, which will guarantee you $1,000.

The principle of rational choice that might be driving this intuition is known as the dominance principle. The basic idea is that if you have a choice between A and B, and you know, or have good reasons to believe, that A is better than B in all the possible ways the world could be, then it's rational to choose A over B.

By contrast, if you think you should choose only the opaque box, then you might think that the rational thing to do is whatever gives you the best evidence that you'll get what you want. Since Apollo is almost always right, you have good evidence that choosing only the opaque box is your best bet for getting $1 million, which is what you want. You also have good evidence that choosing both boxes will get you only $1,000, which isn't what you want.

The principle of rational choice that might be driving that intuition is known as the principle of expected utility maximization. The basic idea is that the rational choice in any situation is the choice that has the greatest benefit, where the greatest benefit is understood in terms of the highest "expected utility." And so, you should make choices that would maximize predicted good outcomes.

The idea behind this principle is that you need to consider both the goodness of the outcome and the chances that the outcome will come about. In the real world, it's often hard to do this, but, in general, calculating expected utility is easy. You multiply the measure of utility for each possible outcome—that is, the value of the outcome—by the measure of the probability of that utility accruing—that is, by the chances that the outcome will come about. Then you add up the products you've gotten for all the possible outcomes, if there's more than one.

Calculating Expected Utility

Although we've been unclear about exactly how accurate Apollo's predictions are, the principle of expected utility maximization seems to tell us that you should choose only the opaque box. To see this, imagine that Apollo is 90% accurate.

If you choose only the opaque box, there's a 90% chance that Apollo predicted that, in which case there will be $1 million in the box. But there's a 10% chance that Apollo got it wrong, in which case there won't be any money in the box. So, there are two possible outcomes.

To find the expected utility of choosing only the opaque box, we do the following calculation:

$1 million × 90% = $900,000

$0 × 10% = $0

$900,000 + $0 = $900,000

So, $900,000 is the expected utility of choosing only the opaque box.

If you choose both boxes, there's a 90% chance that Apollo predicted that, in which case there wouldn't be any money in the opaque box, and you'll only get the $1,000 in the transparent box. But there's a 10% chance that Apollo got it wrong, in which case there will be $1 million in the opaque box plus $1,000 in the transparent box. We again have two possible outcomes.

To find the expected utility of choosing both boxes, we do the following calculation:

$1,001,000 × 10% = $100,100

$1,000 × 90% = $900

$100,100 + $900 = $101,000

So, $101,000 is the expected utility of choosing both boxes. Clearly, $900,000 is a much higher than $101,000, and so the principle of expected utility maximization seems to support choosing only the opaque box.

The Voter's Illusion

If that scenario seems a little far-fetched, consider a real-world approximation of a Newcomb scenario, associated with George Quattrone and Amos Tversky's discussion of the so-called voter's illusion. In big elections, like US presidential elections, it's almost certain that anyone's individual vote, including yours, won't make a difference to the outcome.

But you might think that whether or not you vote is a sign of whether like-minded people are going to vote. You might also think that it's important that people like you actually vote in the election because turnout is what will determine who gets elected. And so, it seems like voting will count as evidence that what you want to happen is what will happen.

As thinkers like Arif Ahmed and others note, this example is a good enough Newcomb's scenario, because it raises the same issues. There are two possible outcomes—your preferred candidate either wins or doesn't. And you have a choice to make—you can vote or not. If we think of the choice to vote in big elections like we thought of choosing between one box or two, voting is like one-boxing, whereas not voting is like two-boxing. And the choice about whether to vote or not is a genuine personal choice with real-world consequences.

Causal versus Evidential Decision Theory

If you're willing to agree that a Newcomb's problem scenario is possible, then what can we say in the face of the problem it presents? Many scholars have said a lot about the problem. One way to clarify what's at issue is to reconsider the dominance principle.

Ahmed notes that there's a potential ambiguity in it, since we can understand the connection between what we do and what will happen in two different ways—either causally or evidentially. By the causal interpretation, the dominance principle is saying that there should be a causal connection between what we do and what ends up happening. It's rational to do the thing that's likely to bring about the best outcome, given the possible ways the world might be. Those who defend a version of causal decision theory like this will choose two boxes, because that will get them an extra $1,000.

By the evidential interpretation, the dominance principle is saying that there should be an evidential connection between what we do and what ends up happening. It's rational to do the thing that provides the best evidence of what you want to happen, given the possible ways the world might be. Those who defend a version of evidential decision theory like this will choose only the opaque box, because that choice is evidence that you're going to get $1 million.

In normal circumstances, we care about both causal and evidential connections. Newcomb's problem powerfully pries apart two conceptions of what it means to make a rational choice—two conceptions that we ordinarily think of as more or less complementary.

Reading

- Ahmed, Arif, ed. *Newcomb's Problem*. Cambridge University Press, 2018.
- Huemer, Michael. *Paradox Lost: Logical Solutions to Ten Puzzles of Philosophy*. Palgrave MacMillan, 2018.
- Nozick, Robert. "Newcomb's Problem and Two Principles of Choice." In *Essays in Honor of Carl G. Hempel*, edited by Nicholas Rescher, 114–146. Dordrecht: Reidel, 1969.
- Weirich, Paul. "Causal Decision Theory." *The Stanford Encyclopedia of Philosophy*. Winter 2020 ed., edited by Edward N. Zalta. https://plato.stanford.edu/archives/win2020/entries/decision-causal/.

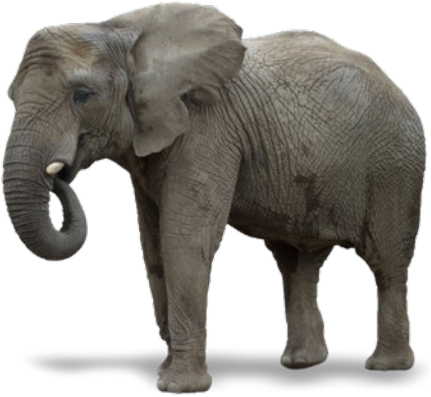
9

STORIES AS THOUGHT EXPERIMENTS

Thought experiments are a lot like stories, and vice versa. But can we learn something about the world from imaginary scenarios? In this lecture, you're going to explore this question while considering some thought experiments along the way. You'll look at the similarities and differences between philosophical and scientific experiments, and you'll see how works of fiction can also help us to understand our world and ourselves.

The Blind Men and the Elephant

Long ago, a king invited all the blind men in his kingdom to his courtyard. When they arrived, an elephant was brought before them. To some of the men, the king presented the elephant's head as though it were the whole elephant. To others, he presented only parts of the elephant, like its side, feet, tail, ears, trunk, or tusks.



He then asked them, "Do you now know what an elephant is like?" Each man answered yes but gave very different descriptions. They shouted at each other, even coming to blows with anyone who disagreed. Delighted, the king watched the spectacle unfold.

This story seems to have first appeared in the Buddhist tradition as a kind of allegory. The Buddha, by telling this story, highlights the absurdity of disagreeing about what we don't really know—because the blind men have only a partial and misleading understanding of how things are—but the Buddha also highlights the real dangers we face when we find ourselves so attached to our opinions that we can't tolerate those who disagree with us.

We don't have to read the story the way the Buddha wants us to. The Jain intellectual tradition uses this parable to explain three of their philosophical doctrines: many-sidedness, perspectives, and conditional predication. The elephant represents the many-sided nature of reality, the blind men represent the many different perspectives from which we might reasonably come to know reality, and what they say about the elephant represents conditionally true statements based on their unique perspectives.

This parable is interesting because it's a piece of fiction that walks a fine line between being a story and being a thought experiment. We can read it in different ways and learn different things, depending on our interpretation, which of course is often how stories and thought experiments work.

But we might worry that if stories and thought experiments are open to interpretation, we can't learn definitive lessons from them. As the philosopher Catherine Elgin asks, "Is it really possible to find out about the world by making things up?" Let's look at an actual experiment Elgin herself focuses on.

Exemplification

This Miller-Urey experiment picked up on the Oparin-Haldane hypothesis that life on Earth gradually emerged from inorganic molecules. The idea was that these inorganic molecules would undergo chemical reactions that would produce amino acids and more complex polymers. The experiment tested this hypothesis and offered the first hard evidence that life could arise from nonliving matter. But because the experimental conditions and the chemical reactions were not what we'd normally find in the world, Elgin noted that "the experiment revealed something important about the natural world." It "exemplified a path from inorganic to organic molecules."

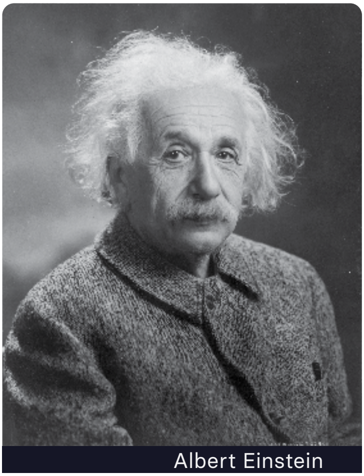
The idea of exemplification is important, according to Elgin. A fabric swatch is a sample of the kind of cloth you might use for a project, so it exemplifies a specific pattern, texture, color, or all three associated with the fabric, providing us with information. What the fabric swatch exemplifies also depends on our interests.

Actual scientific experiments, like the Miller-Urey experiment, are about the actual world, but Elgin notes that they're very distant from "the natural phenomena they illuminate." We make experiments happen; they don't occur naturally. We provide them with a beginning, middle, and end, and we have to think about what their results mean. What this means, Elgin suggests, is that actual experiments have similarities with thought experiments and stories. Thought experiments and stories are fictions, of course, but "the gulf between fact and fiction" might not be so great.

Of course, thought experiments and stories are even further distanced from the real world than scientific experiments. They're not real, and they're not accounts of what actually happened, either. Sometimes thought experiments and stories talk about things that couldn't take place in the real world. What makes for a powerful thought experiment, Elgin argues, isn't free-floating and fanciful, but carefully constrained. "By setting … constraints and drawing out the consequences" of what might happen given those constraints, Elgin notes, "the imagination serves as a laboratory of the mind, a venue in which hypotheses can be contrived, elaborated, and tested."

Einstein's Elevator Thought Experiment

Imagine a scientist in a room with no way to see out and no way to interact with anyone outside the room. Inside the room is a laboratory, so the scientist could do any experiment. As long as the room is moving at a constant speed in the same direction, there's no experiment the scientist could do that would determine whether the room was moving. Now imagine this laboratory is on Earth, and what happens in the laboratory is governed by Earth's gravity. What would happen if the scientist knocked a beaker off a table? Because of gravity, the beaker would fall to the floor with a certain trajectory, right?


Albert Einstein

Now imagine this laboratory isn't on Earth, and instead it's a giant elevator being accelerated at a constant rate through empty space, with no jarring disturbances to give things away. From the outside, if someone could look into the elevator, they would think the falling beaker collides with the floor because the floor is moving upward toward it. But what would the scientist think and observe? If they don't know that the lab is a giant elevator being accelerated at a constant rate, they'll observe exactly what they would have observed on Earth: The beaker will fall to the floor like normal; the floor won't appear to accelerate upward to meet it.

Part of exploring a thought experiment is testing the various assumptions that go into our interpretation, and the various commitments that lead us to the conclusions we draw. That's OK, as long as we're willing to clarify those background commitments and recognize that the story doesn't prove anything. As a thought experiment, it's a starting point, not the final word.

What does this tell us? First, the thought experiment leads to the insight that from inside either lab, the scientist can't do any actual experiment that could provide information about whether the lab was in a gravitational field, like it would be on Earth, or whether the lab was being accelerated upward, like it would be on that giant elevator. From this insight, Einstein teased out the equivalence principle, which roughly says that uniform gravitational forces like those governing the lab on Earth are equivalent to inertial forces like those governing the elevator lab in space. Put differently, the gravitational mass of the beaker on Earth is indistinguishable from and equivalent to the inertial mass of the beaker on the elevator accelerating "upward" through space.

It's OK that we couldn't really do this kind of experiment, because the thought experiment does all the work we need done. It allows us to explore what would happen in conditions that are impossible for us to produce. Of course, it requires the suspension of disbelief as well as the suspension of belief—we're going to have to bracket some things that we think are true so they don't get in the way. But that allows the thought experiment to focus only on what matters.

None of this is too different from an actual experiment. It's an exemplar that gives us a genuine insight into the natural world. Of course, like any experiment, it's up for interpretation and requires us to work through it to articulate its results in the best way, based on a variety of background assumptions and commitments. Einstein's elevator thought experiment is a scientific thought experiment that shares many of the features of actual experiments, despite being the product of imagination.

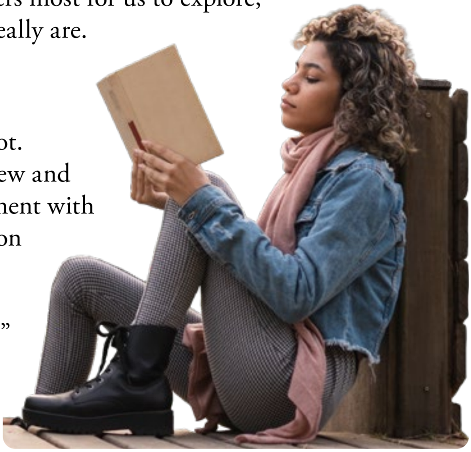
Thought Experiments and Fiction

Elgin notes that the main difference between philosophical thought experiments and their scientific counterparts is the wider scope of things philosophical thought experiments cover. Philosophical thought experiments explore modal, conceptual, and theoretical issues, but also ethical commitments, epistemic standards, metaphysical categories, logical relations, and introspectively available properties about the contents of our own minds.

Philosophical thought experiments also have more freedom. They're often less constrained by theoretical and empirical considerations. It's also one of the reasons why it's so easy to entertain a philosophical thought experiment without awareness of the theoretical commitments and background assumptions we bring to them. According to Elgin, all of this helps us see how fiction works as an extended thought experiment. Works of fiction provide us with "epistemic access to aspects of the world that are normally inaccessible," the same kind of things that philosophical thought experiments help us see.

When you do an actual experiment, you have an actual result, but in the case of thought experiments and fiction, what's exemplified, Elgin suggests, are abstractions about ourselves and the world, offering insights into how things might be, must be, can't be, and are. A powerful story offers us insights partly because it focuses so tightly, blocking out what doesn't matter to focus on what does. Whether the story—the thought experiment—is suitable depends on how well it exemplifies what matters most for us to explore, not whether it describes how things really are.

In fiction, we also have the chance to explore the inner lives of others and ourselves in a way we otherwise cannot. Elgin notes, "we take up a point of view and try it on for size. In effect, we experiment with the perspective." She argues that fiction and thought experiments provide us with opportunities to "recognize and marshal information we already have." We're often not learning something new so much as we're appreciating what we knew in a different way.



As exemplars, Elgin argues, fiction and thought experiments provide indirect evidence about what they exemplify, and they also provide us with evidence about themselves. This is important to keep in mind because we can learn both about the world and about the work. In asking what a short story means, we might be hoping to learn what it tells us about the human condition, or we might be hoping to learn something about the story itself, distanced from what it potentially exemplifies.

Reading

- Breyer, Daniel. *Word Philosophy: 50 Puzzles, Paradoxes, and Thought Experiments*. Routledge, 2023.
- Elgin, Catherine. “Fiction as Thought Experiment.” *Perspectives on Science* 22, no. 2 (2014): 221–241.
- Marsh, Elizabeth J., Andrew C. Butler, and Sharda Umanath. “Using Fictional Sources in the Classroom: Applications from Cognitive Psychology.” *Educational Psychology Review* 24 (2012): 449–469.

Table of Contents

10

EINSTEIN'S REVOLUTIONARY THOUGHT EXPERIMENTS

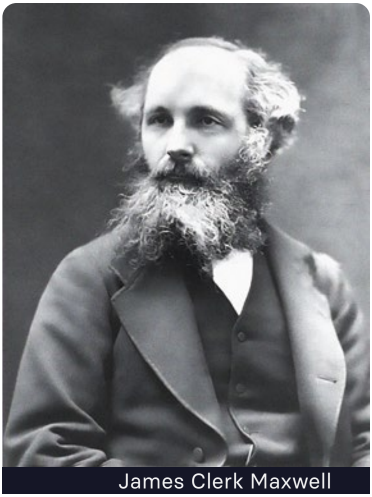
Einstein had a knack for devising penetrating tools for thinking—but very often, their importance is misrepresented or misunderstood. In this lecture, you're going to explore some of his most famous thought experiments. By looking at a few of them, you can get a better understanding of both the power and limitations of thought experiments.

The Principle of Relativity and the Light Postulate

In his *Autobiographical Notes*, Albert Einstein talks about his light beam thought experiment, something that he first considered when he was just 16 years old. He tells us that this thought experiment—which asks us to imagine running alongside a beam of light—contains the germ of special relativity. At the core of this theory are two assumptions, or postulates—one about relative motion, the other about the speed of light.

The first postulate, often called the principle of relativity, is that the laws of physics are the same for everyone in all inertial frames of reference. In other words, the laws of physics are the same for everyone, no matter what their point of view—as long as their relative motion is constant, not sped up or slowed down.

The second postulate is about the speed of light. It says that in a vacuum, like outer space, the speed of light is constant and the same for everyone who measures or observes it—no matter whether they're moving toward the light source, moving away from it, or traveling right alongside it. Einstein based this postulate on two grounds. The first was his confidence in James Clerk Maxwell's 1873 theory of electrodynamics. Einstein maintained that Maxwell had demonstrated that light was a wave moving in an electromagnetic field, and that the speed of that wave had a definite value, *c*.



James Clerk Maxwell

The question then became: Does the measured speed of light increase or decrease depending on our relative motion to it? If we're chasing a light beam, for instance, it seems like we should measure its speed as slower relative to us, since we're moving in the same direction as it. That's the intuition, but it turns out that light behaves differently.

The early experimental evidence for this was provided by the Michelson-Morley experiments in the 1880s, which is the second ground for the light postulate. Those experiments found that there was no measurable difference in the speed of light whether the earth was moving toward the sun or away from it. Light, it turned out, had a constant speed regardless of your motion relative to it.

The Theory of Special Relativity

When we combine these findings for the light postulate with the principle of relativity, we get the foundation for Einstein's theory of special relativity, which looks at motion as it is measured between different frames of reference, or different points of view. That's part of what makes it a theory of relativity. And special relativity is "special" because it focuses on cases where motion is uniform or constant. In 1915, Einstein extended his early theory of relativity (which he only later called special) to include cases of nonuniform motion. That's Einstein's theory of general relativity.

So, how do we get from thinking about a light beam to the beginnings of special relativity? The first thing to note is that what Einstein imagines is based on Maxwell's theory of electrodynamics. In this view, light waves are relevantly like ocean waves. If you were to imagine traveling alongside an ocean wave, you'd observe it as at rest relative to you. It seems that's how it should be for a light wave, too.

Einstein says that there doesn't seem to be any such thing as a static light wave at rest. He offers three reasons: First, it would go against what we know from experience; second, it would go against Maxwell's equations; and third, an observer couldn't know that he's "in a state of fast uniform motion."

We now have the germ of special relativity in the form of a kind of paradox. The principle of relativity tells us that light should behave the same way no matter how fast we're moving. But thinking about what it would be like to chase a beam of light tells us that there's something's wrong with how we're thinking about light, because it does seem to matter whether we're moving at 25 miles per second or the speed of light.

Thought experiments in the natural sciences are rarely, if ever, just imaginary scenarios. It's one thing to imagine chasing a beam of light; it's quite another to grasp the scientific import of the thought experiment. This suggests that, with the right background knowledge, Einstein's thought experiment reveals a problem that only special relativity can solve—which would make it very powerful indeed. Is that true, though?

Let's say you know a lot about late 19th-century physics. You endorse Maxwell's theory of electrodynamics and the view that light waves propagate through the ether. Would Einstein's thought experiment show you that there's a problem with your views?

The philosopher of science John Norton suggests that it might not. Einstein's reasons for why there couldn't be any such thing as a static light wave at rest might not be compelling enough to someone who also held that the ether existed. Norton argues that Maxwell's equations predict that light would become static for observers traveling at the speed of light.

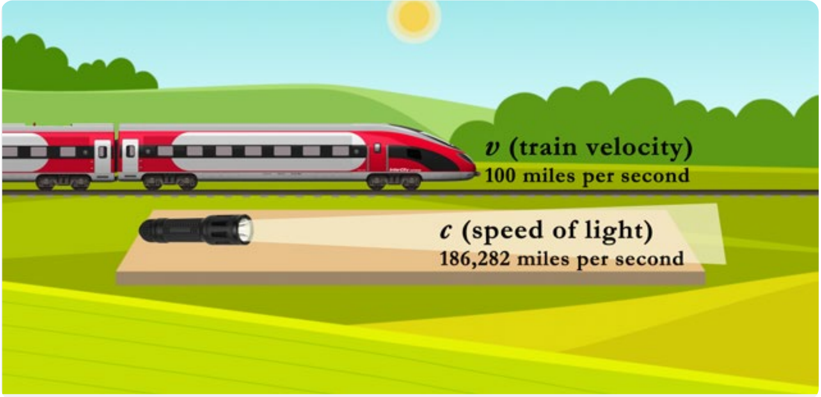
So, someone who endorses an ether-based theory of electrodynamics might just insist that our intuitions go astray when we think about chasing a beam of light. The intuitive idea of the ether makes it seem as though change is essential for a light wave to exist, but Maxwell's equations show otherwise. If the ether is real and Maxwell's equations are right, then what we should do is revise our pre-theoretic intuitions about what it would mean for light to propagate through the ether.

What about Einstein's worry that an observer couldn't know that they're in a state of fast uniform motion? What's doing the real work in Einstein's thought experiment is the principle of relativity. The fundamental problem is that an ether-based theory of electrodynamics makes it possible for us to determine our absolute motion. It's not that we can't determine that we're in a state of fast uniform motion; it's that the then-current way of thinking about light violates the principle of relativity, and that's what the light beam thought experiment reveals.

Relativity of Simultaneity

To reconcile Maxwell's electrodynamics with the principle of relativity, Einstein did a lot of hard technical work, which we can't possibly review here. But the short story includes Einstein's commitment to the light postulate. The problem is that the light postulate and the principle of relativity are, as he stated, "apparently irreconcilable." To explain why, Einstein has a thought experiment.

Imagine a railway embankment in a space-like vacuum. Imagine that a ray of light is sent along the embankment, with the tip of the ray traveling at the speed of light, which is about 186,282 miles per second, relative to the embankment. Now imagine that a train is traveling on the embankment in the same direction as the ray of light, with a velocity of 100 miles per second relative to the embankment. How fast is that ray of light going relative to the train?

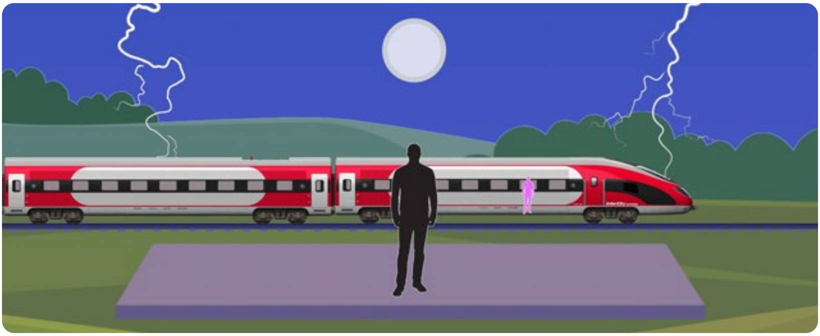


According to classical physics, we should be able to figure this out in the same way we figure out how fast anything else is going relative to the train. The people on the embankment should measure the speed of that ray as 186,282 miles per second, but the people on the train should measure it as 186,282 minus 100, or 186,182 miles per second. The problem is that, according to the principle of relativity, everyone—no matter their point of view—should measure the speed of light in a vacuum as exactly the same, 186,282 miles per second, because the light postulate is a general law of physics. So, the light postulate and the principle of relativity seem incompatible.

To reconcile the two, Einstein introduces the relativity of simultaneity. The basic idea behind it is that whether two or more events happen at the same time depends on the point of view from which those events are observed. What's odd about this is that it's possible for two things to happen simultaneously from your point of view but happen at different times from someone else's point of view—and you'd both be right about when those events happened.

Einstein suggests, or perhaps just stipulates, that what it means for two or more things to happen simultaneously is that someone observes them happening at the same time. The "time" of an event, he notes, is just whatever a clock in the "immediate vicinity … of the event" reads.

With that in mind, consider another thought experiment that Einstein used. Imagine that you're back on that embankment and two lightning bolts strike the train tracks—one to your right, the other to your left, with you exactly in the middle. If the light from each strike reaches you at the same time, you'll say that they're simultaneous.



Now imagine someone else who's moving really fast on a train toward the lightning strike to your right and away from the strike to your left. This person is in the middle of the train when the lightning strikes, at the same place as you standing on the embankment.

We might want to insist that the person on the train will also observe those lightning strikes as happening at the same time, but we can't do that, because the speed of light is constant. For the light from the left-hand strike to reach that person at the same time as the right-hand strike, the light on the left would have to be going faster.

If we stick to the light postulate and the principle of relativity, we have to say that the person on the train will observe the right-hand lightning strike earlier than the left-hand strike and conclude that they did not in fact happen at the same time. As Einstein puts it, "Events which are simultaneous with reference to the embankment are not simultaneous with respect to the train."

We normally think that whether something is simultaneous or not is an absolute fact, but the relativity of simultaneity forces us to think otherwise. The train thought experiment helps us make sense of this strange and defining feature of Einstein's special relativity. That's its real power.

It's tempting to think of Einstein as a singular genius whose imagination provided him with insights no one else had, and while that's not entirely the wrong picture, it misses a lot. He didn't just dream things up. He considered the beam-of-light thought experiment in the context of the cutting-edge science of his time. As John Norton put it, "Einstein could do more than those who came before him precisely because he had absorbed their work and understood it fully. Only then could he see their failings and know just where the new theories are to be found."

Albert Einstein

Reading

- Brown, James Robert. *Laboratory of the Mind: Thought Experiments in the Natural Sciences*. 2nd ed. London: Routledge, 2011.
- Einstein, Albert. *Relativity: The Special and the General Theory*. 100th anniversary annotated ed. Princeton: Princeton University Press, 2019.
- Norton, John. “Chasing the Light: Einstein’s Most Famous Thought Experiment.” In *Thought Experiments in Philosophy, Science and the Arts*, edited by James Robert Brown, Mélanie Frappier, and Letitia Meynell, 123–140. London: Routledge, 2013.
- ———. “How Einstein Did Not Discover.” *Physics in Perspective* 18 (2016): 249–282.

Table of Contents

11

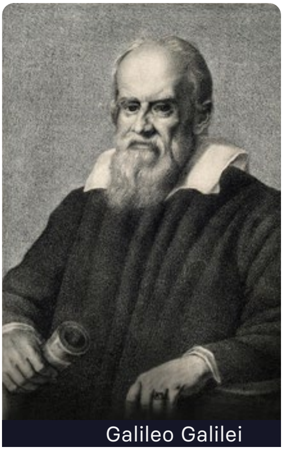
GALILEO'S AND SCHRÖDINGER'S THOUGHT EXPERIMENTS

In this lecture, you're going to look at a few of the most famous thought experiments in the history of science, with special attention given to Galileo's falling bodies and Schrödinger's cat. You'll be able to draw out some lessons about how thought experiments work in the physical sciences, noting their limitations as much as their power.

Galileo’s Falling Bodies

Aristotle

It seems intuitive to think that objects that weigh the same would fall equally fast, and a heavier object would fall faster than a lighter one. The classical Greek philosopher Aristotle noted that “we see the same weight or body moving faster than another for two reasons, either because there is a difference in what it moves through … or because, other things being equal, the moving body differs from the other owning to excess of weight or of lightness.” He claimed that the speed with which an object falls is proportional to its weight. That claim is less intuitive than just that heavy things fall faster than lighter things, and it’s what first encouraged philosophers to push back against Aristotle.


Galileo Galilei

In the 17th century, Galileo Galilei challenged Aristotle, arguing that if we “remove entirely the resistance of the medium,” like air, through which something falls, we’ll find that all materials—no matter what they are, no matter their shape, no matter how heavy they are—fall “with equal speed.” To make his case, Galileo uses a famous thought experiment: Imagine what would happen if a heavy cannonball was attached to a light musket ball, and they were both dropped together, at the same time, from someplace high.

On the one hand, the Aristotelian would expect the cannonball’s speed to be slightly slower than it would be if it were falling on its own, due to resistance from the musket ball. On the other hand, the two balls are heavier together than the cannonball alone, so the cannonball’s speed should be slightly faster than it would be if it were falling on its own.

Of course, that's absurd, and so the Aristotelian view is false. The way out of the problem is to stop insisting that whether an object is heavy or light matters at all. The way out is to assume that "both great and small bodies … are moved with like speeds."

Galileo's thought experiment is powerful because it forces us to reconsider our intuitive sense of how the world works—that heavy objects fall faster than lighter ones. It also opens us up to a new way of thinking about the world that was unavailable before—that the speed of falling bodies might be constant and independent of their weight.

What's even more powerful about the thought experiment is that it shows that any theory which holds that the rate at which bodies fall depends on their weight would be inconsistent. This means that it shows us something about how the world has to be. What more could we ask from a thought experiment than that?

> In the case of Galileo's falling bodies, we see how thought experiments can challenge our intuitions, open us up to new ways of thinking, serve as counterexamples, and reveal something about how the world has to be. We also see that thought experiments, even in the sciences, are at least sometimes preferable to actual experiments.

Quantum Superposition

Some people think that thought experiments in quantum mechanics can reveal deep insights into reality. As the branch of physics that focuses on the microscopic world that we don't experience day-to-day, quantum mechanics emerged in the early decades of the 20th century and is one of the most successful scientific theories ever.

Thomas Young

It's also notoriously mysterious. One of those mysteries is quantum superposition: the notion that quantum phenomena, like photons, seem to exist in different states at the same time. In 1801, Thomas Young performed his famous double-slit experiment to show that light behaves like a wave rather than a particle, but later versions of the experiment revealed something quite unexpected.

If you fire photons, or particles of light, at two parallel slits in an otherwise opaque barrier, the particles going through one slit interfere with those going through the other slit, producing a wave pattern of bright and dark bands on a photographic plate behind the barrier. What's odd is that the same wave patterns occur when a single photon is fired at the double-slit barrier. It seems to pass through both slits at once and manages to interfere with itself. The experiment shows that light has both wavelike and particle-like properties. Crucially, this isn't just about how light behaves; it's about everything—at least at the quantum level—including subatomic particles like electrons.

When we measure which of the two slits a particle of light goes through, we find that it has only gone through one, and the interference pattern disappears. Just as puzzling, when we try to find the position of the photon, we can no longer accurately measure its velocity. And when we measure its velocity, we can no longer accurately measure its position. It seems like the process of measurement "collapses" the superposition. How do we make sense of this?

Schrödinger's Equation and His Famous Cat

In the late 1920s, physicist Max Born suggested that electron waves should be viewed probabilistically. Very roughly, he suggested that where an electron wave is large, it's more likely that we'll find the electron there, and where it's small, it's less likely that we'll find it there. Another physicist, Erwin Schrödinger, then formulated the most fundamental equation in quantum mechanics. The Schrödinger equation, which amounts to a law of nature, describes the shape and development of the waves associated with quantum phenomena, giving probabilistic information about a particle, including where it could be and how fast it might be moving. Schrödinger's equation made it possible for physicists to precisely describe and accurately predict quantum phenomena, but unlike classical mechanics, these descriptions and predictions were probabilistic, not deterministic.

Erwin Schrödinger

$$H(t)|\psi(t)\rangle = i\hbar\frac{\partial}{\partial t}|\psi(t)\rangle$$

In 1927, physicist Werner Heisenberg introduced what's now known as the Heisenberg uncertainty principle, which identifies a relationship between how precisely we can measure something's position and how precisely we can measure its velocity. Roughly, the relationship is that the more precisely we can measure or know one, the less precisely we can measure or know the other.

Werner Heisenberg

The Schrödinger equation and the Heisenberg uncertainty principle mathematically describe the quantum realm, and in that respect, they help us make sense, in a technical way, of what it means for something to exist in different states at the same time, or superposition. But what does it mean to have a law of nature that we can never observe, or to say that reality is blurred and indeterminate? What are the implications?

By 1935, Schrödinger himself was bothered by these questions, so he considered what he called a quite ridiculous case. Imagine that you've conducted an experiment and placed a cat

> in a steel chamber, along with the following device (which must be secured against direct interference by the cat): in a Geiger counter, there is a tiny bit of radioactive substance, so small, that perhaps in the course of an hour only one of the atoms decays, but also, with equal probability, perhaps none; if it happens, the counter tube discharges and through a relay releases a hammer that shatters a small flask of [poisonous gas].

If things exist in superposed states until they're observed or measured, then after an hour, the radioactive substance has both decayed and not decayed; the Geiger counter is both triggered and not triggered; the poisonous gas has been both released and not released. And, as Schrödinger puts it, "the living and the dead cat … [is] mixed or smeared out in equal parts." This is how things are until you open the chamber and check on the cat. Once you do this, though, the cat, which was both dead and alive, jumps from that blurred state to a definite state of being either dead or alive.

The thought experiment highlights that an indeterminacy originally restricted to the microscopic quantum world becomes transformed into macroscopic indeterminacy, which can then be resolved only by direct observation. So, if the Schrödinger equation is right and it captures how the world really is, then we seem to be left with an absurdity that the cat is both dead and alive, and we can't just explain that away by saying that superposed states "collapse" when we observe them. If measurements themselves, like opening the chamber to check on the cat, are ordinary physical interactions that should be described by the laws of nature, then we have to explain why those interactions seem to suspend or dissolve the laws of nature as described by quantum theory.

The power of his thought experiment is that it vividly and memorably demonstrates a serious theoretical and conceptual problem—a problem that might otherwise be easy to miss, given the complexities of quantum theory.

Interpretations of Quantum Mechanics

The standard interpretation of quantum mechanics in Schrödinger's day was the Copenhagen interpretation. It notes that quantum mechanics gives us knowledge about the world, but we don't have to think of it as providing an account of how things really are. It put up a kind of barrier between the mysterious microscopic quantum world and the ordinary macroscopic world, dividing things up into a part where quantum mechanics applies and a part where it doesn't.

Schrödinger's thought experiment, in part, removes that barrier and forces us to confront quantum weirdness directly. Understood this way, Schrödinger's cat is an intuitive counterexample to an instrumentalist interpretation of quantum mechanics. Surely, we can't just accept such an interpretation if it leads to the absurdity that the cat is both dead and alive.

One way to sidestep Schrödinger's cat is to insist that the barrier is best understood as between the observer and the observed. Physicist John von Neumann argued that the superposed state of the cat in the chamber before observation describes how reality is, but that when we observe that system, that superposed state collapses into a definite state. This is the collapse interpretation of quantum mechanics. What makes it an interpretation is that von Neumann doubles down on the idea that subjective experience plays a causal role here. Consciousness causes the collapse, not just measurement. There's a distinction between the observer and what's being observed, and sometimes that distinction includes physical processes as well as mental ones.

Thinking about the boundaries of the observer and the observed might make you think differently about Schrödinger's cat. If the cat were also a conscious observer, would it experience being in a state of superposition—being "smeared out" between life and death?

Wigner's Friend Paradox

In a 1961 article, physicist Eugene Wigner considered questions about observation. Imagine that you and your friend are in the lab doing quantum mechanical measurements together. You step out of the main lab for a moment while your friend measures the decay of an atom or the spin of an electron. Here, your friend is like Schrödinger's cat and you're the observer.

From your perspective, the lab is just a big quantum system, just like Schrödinger's chamber. So, everything in the lab, including your friend and whatever they're measuring, is in a state of superposition. This means that the results of your friend's measurements are indeterminate from your perspective. This superposed state will only collapse when you open the door to the lab and ask your friend what they measured.

Inside the lab, your friend has made a quantum measurement, and so, from their perspective, the small quantum system they've measured has collapsed—it's not in a state of superposition. So, the results of their measurement are determinate from their perspective before you open the door. Before then, it seems like what's going on in the lab was both superposed and collapsed. This is the paradox.

Wigner thinks we can resolve it by being clear about when the collapse actually occurred. With Schrödinger's cat, it seemed clear that the system collapsed when you opened the chamber and looked at the cat, but for Wigner's puzzle, does the collapse occur when your friend takes the measurements or when you open the door?

Wigner thinks that the big quantum system of the lab had already collapsed before you opened the door, and he thinks this gives us some reason to adopt a version of the collapse interpretation of quantum mechanics. It also gives us some reason to conclude that "the being with a consciousness must have a different role in quantum mechanics than the inanimate measuring device" and the laws of physics are violated "where consciousness plays a role."

That might sound like a ridiculous conclusion to you. How could conscious observation play that kind of role in how the world is? You might find yourself questioning the reliability of the thought experiment. Does Wigner's friend really tell us something about how the world is, or does it merely pump our intuitions in the worst sort of way?

In the cases of Schrödinger's cat and Wigner's friend, we see the power of thought experiments to leverage our intuitions in ways that force us to clarify even the most complex scientific theories.

In this lecture, we've been able to witness the power of thought experiments as counterexamples and challenges. What all these examples have in common is they allow us to block out complicating and potentially problematic information, so we can focus on puzzles that are at the very heart of science—and that may be where their true power lies.

Reading

- Galilei, Galileo. *Dialogue Concerning the Two Chief World Systems, Ptolemaic and Copernican*. Translated by Stillman Drake. 2nd rev. ed. University of California Press, 1962.
- Gendler, Tamar Szabo. *Thought Experiment: On the Powers and Limits of Imaginary Cases*. London: Routledge, 2000.
- Maudlin, Tim. *Philosophy of Physics: Quantum Theory*. Princeton: Princeton University Press. 2019.
- Wigner, Eugene. "Remarks on the Mind-Body Question." In *The Scientist Speculates*, edited by Irving John Good. London: Heinemann, 1961.

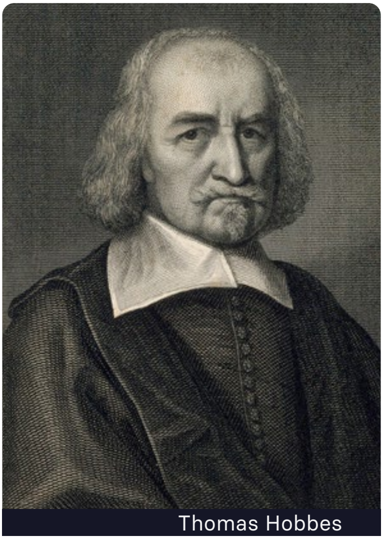
Table of Contents

12

WHAT MAKES IDENTITY THE SAME OVER TIME?

This lecture begins with a discussion of material objects—things like ships and chariots—and ends with a discussion about persons. That's because some of the same underlying problems can be found in each. Topics include what it means for two things to be identical and how parts are related to wholes.

The Ship of Theseus


Thomas Hobbes

We're first told about the ship of Theseus in Plutarch's *Life of Theseus*, but the 17th-century philosopher Thomas Hobbes embellished the story in ways that make it especially compelling. It's now one of the most famous thought experiments in metaphysics.

As legend has it, after the hero Theseus died, his ship was kept in the harbor of Athens. As the ship's planks decayed over time, the Athenians gradually replaced them with new planks, one by one, so that eventually the ship in the harbor was composed of entirely new planks.

What we want to know is whether the original ship of Theseus is identical to the repaired ship or if it is another ship entirely. Can we put a limit on how much change something can undergo and still remain the same thing?

An important fact about identity is that it's transitive, meaning that if A is identical to B, and B is identical to C, then A is identical to C. With that in mind, think about your car. If we can replace one part of your car, like a headlight, and say that it's still the same car, then we can imagine gradually replacing each part of your car, one by one, so that we have a succession of cars—car A, car B, car C, and so on. And every time we replace a part, we want to say that the car remains the very same car. Let's say that car Z is the final car and that it no longer shares any of the same parts with car A. Given the transitivity of identity, we should say that car A is identical to car Z.

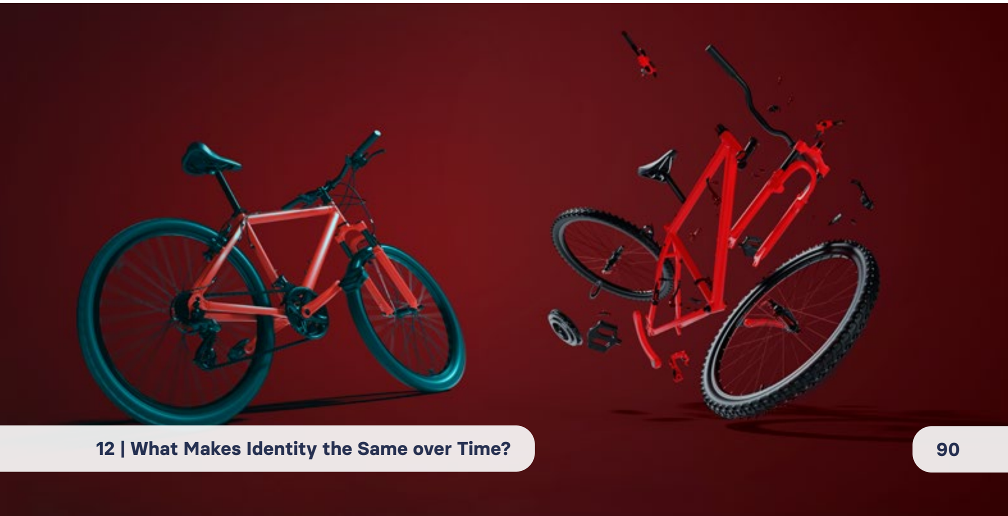
Using this logic, this is also what we should say about Theseus's original ship and the repaired ship—that they're identical despite not sharing any of the same parts. But there's a twist.

The Repaired and Reassembled Ships

Each time the ship received a new plank, the old plank was stored away in a place that naturally preserved the wood, until all the original planks were stored there. Many years later, an archivist discovered the discarded planks and commissioned a team of experts to reassemble them, so that each plank was in the same position it had been in the original ship before it was removed. The museum later unveils the reassembled ship, and the two ships rest side by side in the harbor. Which of the two ships in the harbor—the repaired ship or the reassembled ship—is identical to the original ship of Theseus that made the first journey?

Consider this first: Every winter, you disassemble your bicycle into its component parts and store them in a way that frees up space in your garage. Every spring, you put the parts back together just as they were before you stored them. Surely, the reassembled bicycle is the same bicycle as the one you disassembled at the start of the winter. It's not a new bike! Surely, then, the reassembled ship is identical to the original ship.

The problem is that we have good reasons for saying that both the repaired and reassembled ships are identical to the original ship. But since identity is transitive, we'd have to say that, if the repaired ship is identical to the original ship and the reassembled ship is identical to the original ship, then the repaired ship would be identical to the reassembled ship. But that's nonsense, because the repaired ship and the reassembled ship are two distinct ships, made of different materials, existing in two completely different places.



You might question the point that they are two different ships. There's a good philosophical reason not to question it, captured by what's known as Leibniz's law, or the principle of the indiscernibility of identicals. The idea behind it is that if one thing has a property that some other things don't have, then those two things aren't identical, because if they were the same thing, they'd share all the same properties. The repaired ship and the reassembled ship don't share all the same properties, and so they're not identical.

Leibniz's law, named after the great philosopher and mathematician Gottfried Leibniz, says that if two things are discernible, they're not identical, because they don't share all the same properties.

It seems absurd to say that the ship of Theseus ceases to exist, and we can't just arbitrarily pick one ship over the other. So, where does this leave us? If the repaired ship had never existed at all and the original ship had been disassembled, stored for some time, and finally reassembled—just like your bicycle—then the reassembled ship would have been identical to the original ship, no question about it. So, to pick the repaired ship over the reassembled ship, we'd be committed to saying that whether A (the original ship) is identical to B (the reassembled ship) depends on facts about some other thing, C (the repaired ship), which is not identical to the reassembled ship. This is absurd.

We might think that we need to say something extreme in response to the puzzle. This is often what philosophers have done. For instance, we might say that there really were two ships that coincided with each other, and it was only over time that they separated. Or we might say the original ship was identical to the repaired ship but then stopped being identical to it when the reassembled ship emerged, because being identical to something is the kind of fact that can itself change under the right circumstances. Or we might try something less radical, like arguing that there's an important difference between storing versus discarding parts and whether those parts still belong to the thing.

As a thought experiment, the ship of Theseus is puzzling because it triggers two opposing judgments. The power of it is that it's providing something like philosophical motivation to think carefully about an otherwise mundane thing.

The Chariot

Thus far, we've been assuming that parts make up objects that exist, but is that a reasonable assumption to make? Imagine a chariot. Think about its various parts—axles, wheels, chassis, and so on—and how those parts are arranged. Can you identify the chariot with any of its parts? It doesn't seem like you'd have a chariot if you just had an axle. So, it looks like the chariot isn't identical to any single one of its parts.

Maybe the chariot, like your bicycle, is just the collection of all its parts. But if you took one part away, would there still be a chariot, or would it cease to exist? Just like with the ship of Theseus, we'd still want to say that a chariot can remain the same thing even if it goes through a small change in its parts. So, it doesn't seem like the chariot is identical to the collection of its parts.

Keep in mind that the chariot, as a collection, would be a single thing, whereas its parts are many different things. It doesn't make sense to say a single thing is identical with its many parts. And it doesn't make sense to say the chariot is something completely different from its parts. So, where is the chariot? It seems like there is no chariot.

The chariot thought experiment comes from an important Buddhist text written in the 1st or 2nd century CE by an unknown author. It's a dialogue between King Milinda, a real person who ruled an Indo-Greek empire in the 2nd century BCE, and a Buddhist monk named Nāgasena, who may or may not have been a real person. The later 5th-century Theravada Buddhist philosopher Buddhaghosa endorses and develops the chariot case.

In the Buddhist tradition, the thought experiment is supposed to function as an argument—as a justification for a certain kind of view about the relationship between parts and wholes. In contemporary metaphysics, this view is called compositional nihilism.

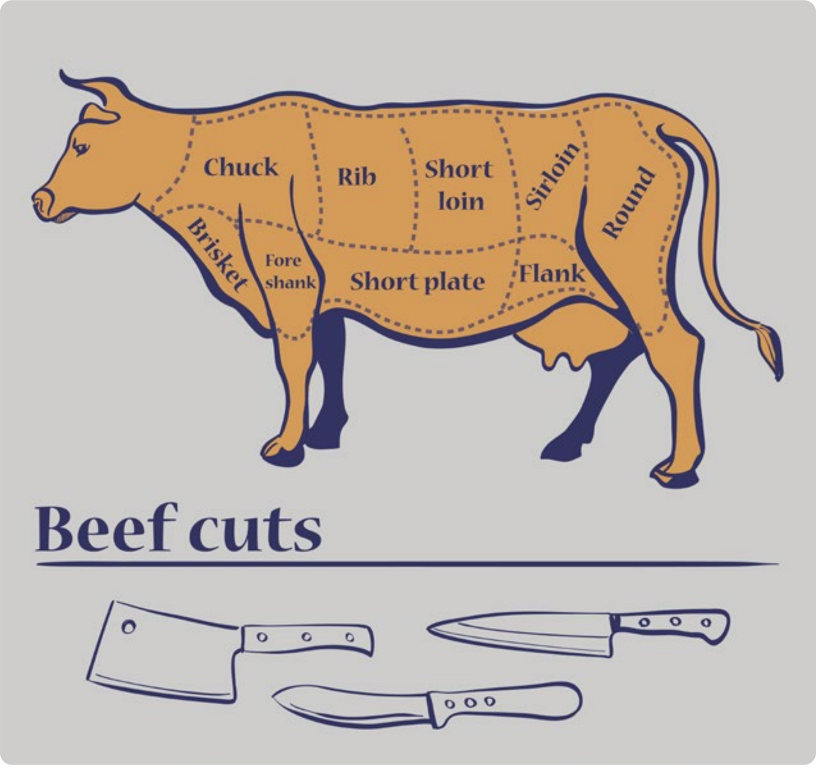
Mereology is the study of parts and wholes—how parts are related to each other to form objects, and what the relationship between parts and wholes is. Mereological nihilism is the view that there are only parts—wholes don't exist.

Compositional Nihilism

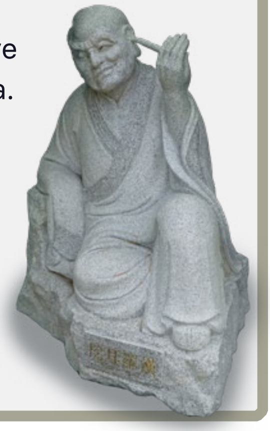
Compositional nihilism, also known as mereological nihilism, is the view that there are no composite objects. Roughly, the idea is that many individual things never come together to compose one thing. When a bunch of parts seem to form a whole, like a chariot, there are actually only partless entities arranged in a certain way. This is not the intuitive view. That view is known as compositional holism, which says that many distinct parts can come together to make a single whole—a composite object made of many individual entities.

Nāgasena points out to King Milinda that, although there are in fact no chariots, strictly speaking, it's still useful to talk about them. The name *chariot* is a verbal handle, a convenient designator. The idea is that chariots have a kind of conventional existence—they don't, strictly speaking, exist at the fundamental layer of reality, but they have a kind of socially constructed existence. So, when we talk about chariots or believe that they exist, we say and believe things that are conventionally true but ultimately false.

Picking up on this, Buddhaghosa offers an analogy. He asks us to consider a butcher who raises cattle and sells meat. While the butcher is feeding a cow, leading it to the slaughterhouse, slaughtering it, and even considering its corpse, he never stops thinking of it as a cow. It's only when he begins carving it up into parts that he stops thinking of it as a cow and starts thinking of it as meat.In that same way, Buddhaghosa suggests, when we think like an ordinary person, attending to our daily business, we think of ourselves as living beings, as persons. It's only when we start thinking of ourselves as we really are—conceptually carving ourselves up into parts—that we move from seeing ourselves as persons to seeing ourselves as merely bundles of impersonal elements. We'll continue to delve into these issues in the next lecture with thought experiments about personal identity.



Compositional nihilism is especially important in the Buddhist tradition because it's associated with the doctrine of no-self. In fact, Nāgasena's chariot thought experiment is supposed to help King Milinda understand this doctrine. The king is confused about how it's possible to refer to Nāgasena by his name if, in reality, there is no Nāgasena—if, that is, there's no self that would serve as the referent of the name Nāgasena. Nāgasena explains that people are just like chariots: Even though there's nothing more to chariots and people than their parts, and the whole chariot or person is technically unreal, it's still useful to talk as though there are people and chariots.



Reading

- Gallois, Andre. "Identity over Time." *The Stanford Encyclopedia of Philosophy*. Winter 2016 ed., edited by Edward N. Zalta. https://plato.stanford.edu/archives/win2016/entries/identity-time/.
- Jones, Nicholaos. "Mereological Composition in Analytic and Buddhist Perspective." *Journal of the American Philosophical Association* 7, no. 2 (2021): 173–194.
- Lowe, E. J. *A Survey of Metaphysics*. Oxford University Press, 2002.
- Rose, David, et al. "The Ship of Theseus Puzzle." *Oxford Studies in Experimental Philosophy* 3 (2020): 158–174.

13

MIND SWAPPING AND PERSONAL IDENTITY

The thought experiments in this lecture seek to answer what makes us the same people over time despite the many changes we go through. You'll consider questions about consciousness and examine different views on bodily and psychological continuity. These types of thought experiments can arouse strong beliefs because they force us to grapple with serious questions about who we really are and what matters to us. And even if our intuitions aren't always clear, they do seem important and informative.

Numerical versus Qualitative Identity

In lecture 12, you might have thought that the repaired ship of Theseus was the same as the original ship because of the chain of continuity that linked one ship to the other. This way of thinking about identity over time makes sense because identity is transitive. What connects you today and the person in your baby photos is a complex chain of continuity. That's not a complete answer to the big philosophical question about what makes you the same person over time, but it's a good place to start.

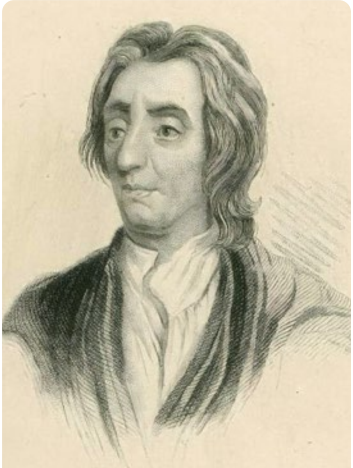
To make sure we're clear about what personal identity is, consider this thought experiment inspired by the ancient comic playwright Epicharmus and suggested by the contemporary philosopher Theodore Sider. Imagine you're on trial for shoplifting and you decide to represent yourself. You claim that you didn't steal anything, despite evidence proving otherwise, because you're a completely different person now. Back when those items were stolen, you had different hair and different interests, and you've changed with age. Therefore, you should be acquitted.

The problem with your defense is that you're confusing two different senses of identity. When we're thinking about personal identity over time, we're primarily concerned with numerical identity—what makes us one and the same person over time, despite all the qualitative changes in identity we undergo. You argued that you're not qualitatively identical to the person who shoplifted, and you should have argued that you aren't numerically identical.

You might think that what makes you the same person over time is something like bodily continuity. After all, we could trace who you are now back to the person who shoplifted and even all the way to when you were just a baby.

It seems obvious, then, that what makes you the same over time is closely associated with—maybe even essentially related to—facts about your body. Of course, you don't have the exact same body you did when you were a baby, or even when you shoplifted, but your body now is appropriately related to your body then, and that's what matters, right? Let's test that hypothesis with a famous thought experiment first introduced by the 17th-century philosopher John Locke.

The Prince and the Cobbler



John Locke

Imagine a prince who's gotten himself into trouble. To escape the consequences of his actions, he arranges what we might call a body exchange with another man, the local cobbler. This case presents a *Freaky Friday*–style consciousness swap. When we're watching a movie like that, we agree with Locke that if the prince's mind is now in the cobbler's body, we think that's the prince, and vice versa. What's important for Locke is continuity of consciousness. When the prince's consciousness—his thoughts, memories, and other aspects of his mind—exchanges places with the cobbler's consciousness, the prince as a person is now in the cobbler's body.

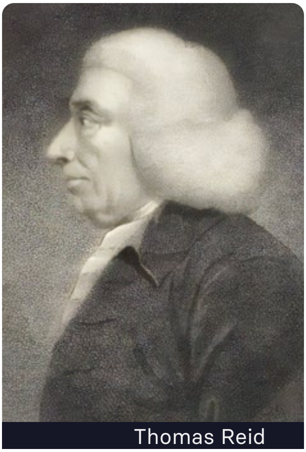
It's important that we track the prince as a person here—not as a man or a human being. Locke notes that "in the ordinary way of speaking, the same person, and the same man, stand for one and the same thing." But this case seems to show that these concepts come apart. The prince is now, we might say, a different human being, a different man—but the same person. The body is what makes the human being, whereas the mind is what makes the person. How do we know that, exactly? Well, Locke notes that after this body exchange, we would think that the individual who would be accountable for the prince's actions would be the one with the prince's consciousness, not the one with the prince's body.

To drive this point home, consider an updated version of the case. As before, the prince and the cobbler somehow swap minds (or exchange bodies). Imagine what happens when the cobbler wakes up and finds himself in the prince's home. When he glances at himself in a mirror, he's startled. He looks different, but he realizes that it's him. It's not him, of course, not his body, but he seems to be looking at himself in the mirror.

As he's puzzling over this, the police arrive to arrest the prince. They break in and throw who they think is the prince down onto the floor. As they handcuff him, the cobbler tells them that they have the wrong man, that he just looks like the prince, but he's really the cobbler.

No one believes him, but we do, because we know there's a distinction here between the body of the prince and the mind of the cobbler. We agree with Locke that the person they've arrested is the wrong person. That person is in a different body, and he's getting away with murder!

Consciousness


Thomas Reid

Thought experiments like the prince and the cobbler are powerfully suggestive. It seems hard to argue with the intuition that personal identity tracks consciousness in something like the way Locke suggests. You might be wondering, however, what exactly *consciousness* means here. The 18th-century philosopher Thomas Reid thought consciousness was memory, and he challenged Locke's view with a thought experiment. In it, a military general remembers some but not all of the events in his life. As a general, he remembered events from when he a was younger soldier. As a younger soldier, he remembered events from his youth. But the general has forgotten events from his youth. Reid reveals a violation of logic—in particular, the principle of the transitivity of identity—because the boy is identical to the soldier, the soldier is identical to the general, but the general is not identical to the boy.

Reid seems to show that there's something deeply wrong with understanding consciousness in terms of memory. Whether that was Locke's intention is questionable, but the lesson is clear: We need to be careful how we understand consciousness when talking about personal identity. Reid's case might also make you question whether we should understand personal identity in terms of consciousness at all. But it's hard to shake the notion that who we are is entangled with our psychology—with facts about what we think, feel, care about, remember, desire, and so on.

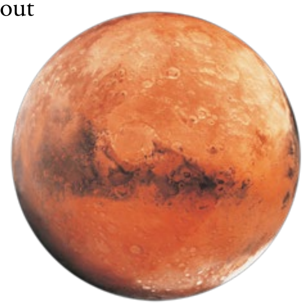
Our intuitions seem to be pulled in different directions, depending on what features of the cases we emphasize. Philosophers like Mark Johnston suggested that our ability to offer counter-descriptions of thought experiments that undermine our initial intuitions calls into question the argumentative use of thought experiments. After all, to draw solid conclusions based on our intuitions, our intuitions would have to be reliable. But, Johnston wonders, how can intuition be reliable if we can be made to react so differently to the same basic case?

We also can't deny that we have genuine reactions that seem to tell us something important about who we are. As the influential philosopher Derek Parfit notes, "By considering these cases, we discover what we believe to be involved in our own continued existence, or what it is that makes us now and ourselves next year the same people." Do you agree with Johnston or Parfit?

Parfit's Teletransportation Case

Consider another thought experiment from Derek Parfit.

You've traveled to Mars before, but always by spaceship, a journey that took weeks. This time, you'll be traveling at the speed of light. All you have to do is push a button, and then you'll lose consciousness before waking up after what will seem like only a few moments, when in fact it will be about an hour. The scanner here on Earth will record all the facts about your body and your brain, but in the process, it will destroy your body and your brain. The machine will then send your information to the replicator on Mars, which will create a new body and brain for you, and it's in that body that you will wake up.



You push the button. You lose consciousness and then wake up in a new room. Instinctively, you examine your new body. Everything is just like it was for you only moments ago. This becomes your routine—your daily commute from Earth to Mars and back again. But one day, you push the button in the scanner, and nothing seems to happen. You ask the attendant what went wrong. They say that it's a new scanner, one that records your information without destroying your body or brain. You're puzzled. If you're here, on Earth, how can you also be on Mars?

Someone in a white coat then tells you privately that, actually, they're having trouble with the new scanner. It's recording and sending your information just fine, and the replicator on Mars will create a new body and brain for you. The problem is that the new scanner is damaging people's cardiovascular system. The person tells you that, although you'll be perfectly healthy on Mars, here on Earth you should expect cardiac failure within the next few days.

Later, in the control room, you see yourself on a screen. It's almost like looking into a mirror, but not quite. Although you're speechless, the image of you on the screen begins to talk.

Bodily versus Psychological Continuity

The first part of the story involves what Parfit calls simple teletransportation. If you saw it as simply a way of traveling, then your intuition might be that what makes us the same over time is psychological continuity, not bodily continuity. However, if you think that the body is what makes you the same person over time, then simple teletransportation isn't a way of traveling at all. It's a way of dying. In that view, the replica on Mars is someone else.

In the second part of the story, which Parfit calls the branch-line case, there are two of you at the same time. If we stick with the view that psychological continuity is what makes someone the same over time, we run into a problem. We have to say that the person on Earth is identical to the person Mars, but that's obviously false, because they're two different people in two different places. The replica is qualitatively, not numerically, identical to you.

Parfit's case presents problems for both the bodily continuity view and the psychological continuity view. It also raises new questions. If the branch-line case happened to you, would you feel like you were dying? Or would you feel like you'd be living on, despite your impending death? How would the person on Mars feel? What would your death mean to each of you?

Parfit thinks the power of these thought experiments is that they get us to think clearly about things that really matter to us, whereas Johnston worries that our intuitions in cases like this are unreliable—or at the very least, that the lessons we draw aren't necessarily the right lessons. Who's right?

Reading

- Gordon-Roth, Jessica. "Locke on Personal Identity." *The Stanford Encyclopedia of Philosophy*. Spring 2020 ed., edited by Edward N. Zalta. https://plato.stanford.edu/archives/spr2020/entries/locke-personal-identity/.
- Olson, Eric T. "Personal Identity." *The Stanford Encyclopedia of Philosophy*. Summer 2022 ed., edited by Edward N. Zalta. https://plato.stanford.edu/archives/sum2022/entries/identity-personal/.
- Parfit, Derek. *Reasons and Persons*. Oxford University Press, 1986.
- Perry, John, ed. *Personal Identity*. 2nd ed. University of California Press, 2008.

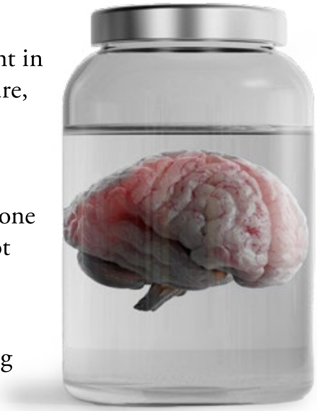
Table of Contents

14

WHO ARE YOU AFTER A BRAIN TRANSPLANT?

The exploration of identity continues with what it means to survive and how identity factors into what we consider survival. In this lecture, you will revisit Derek Parfit's teletransportation case from this new perspective and consider some other thought experiments to gain insight into the criteria of personal identity that we actually use.

Perry's Divided-Self Case



Three individuals—Brown, Jones, and Smith—went in for their annual brain rejuvenations. In the procedure, the brain is removed and put into a machine that scans it completely. Then all the information is encoded into a new brain made of fresh, healthier matter, which is then put back into the skull. Everyone who has this done is cognitively rejuvenated, but not psychologically changed—they retain all the same memories, dispositions, and character.

During the procedures for the three men, something went wrong. Brown's brain and Smith's brain were ruined beyond recovery. To cover up the problem, Jones's brain was used for all three procedures. So, Brown and Smith both got Jones's brain. Jones also got his brain back, but tragically, he died from heart failure before the end of the operation.

When the two survivors woke up, they were confused. They felt cognitively rejuvenated, but their bodies had undergone radical changes. Yet they remembered being Jones. Both claimed to be Jones—even though Jones had died hours earlier.

Have Smith and Brown died? Or do they live on? Has Jones survived, despite his death? Are Smith and Brown new people, or are they both somehow Jones?

This thought experiment is the divided-self case, from contemporary philosopher John Perry. It raises a lot of questions, and it's similar to an even more famous thought experiment that has driven much of the debate about personal identity for the past half century. Let's explore Derek Parfit's fission case.

Parfit's Fission Case

You're one of three identical triplets, and you've all been in a terrible accident. Your body was ruined beyond repair, but your brain was spared. Your siblings' bodies are in good shape, but their brains have been destroyed. You were all rushed to a hospital, and the head of surgery—a pioneering neurosurgeon—decided to perform a daring procedure.

Your undamaged brain was removed and divided in half. The right hemisphere was transplanted into one of your siblings, and the left hemisphere went into your other sibling. The thought was that a split-brain transplant would give you two chances to survive, rather than just one.

Imagine four ways things might unfold:

In scenario 1, neither transplant procedure works; despite the surgeon's efforts, they just couldn't save you. In scenario 2, the left-hemisphere transplant procedure works. In scenario 3, the right-hemisphere transplant procedure works. In scenario 4, both hemisphere transplant procedures work.

In scenario 1, it's clear that you don't survive. In scenarios 2 and 3, you've managed to survive despite the odds. And even though you've changed a lot, you'd still be you, and you'd go back to your family. They'd grieve the loss of your siblings, but wouldn't they also think you had survived?

What about scenario 4? Your family might be confused, but would they think that you had died? You've managed to survive twice! How can you die by surviving?

And where are you now? It doesn't look like we can say that you would cease to exist in scenario 4. But, of course, we also can't arbitrarily stipulate that you survive only in the left-hemisphere transplant, or only in the right-hemisphere transplant, either. There's no fundamental difference between what makes you identical with one survivor as opposed to the other. It seems, then, that in scenario 4 we should say that you survive both transplants. But the problem is that we can't say that either of the survivors is actually identical to you—that would defy logic.

If preoperative you is identical to both survivors, then we'd have to say that the two survivors are identical to each other. But they can't be, because they're two different people, located in different places, now having different thoughts and feelings. We're left with a puzzle. On the one hand, it seems like you have in fact survived twice. On the other hand, it doesn't seem like you could have survived twice, if survival requires identity. But does survival require identity?

Survival versus Identity

Parfit thinks that the lesson of his fission case is that survival isn't the same as identity. In scenario 4, you survive as both people, but you're not strictly identical to either. Parfit thinks the most important lesson is that what matters to us isn't identity at all, but survival by any means. The whole time we've been thinking about personal identity, we'd been thinking that survival requires identity. But maybe we were wrong. Does it really matter to you that you'll be identical to who you were before the transplant, or does it matter to you that you'll survive?

Normally, surviving is a matter of continued existence as the same person. But in the fission case, what matters to us isn't identity at all, but survival in the sense that there's someone around in the future who is psychologically continuous with or connected to us in important ways. You are psychologically connected to someone if, for instance, you seem to remember things they experienced, or if you understand your actions as following through on the plans and intentions that they had for themselves in the future. And you are psychologically continuous with someone if you are linked in a chain of direct psychological relations with them.

Like identity, psychological continuity is a transitive relation. If A is psychologically continuous with B and B is psychologically continuous with C, then A is psychologically continuous with C. Ordinarily, A will be the same person over time as B only if A's psychological composition is related to B's psychological composition by an overlapping series of psychologically connected stages that, as the philosopher Michael Rea puts it, "looks like the development of a single mind over time." But in extraordinary cases, questions about identity become not merely murky but unanswerable.

Recall the branch-line teletransportation case from lecture 13. Your body and brain were replicated on Mars, and on Earth your cardiovascular system was fatally damaged. The person you saw on the screen wasn't you but someone who was psychologically continuous with you. You had survived, in Parfit's sense, on both planets, but you were also going to die on Earth in a few days. We might say that your wish was to not die, not to be identical to your replica on Mars. What you care about is still surviving, not identity.

Let's say that the teletransportation company offers you the chance to go through the old scanner again. This procedure will copy the damage to your cardiovascular system, but on Mars there's a heart surgeon who can fix it. But there's only a 10% survival rate, so they want to make a dozen replicas of you on Mars, in the hope that at least one is successful. What would you do?

Obviously, you can't be identical to all 12 replicas, but you would be psychologically continuous with each of them. And although it's true that most of them would die in a few days, it's also true that, unless you take them up on their offer, you're going to die in a few days anyway. If you care about identity, it seems like you should choose to stay on Earth and live out your final days. But if you care about survival, it seems like you should enter the old scanner and take your chances.

If you think that identity doesn't matter but survival does, then you should also think that, in the original branch-line case, whether you're identical to the replica on Mars doesn't really matter. What really matters is that there's someone who is psychologically continuous with you who will survive. That might sound counterintuitive to you. Parfit thinks these thought experiments can change our intuitions and radically change our sense of ourselves.

What they reveal is that there's no deep fact about who we are—no further fact beyond psychological continuity. The truth is that there's no special thing called personal identity that we want to preserve. You might find that troubling, but Parfit finds it liberating, even consoling.

The philosopher Michael Rea thinks we do care about identity. To show this, he adds a consideration to Parfit's thought experiment: Suppose that, just before the hemisphere transplant, you're told that the survivor of one transplant will receive "your heart's greatest desire," but the other one won't.

Rea thinks you'd hope to be identical to the one who gets your heart's desire. He fears that he would be identical to neither survivor and that the two resulting persons would be other people, not him. So, Rea thinks what matters in survival is identity after all.

He also thinks that the psychological continuity and survival approach is wrong because it leads us to the view that there is, in fact, no self at all. In a way, this is the view Parfit himself endorses. And in other ways, this no-self view is endorsed by the great 18th-century philosopher David Hume and by noteworthy Buddhist philosophers like Vasubandhu and Buddhaghosa, to name just a few. The no-self view, very roughly, is that as persons, we're merely bundles of thoughts; there's no thinker, only thoughts; there's no subject, only experiences.

David Hume

The power of Parfit's fission case is that it's supposed to reveal what we really care about and what we strongly believe. It's supposed to uncover "the criteria of personal identity that we actually use." And part of what gives us such trouble in the fission case is that it's obvious that we can't be identical to both of the survivors.

One Person, One Place

Philosophers Sara Weaver and John Turri have explored the intuition that one and the same person can't be in two different places at the same time. They call this the one-person-one-place rule. Interestingly, they've found that not everybody has this intuition, with some maintaining that one and the same person was in two different places at the same time. This was the case even when the person underwent significant bodily changes at those two locations.

To question this rule, Weaver and Turri asked participants to consider variations on some of the thought experiments we've discussed. In the split-brain transplant study, for instance, participants judged that the individual undergoing the surgery survives in both recovery rooms. And in the teletransportation case, participants judged that one and the same individual was on two different planets at the same time.

Weaver and Turri think these findings provide insights into "the criteria of personal identity that we actually use." Philosophers have insisted that it's a violation of the logic of personal identity for us to say that one and the same person exists in two different locations at the same time. But if our natural tendency is to judge that it's possible, why can't we say that what's really powerful about Parfit's cases is that they reveal not that what we really care about is survival or identity, but that the criteria of personal identity that we actually use isn't quite what we thought it was?

Reading

- Parfit, Derek. "The Unimportance of Identity." In *Identity*, edited by H. Harris. Oxford: Oxford University Press, 1995.
- Perry, John. "Can the Self Divide?" *Journal of Philosophy* 69 (1972): 463–488.
- Rea, Michael. *Metaphysics: The Basics*. 2nd ed. Routledge, 2020.
- Weaver, Sara, and John Turri. "Personal Identity and Persisting as Many." *Oxford Studies in Experimental Philosophy* 2 (2018): 213–242.

Table of Contents

15

WHO ARE YOU RIGHT NOW?

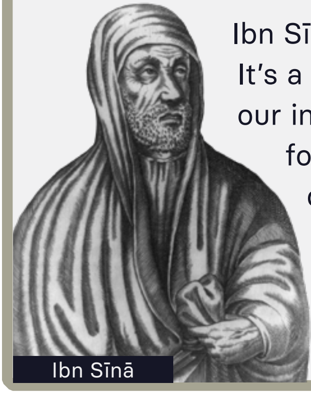
This lecture addresses several questions, including whether self-awareness and bodily awareness are the same thing, and whether our minds can extend into the external world. It looks at how Islamic, Daoist, and modern philosophers have used thought experiments to answer such questions.

The Floating Man

Imagine that someone is created in an instant, but that once they've popped into existence, their senses are blocked off, so they can't see or hear or smell or taste or otherwise sense the external world. Imagine also that they're floating in a void, so they don't touch anything external, and their limbs are stretched out, so they don't even touch themselves. In this state, this newly created person would be deprived of all sensory input. Would they be aware of their own existence?

This floating-man (sometimes also called the flying-man) thought experiment comes from the great Islamic philosopher Ibn Sīnā, who flourished from the 10th to the 11th century. In the European tradition, Ibn Sīnā is better known as Avicenna, which is a Latinized corruption of his name. Ibn Sīnā offers several versions of his thought experiment, but the most important version is found in his *On Psychology*, which is a section of his larger work *The Book of Healing*.

In this section, Ibn Sīnā explores the nature of the soul. Before he considers his floating-man thought experiment, he critically examines previous definitions of the soul, including Aristotle's account. According to Aristotle, the soul and the body are importantly intertwined, but Ibn Sīnā questions whether the soul is necessarily related to the body, and it's in this context that he brings up his famous thought experiment.



Ibn Sīnā

Ibn Sīnā calls the floating man a pointer. It's a thought experiment that sparks our intuitions and helps us see the truth for ourselves. After considering the case, we're supposed to get it on our own, without first having to learn the nuances of logical argumentation.

According to Ibn Sīnā, the case focuses our attention on self-awareness. In this scenario we're supposedly self-aware despite having no bodily awareness at all, so it seems like self-awareness is more fundamental to who we are than bodily awareness. The case is supposed to help us see that our bodies are more like clothes. We're used to taking our clothes off and discarding them, but we're not at all used to discarding our bodies, and so we tend to associate our bodies with who we are, even though they aren't essentially linked. This view helps us see that this insight about self-awareness is supposed to lead to another insight about who we are essentially.

This newly created person can affirm that they exist even without affirming the existence of their body. What this means, Ibn Sīnā suggests, is that our essence (*dhāt*) has nothing at all to do with our bodies. At our core, we're not corporeal beings after all; we're incorporeal. He seems to think that if we're aware of one thing (our existence) but not another (our body), those two things can't be identical.

This line of thought, however, is faulty. To see why, imagine that you're aware of what gold looks like, but you're not aware that gold is element 79 on the periodic table. From that fact, following Ibn Sīnā's lead, you might falsely conclude that gold and element 79 are different things, even though they're actually the very same thing.

Maybe Ibn Sīnā was trying to show that Aristotle's way of thinking about the soul is wrong. For Aristotle, the body and the soul are so intimately connected that the only way we can grasp the soul is in relation to the body. But how do we know that this person floating in the void grasps the existence of their soul rather than the existence of their body? Isn't it possible that the floating man is self-aware because he is aware of his body in some extrasensory way? Ibn Sīnā's answer seems to be no—that the only way we can become aware of our bodies is through the senses—and that seems to be why he carefully crafted the floating man to exclude the senses completely.

Maybe who we are is not essentially connected to our bodies. Still, we might wonder: How much, or what exactly, do we grasp about ourselves through introspective self-awareness?

Zhuangzi's Butterfly Dream

Consider Zhuangzi, a Daoist philosopher who flourished in the 4th century BCE. He was dreaming one night that he was a fluttering butterfly, completely unaware of being Zhuangzi. He awoke suddenly, and he was Zhuangzi once again. He wondered if it had been Zhuangzi dreaming of being a butterfly, or the butterfly dreaming of being Zhuangzi. Surely there must be some difference between them!

A natural way to read this thought experiment is as a skeptical dream argument that calls into question our ability to know what's real and true. How can we tell that we're not in a dream? But perhaps his point is that we might not know who we are. After all, Zhuangzi is confused about who he is in the thought experiment—and you might be, too.

The scholar Hans-Georg Moeller notes that the 3rd-to-4th-century Daoist thinker Guo Xiang did not read the thought experiment like this. He thinks it's important that Zhuangzi doesn't remember his dream at all. Guo Xiang thinks that Zhuangzi is wondering about something he doesn't clearly remember, not doubting his existence. For Guo Xiang, not remembering the dream means that there is no "I" that crosses the boundaries of the dream. As Moeller puts it, "there is no continuous substance underlying the different stages of dreaming and being awake."

The original butterfly dream passage ends with a curious statement: Zhuangzi says that “this is called the transformation of things.” The philosopher JeeLoo Liu suggests that the comment is about how our self-identity (who we are) is something that’s unstable, given that things are constantly changing. At one stage in this process of transformation, we might be a human being, whereas at another, we could be a butterfly. But all these transformations are aspects or elements of the Dao, the way things really are, just manifested in different ways, different shapes.

We could understand this to mean that there’s some underlying thing that we are that transforms from one stage to the next, like a caterpillar becoming a butterfly. But Guo Xiang’s reading points out that the thought experiment isn’t about this kind of transformation. Instead, it’s about transformation as a kind of gestalt shift, where the two ways of being are closed off from each other, not related at all. And yet, Guo Xiang insists, those two unrelated things are equally real.

What’s really important about the distinction between dreaming and being awake, according to Guo Xiang, is that it’s analogous to the distinction between being dead and being alive. When Zhuangzi is asleep and doesn’t know about being anything other than a butterfly, that’s like being dead, and when he is awake and doesn’t know about being anything other than Zhuangzi, that’s like being alive. Both ways of being, according to Guo Xiang, are authentic—one is just blocked off from the other. And so, the idea is, if being a butterfly in a dream is just as authentic as being Zhuangzi while awake, then being dead is just as authentic as being alive.

Otto’s Notebook

Consider Otto now. Otto suffers from Alzheimer’s disease, but he’s resourceful, and he’s developed the habit of relying on a notebook filled with important information. He relies on his notebook the way his friend Inga relies on her long-term memory. Today, Otto and Inga plan to meet at the Museum of Modern Art in New York City. Inga remembers where it is located, but Otto consults his notebook. The notebook, which is external to Otto, seems to play the same functional role as Inga’s internal neural structure.

The information about the museum that Inga can retrieve from long-term memory counts as mental, as something that's part of her mind. It's a dispositional belief—a belief she has but doesn't actually consider until she retrieves it from memory. Can we also say that the information that Otto gets from his notebook is a dispositional belief?

Philosophers Andy Clark and David Chalmers use this thought experiment to make the case that the mind can extend into the external world. After all, when we notice that Inga's neural structure and Otto's notebook play the same functional role, then doesn't it mean that Otto's mind extends into the external world?

And can't we make this same point about important things that make up the kind of person we are—like what our favorite restaurant is, the names of our children, anything that we might otherwise remember the old-fashioned way? We don't normally say that a notebook that plays this sort of role in someone's cognitive life is actually part of their mind, but does this case convince you that we should?

Consider Rick. He reads every copy of *Scientific American* and often consults back issues on his phone when he has scientific questions. Does Rick's mind extend into the magazine? Clark and Chalmers think that it doesn't. They think the difference is that Otto's notebook, but not Rick's magazine, is part of a causally coupled system. When Rick reads *Scientific American*, he thinks about it or in response to it, whereas when Otto consults his notebook, his cognition is bound up with the notebook.

According to Clark and Chalmers, the only relevant difference between Otto and Inga is that her long-term memory is internal while his notebook is external. Both are portable, reliably available, and regularly consulted. None of this, other than being portable, is true of Rick's magazine; it's not part of an integrated cognitive system like Otto's notebook is.

Now consider Jerry. Jerry loves his encyclopedias. They're in his apartment, and when he has a question, he consults them. Like Otto, he's developed the habit of cognitively relying on those encyclopedias. And just like Otto's notebook, Jerry's encyclopedias function a lot like Inga's long-term memory.

Let's say that Jerry and Inga each have a question about polar bears. For Inga, that information is a dispositional belief that polar bears are large predators that live in the Arctic Circle. But it looks like we can say the same now for Jerry about the information he retrieves from his encyclopedias, right? It looks as though Jerry's mind, just like Otto's, extends into the external world. But that's absurd, right? Jerry's encyclopedias aren't part of his mind!

A dispositional belief is a belief that someone is prone to have under the right circumstances but isn't yet considering. Once the person considers it, it becomes an occurrent belief—a belief that's being considered at the moment.

For many, the case raises serious problems with the idea that the mind can extend into the world. Even though digital resources like Google allow us to outsource some of our cognitive work in the same way a calculator does, it might seem that we've committed a category error by extending the mind this far. We don't need to talk about the extended mind to talk about how our minds utilize noncognitive resources like notebooks, computers, search engines, calculators, and encyclopedias.

So, what does Otto's notebook tell us about the self? We might say that it supports a narrative conception of the self, according to which we are, in some sense, the stories we tell about ourselves. And Otto's story about himself and the way he lives his life suggest that his self extends into the external world, because his notebook plays a crucial role in his self-conception, in the story he tells about himself.

This narrative identity interpretation of the case does not force us to resolve problems of demarcation. For philosopher Daniel Dennett, this is because we make a category mistake when we demand to know what the self really is or what the boundaries of the self really are. But this doesn't mean that we need to endorse a kind of anti-realism. Philosopher J. David Velleman, for instance, says that the stories we tell about ourselves provide genuine boundaries for the self, even though those boundaries shift in response to the narrative. Otto's story includes his notebook, and maybe Jerry's includes his encyclopedias.

Another way to think about Otto's case is that it tells us something about our practical identities—who we are as persons. In a minimal sense, the concept of a person captures what's important about being human. By caring about his notebook, by seeing himself bound up in it, and by thinking of himself in terms of it, Otto makes his notebook part of who he is in some important sense. Without it, he'd be a different kind of person.

Read like that, Otto's case tells us that who we are, in terms of our practical identity, goes beyond the skin and the skull and even the mind to extend into the world.

Reading

- Adamson, Peter, and Fedor Benevicht. "The Thought Experimental Method: Avicenna's Flying Man Argument." *Journal of the American Philosophical Association* 4, no. 2 (2018): 147–164.
- Clark, Andy, and David Chalmers. "The Extended Mind." *Analysis* 58, no. 1 (1998): 7–19.
- Moeller, Hans-Georg. "Zhuangzi's 'Dream of the Butterfly': A Daoist Interpretation." *Philosophy East and West* 49, no. 4 (1999): 439–450.
- *Zhuangzi: The Essential Writings; With Selections from Traditional Commentaries*. Translated with Introduction by Brook Ziporyn. Hackett Publishing, 2009.

Table of Contents

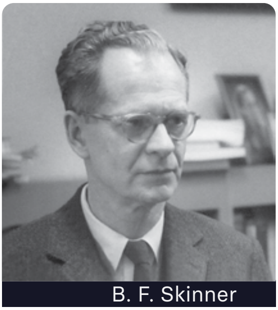
16

EXPLORING THE MYSTERIES OF CONSCIOUSNESS

In this lecture, you will consider some thought experiments about consciousness and the mind. You will also explore whether subjective experience can be explained through scientific study of the mind or whether it's just an unsolvable mystery. And you will look at some influential arguments regarding intelligent machines.

Logical Behaviorism

J. B. Watson



B. F. Skinner

If you wanted to study someone's mind, you might start by studying their behavior. This is the approach that the psychologist J. B. Watson championed in the early 20th century. He suggested a new method that focused on observable and measurable behavior, similar to the way that other sciences work.

To do this, Watson suggested that we should look at the relationship between sensory stimuli, like a pinprick, and our behavioral responses to them, like wincing. The psychologist B. F. Skinner famously picked up on this and adopted a radical form of behaviorism. For Skinner, to be truly scientific in our study of the mind, we can't postulate anything unobservable to explain why you wince when your finger is pricked with a pin.

If we study the mind like this, what can we say about mental states? One suggestion, popular with philosophers in the 1950s and 1960s, was that we could analyze mental states—and internal psychological descriptions in general—in terms of behavioral dispositions. In this view, when we say things like "I'm thirsty," that's just an abbreviated way of talking about our actual or potential behavior. These abbreviated ways of talking would then be captured more completely by a description of what we're actually doing and what we would do in various circumstances.

This approach, known as logical behaviorism, allows us to talk about internal mental states without needing to postulate the existence of some unobservable mental realm beyond the reach of science. An interesting objection to this approach was offered by the philosopher Hilary Putnam in 1965. His thought experiment serves as a counterexample.

Putnam's Super-Spartans

Imagine a community in which the adults have the ability to successfully suppress all involuntary pain behavior. These super-Spartans feel pain, but they don't do any of the normal things associated with pain. They don't have dispositions for pain behavior.

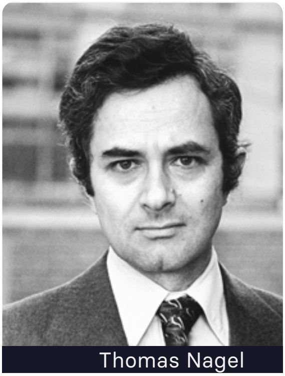
Now imagine super-super-Spartans. They don't even talk about pain. But they still feel it, and we might even imagine that some of them have private thoughts about it. For them, there's such an impermeable divide between their internal lives and their external behavior that they "do not even admit to having pains."

If logical behaviorism were true, and if behavioral approaches could give us everything we want when studying the mind, then super-super-Spartans would be impossible. That's because a description of their actual and potential behavior is the same as a description of someone who's not in pain, even though the super-super-Spartan is, in fact, in pain. Putnam's thought experiment is powerful because it strongly suggests that behavior can't be the whole story about the mind.

It seems like the best way to study someone's mind would be to observe them in the most fine-grained detail—internally and externally. This would include talking to them, noting what they do and say, conducting experiments, and exploring their bodies and brains with advanced technology. That way, we could explain what's going on in the brains of those super-super-Spartans. If they really were in pain, then we should be able to observe the neurological mechanisms associated with feeling that pain, as well as the physical causes of it. And that's basically the approach we find in contemporary cognitive psychology for studying the mind.

Subjective Experience

Imagine that a team of scientists has been studying your mind for your whole life, and they know everything there is to know about you from their objective point of view. They have a scientifically grounded account of your mind, but how could they know what it's like to be you?



Thomas Nagel

According to the philosopher Thomas Nagel, subjective experience is something that objective observation can't penetrate. This subjective experience we're talking about here is associated with what's now called phenomenal consciousness: the private, internal, qualitative side of first-person experience.

If you're inclined to think that the scientific study of the mind can give us everything we need, then you might endorse some form of physicalism or materialism and hold that we can explain subjective experience in physical terms.

Maybe we can explain subjective experience by appealing to neurological facts about the brain in the same way that we can explain the wetness of water by appealing to facts about its molecular structure. It's not some unsolvable mystery.

Not everyone agrees with that notion. To see why, let's explore some influential thought experiments that make a powerful case that physicalism is false. We'll start with one from Gottfried Wilhelm Leibniz.

Leibniz's Mill Argument

Leibniz asks us to "imagine a machine whose structure makes it think, sense, and have perceptions," which is what he calls conscious experiences. This machine would be made of physical material, and it would work like other machines. Imagine enlarging this machine, "keeping the same proportions, so that we could enter into it," in the same way we might enter a mill. If we were to inspect the interior of the machine, we'd find the materials, mechanistically arranged, but we'd never find anything to explain conscious experience, would we? We can extend this idea to the human brain and imagine walking around in it.

Leibniz's thought experiment, sometimes called the mill argument, seems to show that no matter how detailed it is, knowledge of the material and mechanistic structure of the brain, or a thinking machine, wouldn't give us knowledge of consciousness. The philosopher Daniel Dennett notes that we might draw two very different conclusions from the thought experiment. Leibniz clearly wants us to draw the conclusion that "consciousness couldn't be a matter of 'machinery.'" That's a metaphysical conclusion that's supposed to show that mechanistic materialism, or physicalism, is false.

On the other hand, we might conclude that no matter how much we study the brain, "we'll never understand the machinery of consciousness." That's an epistemological conclusion that's supposed to show us something about the limits of understanding.

It's also worth questioning whether we really could observe the complex interactions of the brain in the way the thought experiment suggests. If consciousness emerges from mechanistic interactions in the brain, it wouldn't manifest throughout the brain, and we don't have any good reason to think we'd observe those interactions in their totality.

Mary the Color Scientist

Consider Frank Jackson's knowledge argument, driven by his thought experiment: Mary is confined to a black-and-white room. Having lived there her entire life, she's never seen any colors beyond black, white, and shades of gray. Mary learns "everything there is to know about the physical nature of the world." We might say that if physicalism is true, then Mary "knows all there is to know."

Since Mary knows every physical fact there is to know, she knows exactly what happens when someone sees something that's red—a tomato or an apple, for instance. She's never seen the color red, but she knows everything there is to know about it. One day, Mary leaves the room and sees a brilliant red apple hanging from a tree just outside her door. Does Mary learn something new? If so, then physicalism has to be false. She can know everything objective and measurable about red, but still fail to grasp what it's like to experience it.

Jackson tries to show that the nature of subjective experience is something physicalism can't possibly account for. If it could, then Mary would be able to know what those experiences are like without experiencing them for herself. What's so powerful about this thought experiment is that the intuition that Mary learns something new is hard to resist.

Dennett argues that Mary wouldn't learn anything. What misleads us to think she does is the imaginary scenario itself. We can't really imagine that Mary knows "everything physical." How could she know everything about the physical world if, for instance, the Heisenberg uncertainty principle and the Schrödinger equation accurately describe reality? And how reliable could your imagination be about what Mary knows?

In response, the philosopher Philip Goff suggests that Mary doesn't have to know everything physical. If physicalism is true, what we care about in this case "is neuroscience, not fundamental physics," because neuroscience explains "facts about the human mind."

Philip Goff

All we need to do is "imagine that [Mary] has full knowledge of color experience provided by [a] complete and final neuroscience." Then we can realistically imagine Mary's scenario.

This thought experiment is extremely powerful even if Jackson's knowledge argument fails. It highlights what's been called the hard problem of consciousness, which is about whether and how the scientific study of the mind can explain subjective experience and phenomenal consciousness.

Searle's Chinese Room

John Searle

When we considered Leibniz's thought experiment, we didn't question his suggestion that there could in fact be a thinking machine, but we might wonder whether machines could have minds that really understand anything at all. In 1980, the philosopher John Searle devised a thought experiment intended to show that even the most sophisticated computers couldn't really understand anything—so they wouldn't be intelligent after all.

Imagine a room with someone in it who doesn't speak Mandarin Chinese. There's an input slot on one wall and an output slot on another, and in the middle of the room there's a basket full of cards with Mandarin Chinese characters on them. On the back wall, there's a big chart. It tells the person in the room how to match up Mandarin Chinese characters to other Mandarin Chinese characters, effectively providing instructions for what characters count as output for the characters that come in as input.

一二三四五六七八九十
百千円時分秒朝昼夕夜
日月火水木金土年週東
西南北雨雪左右上下大
中小新古白黒赤青緑黄
茶花草田畑森林石岩人
男女子手足目耳鼻口魚
肉米歩走食見聞市町村
Output
Input

Outside the room, there are native Mandarin Chinese speakers who write characters on cards and then slip them through the room's input slot. When the person inside receives a card, he follows the instructions on the big wall chart and finds the corresponding output card, which he then slips through the output slot. In this way, he carries on conversations with the native Mandarin Chinese speakers outside the room. But, of course—and this is key—the person doesn't understand anything that comes through as input or leaves as output. He just follows the functional instructions carefully. We have the mere appearance of understanding.

Computer scientists objected when Searle first presented the thought experiment, arguing that it's not the person's understanding that matters, but the whole system, which includes the person plus the room with its instructions, cards, inputs, outputs, and everything else. Searle calls this objection the systems reply.

Searle's view is that it's intuitively implausible to say that the room, understood as a complex system, understands Mandarin Chinese. He notes that he could get rid of the room altogether and just memorize all the instructions and do everything in his head. He'd follow all the rules, but he still wouldn't understand Mandarin Chinese, would he?

The Chinese room is one of the most famous thought experiments in the philosophy of mind and computer science, and it's generated decades of important debate about mind and consciousness. The takeaway from this lecture should be how a well-conceived thought experiment can push the boundaries of our knowledge. This is even more important when the topics are as difficult as the mind and how we experience the world.

Reading

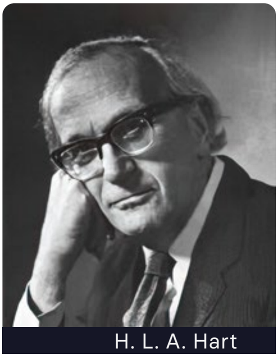
- Dennett, Daniel. *Consciousness Explained*. Boston: Little, Brown and Co., 1991.
- Ibn Ṭufayl. *Hayy Ibn Yaqzan: A Philosophical Tale*. Translated by Lenn Evan Goodman. Updated ed. Chicago: University of Chicago Press, 2009.
- Jackson, Frank. "What Mary Didn't Know." *Journal of Philosophy* 83, no. 5 (1986): 291–295.
- Nagel, Thomas. "What Is It Like to Be a Bat?" *Philosophical Review* 83, no. 4 (1974): 435–450.
- Searle, John. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (1980): 417–424.

17

WHEN ARE YOU MORALLY RESPONSIBLE?

This lecture looks at various ways the word *responsibility* is used. You will consider what it means to be morally responsible and whether that requires the ability to do otherwise. You will also look at how a specific kind of thought experiment called a Frankfurt case challenges traditional views about moral responsibility.

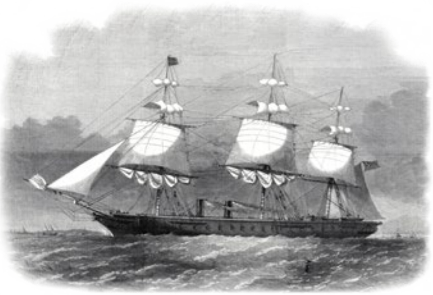
Causal versus Legal Responsibility



H. L. A. Hart

The 20th-century philosopher of law H. L. A. Hart tells the story of Captain X. As a sea captain, he was responsible for the safety of everyone aboard his ship. On his final voyage, however, he behaved irresponsibly and ended up running the ship aground. Some say that he lost his mind, but psychiatrists disagree. They say he was responsible for behaving irresponsibly. Captain X claimed that an unforeseen storm was responsible for the shipwreck, but the courts found him responsible for what happened. To this day, he's still alive and responsible for the deaths of many.

This thought experiment is meant to highlight the various ways we talk about responsibility. When Hart tells us that Captain X is responsible for the safety of everyone aboard his ship, he's talking about role responsibility. In his role as captain of a ship, Captain X has an obligation and duty to ensure their safety. The problem with Captain X is that he failed to fulfill his role in a responsible way.



Captain X claimed that a storm was responsible for what happened. The issue here is causal responsibility, and the important question is this: Was the storm the real, or primary, cause of the shipwreck or was Captain X?

The court decided that he was both causally and legally responsible. So, it judged that he failed to meet his legal obligations and determined that, because of this failure, he deserved some kind of punishment.

Captain X could have responsibly decided that the best bet for surviving the storm was to sail directly through it. In that case, the court might have determined that, although he was causally responsible for the shipwreck, he wasn't legally responsible because he did his best to ensure the safety of everyone aboard.

But some people heard that Captain X had suffered from a mental lapse that rendered him temporarily irresponsible. If so, he might have been causally responsible for the shipwreck but not legally responsible, because he would have lacked what Hart calls capacity responsibility, which refers to possessing the psychological capacities to make good choices and exercise self-control. If someone is responsible in this sense, they would have the capacities necessary to be legally responsible.

Moral Responsibility

Captain X was also found to be morally responsible. Moral responsibility is about meeting or failing to meet our moral obligations and then being responsible for that fact—that we either did or did not do what we were supposed to, morally speaking. Like legal responsibility, moral responsibility requires capacity responsibility. Unlike legal responsibility, however, it isn't determined in a court of law. Someone like Captain X can be morally responsible for something even if no one else knows about it.

Moral responsibility is also about living up to our moral obligations, maybe even exceeding them. If Captain X did everything in his power to save the ship and sacrificed his own life to make a sharp turn that avoided the storm, he would also be morally responsible, but instead of being blameworthy, he would be praiseworthy, even if no one knows what he did.

The power of Hart's thought experiment isn't just that it helps clarify the various ways we use the word *responsibility*; it also—and more importantly—helps us identify what philosophers mean when they talk about moral responsibility.

It also raises a question about the relationship between moral responsibility and capacity responsibility. What exactly are the capacities we need to qualify as morally responsible agents?

The Principle of Alternative Possibilities

For a long time, the traditional view among philosophers was that the most important capacity was "the ability to do otherwise"—the ability to exercise control over our actions in such a way that we could have avoided performing them. What if the storm was unavoidable and Captain X had to steer the ship directly into the storm?

The intuition that, if he couldn't have done otherwise, Captain X wouldn't have been responsible for the shipwreck is captured by what's known as the principle of alternative possibilities, or PAP. This principle tells us that we're only morally responsible for our actions if we could have avoided doing them. We might also need other capacities, but the point is that we need the ability to do otherwise to be morally responsible for our actions. But is that true?

Imagine that you're driving home from a friend's house. You never go above the speed limit, and you're generally careful, taking the route that allows you to make only right turns, because they're safer than left turns. Everything goes smoothly, and you get home. What if, the next morning, you discovered that your car's steering system was broken such that you couldn't make left turns? What would you think about your trip the previous night? You might think that you were responsible for driving home safely and conscientiously. After all, you got home exactly like you planned, and you drove the car to the best of your abilities.

This case comes from the contemporary philosopher John Martin Fischer. It suggests that maybe we don't need the ability to do otherwise after all. Maybe the principle of alternative possibilities is false. Did you really need to turn left to be responsible for turning right?

This control over our actions that involves the ability to do otherwise is what Fischer calls regulative control. You would have had regulative control over making a right turn if your car's steering system had been working properly. You would have had the ability to turn left—the ability to do otherwise. Fischer contrasts regulative control with what he calls guidance control, which requires that you act based on your intentions. You turned right because you intended to.

Though suggestive, this thought experiment is flawed, as it doesn't actually show that all we need is guidance control. After all, you still could have tried to turn left; you could have tried to do otherwise. Because of that, we might think that you're responsible for driving home safely. So, we need a better thought experiment to show that the principle of alternative possibilities is false.

Frankfurt Cases

In an incredibly influential 1969 article, the philosopher Harry Frankfurt introduced a certain kind of thought experiment specifically for the purpose of showing that the principle of alternative possibilities is false.

Frankfurt introduces two characters, Jones and Black. Black really wants Jones to do something, and he's willing to do whatever it takes to make sure Jones does it. But Black would prefer to stand back and see whether Jones will act on his own. Black doesn't want to interfere unless it's absolutely necessary. So Black waits until just before Jones is going to make up his mind about what to do. And if Black notices that Jones is not going to do what he wants, he'll intervene and force Jones to act how he wants him to. As things turn out, however, Jones does exactly what Black wants, and Black doesn't have to intervene.

In this Frankfurt scenario, we have circumstances that in no way bring it about that an agent performs an action, even though those very circumstances make it impossible for the agent to avoid performing that action. In this case, Black doesn't actually force Jones to do anything. Jones does what he does on his own. But Black creates circumstances that make it impossible for Jones to avoid doing what Black wants him to do, because Black would intervene if Jones were to choose otherwise.

Now imagine that Black wants Jones to vote a certain way in the upcoming election. Black implants a device in Jones's brain that will allow him to monitor his thoughts. The only way that Black can disrupt Jones is if Black chooses to force Jones to vote a certain way. He waits to see what Jones is going to do, preferring not to interfere. But if Jones shows any sign that he'll vote against Black's wishes, Black will force Jones to vote for his preferred candidate. As things turn out, however, Jones does exactly what Black wanted him to do without Black needing to intervene at all.

Jones seems to have acted all on his own, meaning he appears morally responsible for his action, doesn't he? But Jones can't do anything other than vote as Black wishes, because Black plays an important role. His presence functions as an ensuring condition that guarantees that Jones could not have done otherwise.

What's ingenious about the thought experiment is that Black never actually does anything. In the actual sequence of events, he's not an actual intervener but a counterfactual intervener. Jones doesn't show a prior sign that he plans to vote against Black's wishes, so Black doesn't intervene. Yet, because Black would have intervened if Jones showed any prior sign, Jones could not have voted against Black's wishes. So, this powerful thought experiment gives us a reason to reject the principle of alternative possibilities, along with the traditional view that moral responsibility requires the ability to do otherwise.

The Flicker Defense

You might think that if Jones exhibits some prior sign that he's going to vote against Black's wishes, it's just like you trying to make a left turn. If so, then Jones has what philosophers call a flicker of freedom. You might think that's what makes Jones responsible for voting as he does, just like you might think that being able to at least try to make a left turn is what made you responsible for making only right turns. This objection to Frankfurt-style cases is known as the flicker defense or the flicker-of-freedom strategy. It will help us clarify how Frankfurt-style cases are supposed to work.

Let's stipulate that Jones voluntarily exhibits a prior sign, just like you voluntarily tried to turn left. If that's the case, can't we say that Black can keep watch for an even earlier prior sign? But then Black would intervene earlier, and we'd still have a Frankfurt scenario, where Jones lacks the ability to do otherwise. So, it doesn't look like we can object to the thought experiment by insisting that Jones voluntarily exhibits a prior sign.

If Jones involuntarily exhibits a prior sign, we might admit that Jones would have an alternative possibility available to him, but he doesn't really have the ability to do otherwise if it's involuntary. And we couldn't say that he's morally responsible for exhibiting this sign, because he doesn't exercise any control at all.

The important lesson is that good Frankfurt-style cases don't need to rule out every alternative possibility; they just need to rule out possibilities associated with voluntary actions that are, as Fischer puts it, robust enough to ground responsibility. Frankfurt-style cases are powerful thought experiments not only because they seem to show that we don't need the ability to do otherwise to be morally responsible, but also because they tell us something important about the relationship between moral responsibility and causal determinism.

Determinism, very roughly, is the view that there is precisely one possible future. Causal determinism is the view that the past and laws of nature causally necessitate the future down to the last detail.

The traditional view was that causal determinism threatened moral responsibility by ruling out the ability to do otherwise and closing us off from alternative pathways of action. But if Frankfurt-style cases are sound, they challenge the traditional view and shift a long-standing debate. They shift our attention from alternative possibilities to the actual sequence of events. They shift attention away from the ability to do otherwise toward a different ability—the power to be the sources of our actions.

Reading

- Boxer, Karin. "Hart's Senses of 'Responsibility.'" In *Hart on Responsibility*, edited by C. G. Pulman, 30–46. Palgrave-Macmillan, 2014.
- Fischer, John Martin. "The Frankfurt Cases: The Moral of the Stories." *The Philosophical Review* 119, no. 3 (2010): 315–336.
- Frankfurt, Harry. "Alternate Possibilities and Moral Responsibility." *The Journal of Philosophy* 66, no. 23 (1969): 829–839.
- Hart, H. L. A. "Postscript: Responsibility and Retribution." In *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford University Press, 1968.
- Kane, Robert. *A Contemporary Introduction to Free Will*. Oxford University Press, 2005.
- Talbert, Matthew. *Moral Responsibility*. Polity, 2015.

18

HOW LUCK CHANGES MORAL THOUGHT EXPERIMENTS

Can we hold people responsible for things that are out of their control and morally evaluate people differently when the only difference between them is a matter of luck? In this lecture, you will dive into questions like this and examine how luck might affect our intuitions. You will also explore intuitions around manipulation and the importance of historical considerations when thinking about freedom and responsibility.

The Conscientious-Drivers Case

Imagine you're driving home through your neighborhood, and your neighbor is a second driver who's just as careful and on the same route as you. The only difference between your trips is that a dog bolts into the street and runs directly under your neighbor's car. Because there's no time to stop, the dog is sadly killed. Do you think your neighbor is a worse person, morally, than you?

Philosophers like Dana Nelkin have used thought experiments like this one to support three intuitively appealing moral principles. The first one is called the control principle. The idea behind it is that we can only hold people morally responsible for what's under their control. If you think it's unfair to hold your neighbor morally responsible for killing the dog, you probably like the control principle.

If you think it would be unfair to evaluate you and your neighbor differently, you might also like the difference principle, which says that we shouldn't morally evaluate two people differently if the only differences between them are due to factors beyond their control.

Lastly, there's the equal responsibility principle, which says that your neighbor is no more responsible for killing the dog than you. You were equally conscientious drivers; it's just that one of you ran into some bad luck.

The thought experiment does a good job of providing intuitive support for the three principles, but it also highlights what the philosopher Robert Hartman calls the luck-free intuition: the intuition that moral judgments should be immune from luck. Your neighbor couldn't control the fact that a dog ran out into the street, and you couldn't control the fact that nothing out of the ordinary happened.

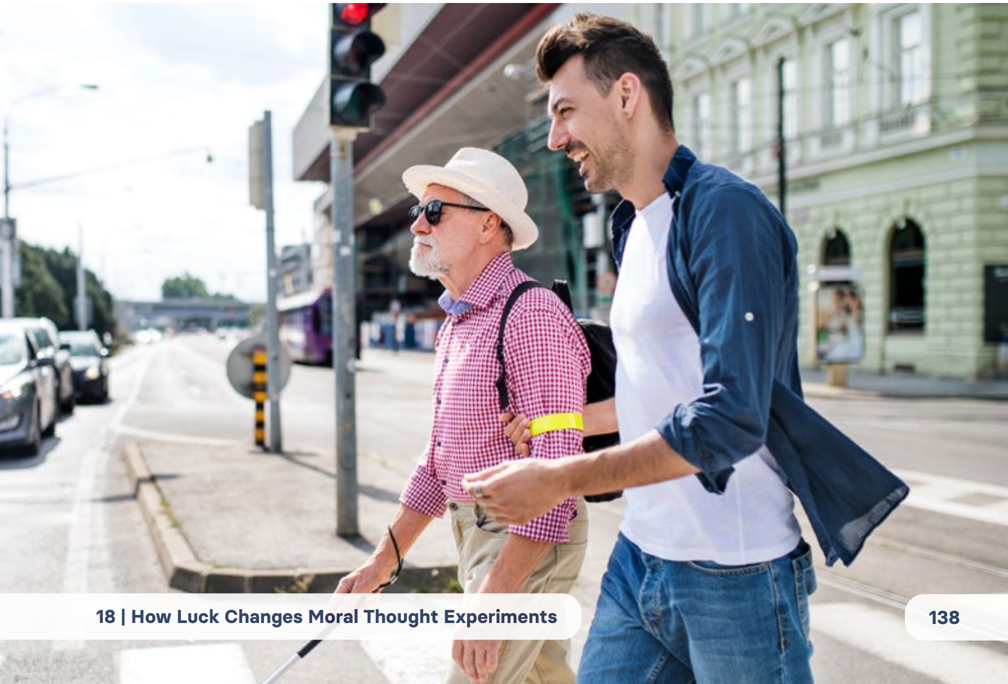
Robert Hartman

The Careless-Driver Case

The careless-driver case is a variation on a case first introduced by the philosopher Thomas Nagel: You're driving to work, and you're expecting an important call. You've set up your phone so when the call comes in, you can answer it without taking your eyes off the road. Still, when it does come in, you're so focused on it, you drive through a red light. Luckily, no one was in the intersection, and you get to work safely.

You're just lucky no one was in the intersection. If someone had been there, you would have hit them, and you would have been blameworthy for hitting them. If someone told you that this had happened to them, you wouldn't praise them, but you probably wouldn't blame them in the same way you would blame someone who struck and killed a pedestrian.

This tells us that sometimes we can reasonably hold people responsible for things that are out of their control and morally evaluate people differently when the only difference between them is a matter of luck. This case taps into what Robert Hartman calls the moral-luck intuition: the intuition that luck can affect our moral judgments.



The Problem of Moral Luck

Thomas Nagel, along with the late Bernard Williams, was the first to identify what's known as the problem of moral luck, which involves conflicting intuitions: the luck-free intuition that moral judgments should be immune from luck and the moral-luck intuition that moral judgments aren't immune from luck.

Common sense tells us that we can't hold people responsible for what's outside their control. If you lean toward the moral-luck intuition, you might conclude that there's something unfair about morality—and that's OK. Although we might hope our moral judgments will always be immune to luck, it turns out that they aren't. Sometimes it's reasonable to hold people morally responsible for what's outside their control; the only leftover question is where to draw the boundaries.

If you lean toward the luck-free intuition, you might conclude that there's something unfair about our ordinary moral practices, like holding people morally responsible for their actions—and that's a problem. To solve it, you might think we'd have to revise those practices.

The power of these thought experiments is that they help us clarify our own views about the role luck should play in moral judgments, while highlighting that our intuitions aren't always clean.

The Ann-Beth Case

Ann is smart and hardworking. She's an accomplished philosopher who loves her job, and the dean wishes everyone was like her. By contrast, Ann's colleague, Beth, isn't all that hardworking and she hasn't accomplished much. The dean works with a team of neuroscientists and psychologists to manipulate Beth into becoming more like Ann. One morning, Beth wakes up and she's psychologically identical to Ann. She feels a new zest for philosophy and wholeheartedly endorses her new preferences and values as her own.

This is a variation on one of Alfred Mele's famous thought experiments about manipulation. The important difference between Beth and Ann is that Beth has been manipulated to have the preferences and values she has, whereas Ann hasn't; Ann acquired hers on her own. Do you think Beth is living her own life? Mele suggests that there's something less than genuine about Beth's ability to govern herself going forward.

And yet Ann and Beth are psychologically identical. They both meet the internal conditions for responsible and autonomous agency. The lesson according to Mele is that we don't just care about how people are from the inside, we also care about "how they came to have … their psychological features." By considering a case of manipulation, we can get a better understanding of what genuine agency looks like. Being a free agent isn't just about having a certain internally coherent constitution; it's about historical factors, too.

The Four-Case Argument

The philosopher Derk Pereboom expands on this lesson in his famous four-case argument, which uses characters from the board game Clue. In each case, Professor Plum kills Mrs. White for the sake of some personal advantage. In each case, Plum will meet all the internal conditions for responsible and autonomous agency. These internal conditions are the kinds of things compatibilists like to emphasize. Compatibilists are philosophers who think free will is compatible with determinism—the view that our actions are the necessary consequence of antecedent causes and the laws of nature. By contrast, incompatibilists are philosophers who think free will is incompatible with determinism.

Each case demonstrates that Professor Plum's decision to kill Mrs. White is determined by factors beyond his control. It's not manipulation that's the problem; it's luck. As a result, the argument is supposed to show that freedom and responsibility aren't just incompatible with manipulation, but determinism as well. It's supposed to show that the problem with manipulation generalizes to determinism; it's an argument against compatibilism and for incompatibilism.

In the first case, a team of nefarious neuroscientists manipulates Plum's brain remotely. Pereboom is betting that your intuition is that Plum is not responsible for killing Mrs. White in this case. And that's because he was directly manipulated to do so!

In the second case, the neuro team programmed Plum when he was still in the womb. In general, Plum is causally determined to make the immoral choice to kill White. This case is what we call indirect global manipulation. Pereboom bets that your intuition is that Plum isn't free and responsible, despite meeting all the compatibilist-friendly conditions.

In this third case, the neuroscientists are absent, and Plum instead had a specific upbringing where people raised him in a particularly rigorous, and perhaps immoral, way. There's no global or local manipulation, just communal and cultural conditioning that causally determines him to act as he does in this specific situation. Plum seems free to act as he does. The compatibilist will insist that Plum isn't being manipulated by other people and is responsible, but Pereboom's point is that this judgment seems intuitively unreasonable. The reason Plum still isn't free is that factors beyond his control causally determine Plum to act as he does.

In the fourth case, the world is deterministic, so everything that happens "is causally determined by virtue of its past states together with the laws of nature. Plum is an ordinary human being, raised in normal circumstances." Pereboom argues in the final case that Plum isn't free for the same reasons he isn't free in the first case.

Pereboom's four-case argument uses thought experiments in a clever way to reach a powerful conclusion. But a lot depends on our intuitive reactions to the first two cases. And a lot depends on whether we agree with Pereboom about whether there's no salient difference between those first two cases and the final two.

Soft-Line and Hard-Line Strategies

Picking up on this, the philosopher Michael McKenna distinguishes between two kinds of responses a compatibilist might have to this argument. Soft-line strategies argue that there is a salient difference between Pereboom's first and final cases. In the first case, Plum's agency is bypassed altogether. McKenna argues that this soft-line strategy is doomed to fail, because we can always modify the first case to avoid whatever worry the soft-liner has. He thinks this will just end in a war of intuitions. McKenna prefers the hard-line strategy, which accepts that the first case is a case of manipulation but claims that Plum is still free and responsible.

Michael McKenna

That might sound like a stretch, but McKenna and other philosophers like Nomy Arpaly point out that there are actual cases similar to Pereboom's first scenario, especially if we make it more attractive to soft-liners. If we're talking about factors beyond people's control that initiate processes that causally determine certain decisions, we don't have to look beyond brain injuries, traumatic accidents, or intense grief. But in cases like these, we don't usually think people are undermined as agents, unless they're no longer able to meet compatibilist-friendly conditions on freedom.

In the end, the philosopher Kristin Demetriou points out that Pereboom's argument faces a dilemma. In the first case, either Plum meets all the internal conditions on free agency, meaning a compatibilist can safely adopt a hard-line strategy and claim that Plum is free, or he fails to meet those same conditions, meaning a compatibilist can adopt a soft-line strategy.

The thought experiments that make up Pereboom's four-case argument are powerful, even if the argument is unsuccessful. It forces us to clarify our intuitions about freedom, autonomy, and responsibility.

Reading

- Breyer, Daniel. *How Luck Changes the Way We View the World*. Audible Originals, 2021. 5 hr., 47 min.
- McKenna, Michael. "A Hard-Line Reply to Pereboom's Four-Case Manipulation Argument." *Philosophy and Phenomenological Research* 77, no. 1 (2008): 142–159.
- Mele, Alfred. *Manipulated Agents: A Window to Moral Responsibility*. Oxford University Press, 2019.
- Nagel, Thomas. *Mortal Questions*. New York: Cambridge University Press, 1979.
- Nelkin, Dana K. "Moral Luck." *The Stanford Encyclopedia of Philosophy*. Summer 2021 ed., edited by Edward N. Zalta. https://plato.stanford.edu/archives/sum2021/entries/moral-luck/.
- Pereboom, Derk. *Free Will, Agency, and Meaning in Life*. Reprint ed. Oxford University Press, 2016.

19

Table of Contents

CHALLENGING WHETHER YOU HAVE FREE WILL

In this lecture, you will explore thought experiments that focus on our ability to choose. You will consider the role that indeterminism plays in debates about free will and examine different views on the significance of self-forming choice, plural voluntary control, and objective worth.

The Liberty of Indifference

Imagine that you're hungry and you see two delicious dates in front of you. Normally you'd pick one, but they both look equally good. Would you just stand there, trapped in frictionless deliberation until you starve, or would you pick one?

Al-Ghazālī

The Islamic philosopher and theologian Al-Ghazālī was the first thinker to present a thought experiment like this. He's certain that you'd pick one or the other. Even though there's no difference between the dates, you nonetheless have the power to choose without needing them to have some determining feature that makes one more appealing than the other.

Al-Ghazālī uses this thought experiment to make the point that if human beings have this ability, why wouldn't God? He uses the case to show that there's no good reason to deny that God could have freely chosen to create the world at one point in time rather than another. Why would God need some determining factor if human beings don't? Al-Ghazālī insists that you can choose one date over the other because you have a special capacity, a psychological faculty, that allows you to choose between indiscernible things. This capacity is the will (*irāda*).

This case is one of the earliest versions of a thought experiment known as a Buridan's-ass case. Imagine a hungry donkey standing at an equal distance between two equally appealing bales of hay. The donkey has a reason to eat each bale of hay, but it doesn't seem like the donkey has a reason to choose one bale over the other. And because the donkey can't break the tie between the two bales of hay, he sadly starves to death. Unlike the donkey, you wouldn't have starved, because, as a human being, you have a will, whereas, it's thought, the donkey doesn't. In medieval philosophical tradition, the freedom you have but the donkey lacks is known as the liberty of indifference, or the free choice of indifference—the freedom of the will to make a choice between alternatives without a determining reason.

We might agree with Al-Ghazālī that you can choose between those dates, but we might wonder whether the choice would be reasonable. If it's arbitrary, is the liberty of indifference worth wanting? Do we really want the freedom not to have our choices determined by reasons? As the philosopher Eugene Chislenko notes, we might worry that these cases raise a practical problem: "the problem of how to act when no single alternative seems to be better, or more worthy of choice, or the one we ought to take."

We could try solving the problem by appealing to randomization, as the philosopher Nicholas Rescher famously suggested. His view is that the rational way to decide in cases like this—in what he calls "the problem of choice in the absence of preference"—is to use random selection. For example, you could flip a coin to choose one of the dates. It might bother you that this process seems arbitrary. So, if Al-Ghazālī is right, and experience tells us that he is, and we can make choices in the absence of any relevant difference, then it's not clear why we should resort to this kind of process.

Chislenko points out that we can't avoid Al-Ghazālī's date case by appealing to randomization. If you flip the coin, you'd still have to decide, without good reason, which of the two dates you'd assign to heads. So, as Chislenko argues, for that to work, we'd already have to have solved the practical problem the case presents.

Kane's Businesswoman

The influential contemporary philosopher Robert Kane is skeptical about whether we should talk about free will as the liberty of indifference. He thinks we should focus on cases where the alternatives we're considering aren't identical, but rather incommensurable. Those cases will help us see the real significance of free will, because when we make a choice to go with one incommensurable option over another, there are real differences at stake, whereas when we go with one equivalent option over the other, there's nothing at stake.

Kane has us consider self-forming choices. These are the kind of difficult choices we make "when we are torn between competing visions of what we should do or become." To illustrate this sort of choice, Kane presents a thought experiment about an ambitious businesswoman.

Imagine Kate, who is running behind on her morning commute. She knows that if she misses her train, she'll miss her business meeting, which will likely cost her a promotion. But as Kate's rushing to the train, she notices someone being assaulted in an alley. She realizes that she can't both stop to help and make her train. Kate is torn because she wants to do both but can't make both decisions.

Kate could make either choice, but none of the reasons she has are sufficient to ensure that she'll make one choice or the other. Her undetermined choice will shape who she is going forward, either by forming her moral personality as someone who helps people even when it's hard, or forming her character as someone who values their career more than helping people.

This case helps highlight the role indeterminism plays in debates about free will. Not all theories of free will are indeterministic, but Kane's is. After all, it's indeterminism that makes it possible for Kate to make that self-forming choice, because indeterminism allows Kate to exercise what Kane calls plural voluntary control. By exercising plural voluntary control, Kate can bring about more than one outcome. She has control over whether to keep running and control over whether to stop, because she can bring about either one of these choices on purpose and for her own reasons. Kane thinks that a deterministic world would rule out plural voluntary control, and so for him, the world has to be indeterministic for us to be truly free.

Rollback Thought Experiment

Not everyone thinks free will and indeterminism are compatible, however. The philosopher Peter van Inwagen asks us to imagine that the universe is indeterministic, meaning that not everything is determined by prior causes, and so Kate's choices might not be determined by her reasons or motives or desires—or anything. Now imagine Kate is riding home on the train after getting her promotion, considering what to order for dinner: steak or pad thai. She's torn, but after a little deliberation, she decides on the steak.

But just as Kate makes her decision, God rolls back time to the very moment she made her choice. Now, if the world is indeterministic and Kate really has the ability to order otherwise, then this time around, Kate might order pad thai rather than the steak, even though she has the same reasons and all the facts about her and the world remain exactly the same. The point is that when God rolls back time, Kate can choose her dinner order, and whether things turn out the same way isn't guaranteed.

Van Inwagen uses this rollback thought experiment to pump the intuition that Kate's choice is really just lucky. In fact, the thought experiment is often used as an argument to show that indeterministic theories of free will are either incoherent or mysterious.

The philosopher Christopher Franklin's view is that it's not so much an argument as a description that clarifies what indeterministic libertarian theories of free will are committed to. These theories say that we at least sometimes have free will and that determinism is false—not all our choices are determined by prior causes. Libertarian theories of free will are one brand of incompatibilism, which is the view that free will is incompatible with a deterministic universe.

Franklin notes that we can't understand what libertarianism is without realizing that there will be observed variability from replay to replay because, according to libertarianism, the presence of indeterminism is among the grounds of freedom. But can we really be free if our actions are undetermined like this?

Red Light, Green Light

Imagine a device that has a green and red light and a button you can push. The instructions say that if you press the button, one of the two lights is guaranteed to flash. The red light might flash, or the green light might flash. Put differently, there's an objective probability of less than 1 that the red light will flash, and the same goes for the green light. You press the button. Do you think you have a choice about which light will flash?

INSTRUCTIONS:
If you press the button, one of the two lights is guaranteed to flash.

This is another one of Peter van Inwagen's thought experiments, and he thinks it's "obvious that you have no choice about this." Even if you really want to see the red light flash and you push the button with this in mind, have you chosen which light will flash? It might be the red light, but it also might be the green light, because the connection between pressing the button and which light flashes is genuinely undetermined.

We might interpret the thought experiment as an argument against indeterministic libertarian theories of free will. But we might read it as a challenge to clarify something about plural voluntary control.

Think back to Kate. If she were to decide on steak for dinner, would it be genuinely undetermined whether she would then actually order steak or pad thai? Would indeterminism come between her choice and her action? If so, then that disconnect sounds bad, right?

But we don't have to think of Kate's choice like that. According to Kane, we're supposed to "think of [Kate's deliberative] effort and the indeterminism as fused." Kate's deliberative effort is indeterminate. It's genuinely undetermined which effort will win out, but once one effort wins out, Kate makes that choice, not the other.

Objective Worth

So, why is this sort of freedom worth wanting? In part, Kane's answer is that this kind of free will is valuable because it allows us to shape who we are and be ultimately responsible for our actions. It gives our lives objective worth. To understand this idea, Kane introduces another thought experiment.

Alan is a struggling artist who's feeling despondent because no one appreciates his paintings. When his wealthy friend finds out how Alan feels, he goes to an upscale gallery and arranges for Alan's paintings to be purchased under an assumed name at $10,000 apiece. When Alan learns about this, he's elated and feels like his work is finally getting the attention it deserves.

Now imagine two different worlds. In one, Alan isn't much of an artist, but this attention makes him feel like one, and he ends up dying "happily, believing he is a great artist." In the other, Alan really is a great artist, and this attention makes him feel like one, so he dies "happily, believing he is a great artist."

Kane bets that you would prefer the second world. That's the world with objective worth, where Alan doesn't just feel like he's a great artist—he really is one! Alan's experience is the same in both worlds, so the difference between them is something objective. And we value that over mere subjective experience. Kane suggests that "the notion of ultimate responsibility is of a piece with the notion of objective worth."

Some philosophers challenge this thought experiment. Tamler Sommers presents the case of Joe the naturally funny guy. Sommers thinks there's objective worth to being funny that goes beyond the subjective satisfaction that comes along with it. But that objective worth isn't connected with being ultimately responsible for being a funny person—because Joe is just naturally that way. So, it seems like Kane's wrong: Objective worth doesn't seem to be "of a piece with" ultimate responsibility and indeterministic free will at all.

Where does this leave us? Whether we're making a choice about what to eat or about how to act, some philosophers think what we do is ultimately up to us. And that power, which we call free will, is worth wanting, even if it remains elusive.

Reading

- Chislenko, Eugene. "A Solution for Buridan's Ass." *Ethics* 126 (2016): 283–310.
- Franklin, Christopher. *A Minimal Libertarianism: Free Will and the Promise of Reduction*. Oxford University Press, 2018.
- Hasan, Ali. "Al-Ghazali and Ibn Rush (Averroes) on Creation and the Divine Attributes." In *Models of God and Alternative Ultimate Realities*, edited by Jeanine Diller and Asa Kasher, 141–156. Springer, 2013.
- Kane, Robert. *The Significance of Free Will*. Oxford University Press, 1999.
- Kaye, Sharon. "Why the Liberty of Indifference Is Worth Wanting: Buridan's Ass, Friendship, and Peter John Olivi." *History of Philosophy Quarterly* 21, no. 1 (2004): 21–42.

Table of Contents

20

SUPPOSE YOU’RE IMMORTAL. WHAT DO YOU VALUE?

If you could experience “a lifetime of bliss” by plugging into a simulator, or if you could live forever by taking a pill, would you choose to do either? Why or why not? By imagining what it would be like to live forever—and whether we’d want to—we can shed light on what it means to live a meaningful life.

Nozick's Experience Machine

Imagine that scientists have developed a cutting-edge machine that can give you any experience you want. Neuropsychologists have figured out how to stimulate your brain so that you can think and feel whatever your heart desires. You just have to come in to plug in!

You won't even know that you're in the machine while you're in it; it's completely immersive. But that requires more than putting on a VR headset. You would need to enter a high-tech sensory deprivation tank, where you'd be suspended in a special fluid, with electrodes attached to your brain. This ensures that you're completely closed off from the outside world.

How long you plug in is up to you. An experience could last 30 minutes or an entire day. The extended immersion package allows you to remain in the experience machine for years at a time. Afterward, you'll have the option to exit the tank for a short period of time to select your next set of experiences. The lifetime package allows you to enter the experience machine for "a lifetime of bliss." The team will work with you to curate the best possible life for you based on your preferences.

This is Robert Nozick's experience machine thought experiment, first introduced in his 1974 book *Anarchy, State, and Utopia*. Would you sign up? You'd probably be curious about that short-term experience. It sounds kind of amazing, doesn't it?

The bigger questions are about extended immersion and the lifetime packages. Would you spend two years, or a lifetime, in the experience machine? Before you answer, keep in mind that while in the machine, you won't know that you're there; you'll think it's actually happening. You won't even remember that you signed up to be in it. And your experience is guaranteed to be great.

If experiences are what's most important to you, then it seems like you should definitely plug into the machine. But Nozick wagers that you care about more than just your experiences. Don't you care about actually doing things, and not just having the experience of doing them? Other things might matter to you too, like actually connecting with other people. If you use the machine to forge a friendship with someone you admire, wouldn't your friendship be ruined because it's just a simulation?

If you wouldn't plug in, and you think this is the best choice based on good reasons (rather than fear), you might also think that you should not plug in—that no one should. If this is your reaction to the experience machine, Nozick thinks you've learned something important: You've learned that something matters to you in addition to experience. This is significant because it shows that views about what's valuable that focus only on internal states are wrong.

In his original presentation of the thought experiment, Nozick says that "others can also plug in to have the experiences they want, so there's no need to stay unplugged to serve them." We can read this as a way of dampening fears of being irresponsible for plugging in and leaving family and others important to us behind.

Your gut may still be telling you that you wouldn't and shouldn't plug in, possibly because you want to live an actual life with those people, rather than a shared simulation. But are your instincts here really a window into what you value, or is something else going on?

Psychological Biases in Thought Experiments

Law professor Adam Kolber suggests that psychological biases might better explain these instincts than appealing to what we really value. For instance, some people would hesitate to plug in because they find the thought experiment weird, but their aversion wouldn't tell us anything about what they really care about. Most importantly, Kolber suggests that our instincts telling us not to plug in are driven by status quo bias: the unjustified preference that things remain how they are.

To make his case, he offers a twist on Nozick's original thought experiment. Imagine that the scenario is reversed. You're already in the experience machine and considering whether you should unplug. When you ask about life outside the machine, you find out that your experiences will be worse than they've been in the machine, but you'll be in the real world. Kolber bets that you'd choose to stay plugged in. Would you?

The philosopher Dan Weijers, who tested another variation of the scenario, thinks that the thought experiment doesn't provide evidence that we care about contact with the real world or that we care about anything beyond "how our experiences feel to us on the inside." He points us to an important lesson about the power of thought experiments more generally. He notes that, for a long time, especially in the classroom, philosophers have used Nozick's experience machine as a knockdown argument against hedonism, showing that pleasure isn't the only thing we care about.

The problem, Weijers suggests, is that treating thought experiments as knockdown arguments is misleading, especially when introducing people to philosophical views. What we should do is warn people, students in particular, "about the many biases and other irrelevant factors that might be affecting our judgments about these kinds of scenarios." To figure out just how powerful a thought experiment really is, we have to know how thought experiments like these might go wrong.

Williams's Immortality Pill

Nozick's experience machine might offer a lifetime of bliss, but in the end, you'd still die. One of the reasons we might think death is bad is because it deprives us of the goods of life. Maybe the best life is one that doesn't end?

Philosopher Bernard Williams bases his immortality-pill thought experiment on the story of Elina Makropulos, a character from Leoš Janáček's opera *The Makropulos Affair*. In the opera, Elina's father gives her an "elixir of life," which allows her to live for 300 years at her current age, after which she can take it again. Elina takes the elixir, but then, after those 300 years, chooses to die, but not before destroying the elixir itself.

Imagine that someone has offered you an immortality pill. You'll live forever in a healthy state, with the ability to learn and experience new things. Pretty amazing, isn't it? So, would you take it?

You might have some concerns, so let's just stipulate that if you take the pill, you'll always have your basic needs met and basic freedoms. Let's also stipulate that you can give a pill to anyone you love who wants it. Of course, some bad things will happen, you'll be sad at times, and things won't always go your way—that's life, after all—but let's pretend that the pill also comes with a no-awful-life guarantee.

Let's sidestep any other worries by following the way the philosopher Bernard Williams thinks of your choice. Williams imagines that you have the choice to take a pill that will make you immortal for 300 years at a time. After those 300 years, you have the option to take another pill or die. Would you take it?

Finding Meaning in Immortality

Williams himself wouldn't take that pill. He thinks an everlasting life would be too boring to be meaningful. He argues that a meaningful life is a life spent satisfying what he calls categorical desires: desires that give us reasons to live and pursue projects that move us into the future.

We also have conditional desires, which are desires for things to go on a certain way without disrupting the way we continue to live. The fundamental difference between categorical and conditional desires is that categorical desires provide us reasons to continue our lives, whereas conditional desires don't. Williams worries that everlasting life would exhaust our categorical desires, plunging us into a state of meaninglessness, making life no longer worth living.

The philosopher Iddo Landau would take the pill and doesn't think he'd get bored at all. To show why, he considers something he really likes: reading books. Forever is a long time, so across the ages, you probably could read every book in existence. And when you finish all of them, you could just reread them! That would be a fun project, because rereading a book is a different experience from the first reading, even if we remember the whole book. And people will likely write new books for you to explore.

Moreover, the philosopher John Martin Fischer points out that as long as we can retain our individuality while still forgetting enough about ourselves, like our experiences and the world, we could endlessly return to previous projects (or new iterations of them), because we will have forgotten that we pursued them.

Williams, however, worries that immortality is only going to be desirable if we can maintain our sense of self throughout it. We can't satisfy all of our desires unless we have a consistent sense of self throughout. The problem is that if someone were to live an immortal life, they would exhaust their projects—not all the projects in the world, but all of the categorical desires that give them a reason to live.

You might think that Williams is wrong here. Why can't we just add more projects we didn't start out with when we decided to take the immortality pill? Even if we don't forget too much about ourselves, don't we still get new interests, or see things in new ways? Even if we can't add more projects, Fischer thinks Williams is too pessimistic. When we do things we enjoy, it often isn't to get something out of it, but because the experience itself is valuable and worth pursuing. Fischer thinks valuable experiences like these could be "reliably repeatable in an immortal life."

He also thinks the "magic and mystery of love and the compelling beauty of friendship" have an irreducible quality to them and that, because of this, they "would not lose their transformational and inspiring qualities in an immortal life."

Return to your choice and imagine, following Fischer, that you've already taken the first pill of immortality. You've lived through 300 years with the no-awful-life guarantee and still have the people you love with you. You've essentially lived a life of your design, doing everything you wanted. Now imagine it's time to either opt back in or opt out. Would you rather live another 300 years or die?

Reading

- Anderson, R. Lanier. "Friedrich Nietzsche." *The Stanford Encyclopedia of Philosophy*. Summer 2022 ed., edited by Edward N. Zalta. https://plato.stanford.edu/archives/sum2022/entries/nietzsche/.
- Fischer, John Martin. *Death, Immortality, and Meaning in Life*. Oxford University Press, 2019.
- Kolber, Adam. "Mental Statism and the Experience Machine." *Bard Journal of Social Sciences* 3 (1994): 10–17.
- Nozick, Robert. *Anarchy, State, and Utopia*. Blackwell, 1974.
- Weijers, Dan. "Nozick's Experience Machine Is Dead, Long Live the Experience Machine!" *Philosophical Psychology* 27, no. 4 (2014): 513–535.
- Williams, Bernard. "The Makropulos Case: Reflections on the Tedium of Immortality." In *Problems of the Self: Philosophical Papers 1956–1972*. Cambridge University Press, 1973.

Table of Contents

21

VISIT TWIN EARTH TO EXPLORE MEANING

Have you ever wondered how your thoughts and words have meaning? And what is meaning, anyway? In this lecture, you're going to look at some of the famous thought experiments that have tried to answer these questions, which are among the most central in philosophy.

Reference and Sense

Imagine that you don't know much about astronomy, but you love to look at the stars. At some times of the year, you notice something bright in the sky just before sunrise, and at other times, you see something bright in the sky just after sunset. You call one the morning star and the other the evening star. But then you take a class, and you learn that these two objects are not stars at all. They are one and the same thing—the planet Venus.

This is a little puzzling, because before the class, it seems like you're not talking or thinking about the same thing. But after the class, you realized that you were in fact talking and thinking about the same thing all along. How can we make sense of this?

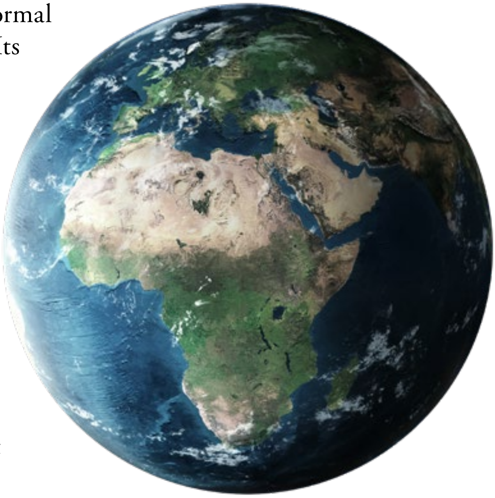
The German philosopher and mathematician Gottlob Frege had an influential solution to this puzzle. His suggestion was to make a distinction between two aspects of meaning: reference and sense. For Frege, reference determines the truth value of a sentence, whereas sense is how the reference is presented—it's the thought the sentence expresses. Frege also holds that two words or phrases or expressions that have the exact same reference don't have to have the same sense.

So, when you talked about the morning star and the evening star, you were talking about the same thing—those expressions had the same reference but not the same sense. And that's really what mattered for understanding what you meant by talking about them. But what you meant changed when you learned that they were both Venus. That happened because the sense, not the reference, of your words changed.

According to Gottlob Frege, sense determines reference. If two words have the same sense, then they'll also have the same reference—assuming, of course, that there's something in the world they refer to. Some words might not have objects they designate—like a unicorn.

Putnam's Twin Earth

Famously, the philosopher Hilary Putnam considers a Twin Earth thought experiment that puts pressure on Frege's view. Imagine that there's a planet just like Earth where everything is duplicated except for one important difference: Instead of water, a different liquid flows in rivers and falls from the sky on Twin Earth. It is indistinguishable from the water on Earth under normal circumstances. But it's not H_2O. Its complicated chemical formula is abbreviated as XYZ. On Twin Earth, there's also a doppelgänger—an identical copy—of everyone on Earth. Some of these Twin Earthlings speak English, and when they talk or think about this liquid, they call it water.



Suppose that a spaceship from Earth lands on Twin Earth. Wouldn't everyone on that spaceship initially think that the word *water* had the same meaning on Twin Earth as it has on Earth? After all, every other English word seems to have the same meaning on both planets. But scientists soon realize that the word *water* means H_2O on Earth and XYZ on Twin Earth.

Now imagine this thought experiment back in the year 1750, when chemistry had not yet developed on either planet. The typical English speaker didn't know that water on Earth was made up of hydrogen and oxygen, and on Twin Earth, they didn't know it was made up of XYZ. Suppose that Oscar and Twin Oscar are exact physical and psychological duplicates, living on Earth and Twin Earth in 1750. In every way, they share all the exact same beliefs about what they call water. Do they mean the same thing when they call something water?

Putnam's view is that they don't mean the same thing, despite being psychologically identical. This is important, because it's supposed to show that "the psychological state of [a] speaker does not determine" the meaning of a word like *water*. In other words, what we mean isn't solely determined by what we think; it's partly determined by how the world really is.

The thought experiment pushes back against the Fregean picture of how meaning works. It suggests that sense—what a word or thought expresses—doesn't fully capture meaning, and that sense doesn't actually determine reference.

Intension and Extension

Putnam's thought experiment targets a psychological understanding of sense, which isn't quite how Frege himself understood sense. Because of this, Putnam talks about sense as *intension*, which is a technical term that picks out the psychological content of a word or concept or thought. He talks about reference as *extension*, a technical term that picks out the objects in the world a word or concept or thought designates.

With this in mind, consider a thought experiment Putnam uses to supplement Twin Earth. Imagine that you're walking in a forest, looking at trees. Suppose you can't distinguish between elm trees and beech trees. Your concept is the same for both trees, and when you think about one, you could just as easily be thinking about the other.

Still, if you were to call a beech tree an elm, you'd clearly be wrong, because the word *elm* has the same extension for you as it does for everyone who uses the word. The word *elm* refers to the set of all elm trees, or something like that. And we can say the same thing about the word *beech*.

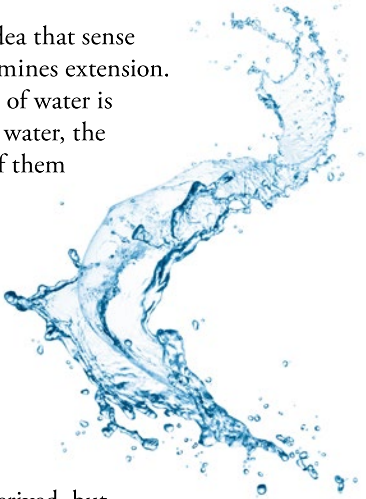
Is it at all plausible that the difference in extension between the words *elm* and *beech* arises from any psychological or conceptual difference? After all, your concept is the same for both. It doesn't seem like sense can determine reference—or that intension can determine extension.

Putnam's two thought experiments ultimately make the case for semantic externalism over semantic internalism. Semantic internalism is the view that intension determines extension and that psychological factors fully account for meaning. In slogan form, it's the view that meaning is in the head. By contrast, semantic externalism is the view that intension doesn't determine extension and that meaning is at least partly determined by external factors.

When you see a tree that you think is an elm and call it an elm, what do you mean to talk about, whatever it is that you have in your head or the kind of trees that are scientifically recognized as elms? Putnam's hunch is that you intend to get things right. He thinks the lesson here is that there's a division of linguistic labor. We can't talk about elm trees without having good ways of distinguishing elm trees from beech trees, and for that we need specialists, the people who do the nonlinguistic labor. Putnam suggests that meaning isn't just determined by factors external to psychology; meaning is fundamentally social.

If we were to return to Twin Earth now and compare scientific notes, we'd find that we don't mean the same thing by *water*. When we use that word, we mean to talk about the stuff we find on Earth, whereas when our counterparts use the word, they mean to talk about the stuff they find on Twin Earth. What this means is that the meanings of our words depend on the things in the world that we interact with. So, meaning isn't just social; it's environmental—it's determined by things "around here," at least in the case of natural kind terms like *water* and *elm*.

This point allows Putnam to revise Frege's idea that sense determines reference, or that intension determines extension. When Oscar talks about water, the reference of water is H_2O, whereas when Twin Oscar talks about water, the reference of water is XYZ—even if neither of them knows that. Because reference determines sense, or extension determines intension, we won't always know what is in fact meant by a term until we know more about its extension. In the case of water, we didn't know what that word really meant until we figured it out, scientifically and empirically. The wet stuff on Earth is really H_2O.



Putnam's Twin Earth thought experiment makes a powerful case for how meaning is derived, but the way he originally frames it depends on Oscar and Twin Oscar having the exact same psychological states. If we stipulate that, we might think that everything in their heads was the same, but as some philosophers have pointed out, it looks like the content of their thoughts about water differ.

This observation expands the thought experiment. Now it seems like it contains the intuitive grounds for a different kind of externalism: content externalism, which is the view that at least some mental content—some of what our thoughts, beliefs, desires, and things like that are about—depends on facts about the world. Putnam's thought experiment focuses on natural kind terms like *water*, *elm*, and *gold*—these are things we find out in the world and that, at least if we're realists about them, we think have a certain kind of nature or essence.

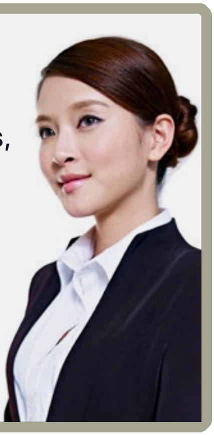
Moral Twin Earth

Now let's consider whether moral terms are similar. The philosophers Mark Timmons and Terrence Horgan present Moral Twin Earth, which is just like Earth in almost all respects. You even have a doppelgänger here who thinks about moral issues in much the same way that you do.

Although the term *morally right* plays the same social role on Moral Twin Earth for your doppelgänger as it does for you on Earth, its meaning is determined by different natural properties.

Imagine that the two of you meet. You discover that when your Moral Twin says that something's morally right, they mean something rather different from what you mean. You think that it's morally right to lie when it brings about good consequences, but your Moral Twin disagrees. Would you conclude that the two of you just don't mean the same thing when talking about what's morally right and wrong? Or would you think that you two disagree about what's morally right in that situation?

When consequentialists say that something's right, they mean that it produces the best possible consequences, increasing the total welfare for everyone. When deontologists say something's right, they mean that it adheres to a fundamental moral law—doing the right thing means doing their duty, no matter what the consequences are.



Timmons and Horgan think that you are having a genuine moral disagreement about what's morally right in that situation. And if that's right, then even if Putnam's Twin Earth thought experiment tells us something about how language works when it comes to natural kind terms like *water*, it leads us astray when we think about language more generally—in particular, when we think about how moral discourse works.

Reading

- Horgan, Terry, and Mark Timmons. "New Wave Moral Realism Meets Moral Twin Earth." *Journal of Philosophical Research* 16 (1991): 447–465.
- Putnam, Hilary. "The Meaning of 'Meaning.'" *Minnesota Studies in the Philosophy of Science* 7 (1975): 131–193.
- Speaks, Jeff, "Theories of Meaning." *The Stanford Encyclopedia of Philosophy*. Spring 2021 ed., edited by Edward N. Zalta. https://plato.stanford.edu/archives/spr2021/entries/meaning/.
- Zalta, Edward N., "Gottlob Frege." *The Stanford Encyclopedia of Philosophy*. Fall 2022 ed., edited by Edward N. Zalta. https://plato.stanford.edu/archives/fall2022/entries/frege/.

22

HOW DO YOU KNOW WHEN YOU KNOW SOMETHING?

What is knowledge? Philosophers have been thinking about this issue for a long time. In this lecture, you will consider scenarios that explore the relationships between belief, knowledge, luck, and understanding. These thought experiments provide intuitive motivation to move beyond long-standing debates about the nature of knowledge to investigate something new—which is, ideally, what thought experiments are supposed to encourage in the first place.

Gettier Cases

Imagine that Smith and Jones work for the same company and they both want the same promotion. Smith finds out that Jones is probably going to get the promotion, and he has good reasons for believing that. A little defeated, Smith remembers that Jones also has a nice car. He thinks that the person who got the promotion has a nice car.

But Jones didn't actually get the promotion. Smith did; he just doesn't know it yet. It also turns out that, unbeknownst to Smith, his wife just bought him a nice car. So, even though Jones didn't get the promotion, it's true that the person who got the promotion owns a nice car. Does Smith know that the person who got the promotion owns a nice car?

Smith has what philosophers call a justified belief. Sure, it's false that Jones got the promotion, but still, Smith's belief is justified. And, though he doesn't know it, Smith also has a justified true belief that the person who got the promotion owns a nice car.

One influential way of analyzing knowledge, often known as the standard analysis, is to say that knowledge is justified true belief. This thought experiment is a counterexample to that view, because although Smith has a justified true belief that the person who got the promotion owns a nice car, he surely doesn't know that. Justified true belief isn't enough for knowledge.

Gettier cases challenge the view that knowledge is justified true belief. Introduced in 1963 by the philosopher Edmund Gettier, they have generated perhaps more discussion than any thought experiment since that time.

In Gettier cases, someone has a justified belief, but that belief is the result of some bad luck, and so the justification associated with the belief isn't closely connected with the truth of the belief. Instead, as a result of some good luck, the belief just happens to be true.

Edmund Gettier

Smith's epistemic bad luck is that Jones didn't get the promotion, since it undermines what he thinks he knows. Smith's good luck is that, unbeknownst to him, the person who got the promotion does in fact own a nice car, even though that person isn't Jones. This good luck offsets the bad luck, and we're left with a situation where someone has a justified true belief but not knowledge. Are you convinced?

Goldman's Fake Barn County

You might be worried that there's false evidence in that thought experiment, and you can't use false beliefs, even if they're justified, to arrive at genuine knowledge. If justified true belief is arrived at legitimately, then there's knowledge. But we can't dismiss Gettier cases so easily, because we can devise thought experiments that don't make use of false evidence.

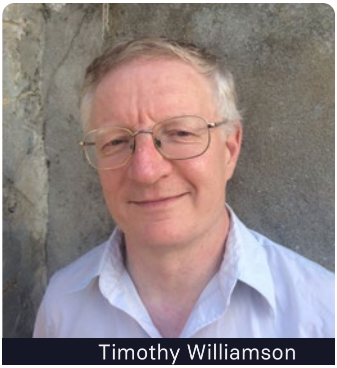
Imagine that you're driving with your family and you're looking out the window. You point out a barn in the distance. But you don't know that this is Fake Barn County, which is littered with barn facades. There's only one real barn to be found, and you just happened to see it. Do you know that there's a real barn in the field?

In this case, you don't rely on any false evidence. You base your belief on seeing an actual barn. It seems like you have justified true belief, and yet it doesn't seem like you do know that there's a barn in the field. The philosopher Alvin Goldman came up with this thought experiment in 1976. What causes you trouble is the weird environment you find yourself in. Your eyes work just fine, and there's actually a barn there. But it seems like it's just lucky that you saw the one real barn.

We might then think that the lesson from these Gettier thought experiments is that any good theory about what knowledge is should include a no-lucky-belief requirement. According to this anti-luck way of thinking about epistemology, knowledge is justified, non-accidental, true belief.

Though powerful, the thought experiments call into question just about any plausible theory of knowledge that appeals to justified true belief—or even non-accidental justified true belief. So, although these thought experiments push new lines of research, they also remain a persistent problem that at least some philosophers don't think we're going to solve.

One of those philosophers is Timothy Williamson, who thinks that Gettier-style thought experiments show that knowledge is unanalyzable. We should instead take knowledge as conceptually primitive and use it to understand concepts like epistemic justification, belief, and evidence—not the other way around. This is knowledge-first epistemology, which moves away from the traditional justified true belief model.


Timothy Williamson

Context-Sensitive Knowledge

Imagine that it's 1990, before mobile banking, and you're driving home with your roommate after work. It's Friday, and you both just got paid. You'd like to deposit your paychecks at the bank before the weekend, but it's not especially important, because you won't need the money until next week. Your roommate suggests that you go to the bank tomorrow. She says her friend Erin told her it's open on Saturdays. Erin knows this because she went there last weekend. You agree to go tomorrow, because Erin knows the bank is open.

In a variation of this scenario, imagine that it's very important to deposit the checks so you can pay your rent on time. When your roommate tells you about what Erin knows, you question whether she really knows it. You decide to go to the bank and not wait until tomorrow.

In the first scenario, you were willing to say that Erin knows that the bank is open on Saturday, but in this second scenario, you're not. Why? Your circumstances are difference in each case. The first scenario is a low-stakes situation for you, whereas the stakes are a lot higher for you in the second scenario. As a result, it's easier to attribute knowledge to Erin in the first scenario than it is in the second.

The philosopher Keith DeRose first introduced thought experiments like this to make the case that knowledge is context-sensitive. He argues that the ordinary way we use the word *knows* is context-sensitive. Whether Erin knows that the bank is open or not depends on the context. In the first scenario, the context allows for you to say that Erin knows, but not in the second.

Keith DeRose

For DeRose, the speaker's context is what matters, the person attributing knowledge to someone—not the context of the person having knowledge attributed to them or not. According to this view, known as contextualism, you correctly attribute knowledge to Erin in the first scenario, and in the second, you correctly refuse to attribute knowledge to her.

We can also consider whether you know that the bank is open on Saturday. Let's say you're driving alone and that you're the one who was at the bank last weekend. If contextualism is right, you're correct to say that you know in the first scenario and correct to say that you don't know in the second scenario. And in these new scenarios, you're both the speaker and the potential knower. It's clear that what we care about is whether you know.

Grimm's "Whose Stakes?" Problem

Some philosophers have thought that what matters is whose practical interests matter most when it comes to knowledge. The philosopher Stephen Grimm calls this the "whose stakes?" problem. He offers another variation on the bank thought experiments. Imagine that it doesn't matter to you, as the one attributing knowledge, or to Erin, as the potential knower, whether the bank is open on Saturday. But it's extremely important for your friend Ayushi. She told you that she will be evicted from her apartment if she doesn't deposit her check before Monday. She asked whether the bank would be open on Saturday, and you said that you didn't know, because you had no evidence either way. Then Erin tells you that she was at the bank last Saturday, and so it will be open this Saturday, too.

Grimm thinks that it's natural for us to say that Erin doesn't know that the bank is open. Even if she's right and has good reasons for thinking it is, Grimm's view is that, given Ayushi's dire situation, Erin's reasons aren't "good enough for knowledge."

The power of the banking cases is that they've put pressure on different views of knowledge. Given all the problems that arise when focusing on knowledge, since the mid-1990s, philosophers have started to focus on other important cognitive states, like understanding.

Understanding

Imagine that you're in the library, researching the role women played in the American civil rights movement. You pick up a book whose cover just happens to grab your attention. You carefully read its chapter on Daisy Lampkin. As you read, it seems to you that you start to understand why Daisy Lampkin was such a successful organizer and why her efforts mattered to the civil rights movement.

The thing is, although the book you happened to pick up is accurate, up to date, and reliable, every other book on that shelf is not. If you had picked up any one of those books, then you'd believe a host of falsehoods about your topic. So, it seems like everything you believe based on reading that chapter is a matter of luck. That said, what you believe is true and justified.

On the one hand, it seems like you understand why Daisy Lampkin was such a successful organizer and why her efforts mattered to the civil rights movement. On the other hand, it seems like you don't know these things. As a result, it looks like luck undermines knowledge, but it doesn't undermine understanding. Whereas knowledge is incompatible with accidental belief, understanding isn't.

The philosopher Jon Kvanvig first suggested this sort of thought experiment to make the case that understanding is not a form of knowledge. When it comes to understanding, it doesn't matter how our beliefs came to be true. What matters is cognition. If what you read in that book is true, you can understand it no matter how the information ended up being true.

Stephen Grimm resists Kvanvig's conclusions. Grimm thinks understanding seems just as incompatible with luck as knowledge. Following the philosopher John Hawthorne, Grimm also suggests that knowledge is at least sometimes compatible with lucky environments. Hawthorne has us imagine that he gives six books about Austria to six children and asks them to each pick one at random. Five of the books get the capital of Austria wrong; one gets it right. He then asks each child to look up the capital of Austria and tell him what it is.

Would you say that the child who says Vienna is the child that knows? Or would you say that none of them know, including that child? Hawthorne's experience is that people say the child who gets the answer right knows what the capital of Austria is. They don't care that the kid was lucky to get the only book with the right information in it.

Grimm suggests that it was lucky for that child to get the reliable book, but the fact that the book was reliable wasn't itself a matter of luck. In the end, Grimm thinks that our judgments about whether someone understands "sway together with" our judgments about whether they know.

The power of Kvanvig's library thought experiment is that it invites us to explore the relationship between knowledge and understanding further, while pressuring us to think about the nature of understanding and get clear about whether it's compatible with luck or not.

Reading

- DeRose, Keith. "Contextualism and Knowledge Attributions." *Philosophy and Phenomenological Research* 52, no. 4 (1992): 913–929.
- Gettier, Edmund. "Is Justified True Belief Knowledge?" *Analysis* 23, no. 6 (1963): 121–123.
- Grimm, Stephen. "Understanding." *The Stanford Encyclopedia of Philosophy*. Summer 2021 ed., edited by Edward N. Zalta. https://plato.stanford.edu/archives/sum2021/entries/understanding/.
- Kvanvig, Jonathan. *The Value of Knowledge and the Pursuit of Understanding*. New York: Cambridge University Press, 2003.
- McCain, Kevin. *Epistemology: 50 Puzzles, Paradoxes, and Thought Experiments*. Routledge, 2021.

23

HOW TO CREATE CIVILIZATION FROM CHAOS

Thomas Hobbes famously noted that life in a state of nature would be "solitary, poor, nasty, brutish, and short." Would you agree? In this lecture, you will explore his views on human nature as well as those of the Confucian philosophers Xunzi and Mengzi. Using some famous thought experiments, including the prisoner's dilemma and John Rawls's veil of ignorance, you can imagine what it would be like to negotiate our way to social harmony, even morality.

State of Nature

What would things be like without any laws, nations, governments, institutions, or even social norms? What would this "state of nature" be like? We might imagine a precivilized state of nature—a time before governments and social customs were established—or a postapocalyptic state of nature, a time after these things have collapsed.

The great Confucian philosopher Xunzi, who flourished in the 3rd century BCE, imagined a state of conflict where people would struggle with one another to satisfy all their desires. This would naturally lead to disorder, which would lead to everyone's impoverishment.

The 17th-century English philosopher Thomas Hobbes also imagines that the state of nature is a state of conflict, agreeing with Xunzi that people are fundamentally self-interested, and that selfishness is grounded in our nature as human beings. To illustrate this, Xunzi considers another thought experiment. Imagine two brothers in the state of nature. There are no laws or customs to guide their behavior. If they had to divide some property between themselves, Xunzi thinks that they'd fight over the property, seeking to satisfy their own personal interests.

Xunzi's thought experiment highlights the notion that, because we're selfish, we're also prone to conflict. And Hobbes agrees. They have what we might call a pessimistic view of the state of nature, and their intuitions about what the state of nature would be like are at least partly driven by their views about human nature.

For Xunzi, ritual can improve our situation. "Ritual" is a translation of the ancient Chinese word *li*, and in the Confucian tradition, *li* came to refer not just to things like religious ritual but to propriety more generally, including not only matters of custom and decorum but also legal and moral standards. Xunzi's view is that what's good in people isn't their nature but their "artifice"—all the habits and qualities that they acquire through deliberate effort. In this sense, Xunzi thinks that morality, along with custom and law, is manufactured rather than natural.

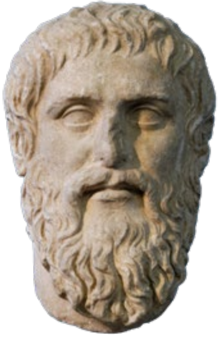
Both Hobbes and Xunzi lived during times of crisis, but Hobbes had a different suggestion for how to emerge from the state of nature. He thought that the state of nature was a problem not just because resources are scarce and human beings are selfish and prone to conflict, but also because we're all more or less equal in the state of nature.

For Hobbes, this means that no one is going to naturally emerge as the sovereign. Something else has to happen, and Hobbes suggests that it's a social contract: an agreement among the individuals in the state of nature to establish something better, as a way of resolving the state of constant conflict.

Hobbes's Social Contract

Hobbes thinks that, in the state of nature, everyone has certain "natural rights." They include things like unrestricted individual liberty and the right to defend yourself and your family and your possessions by any means necessary.

The problem is that they lead to a "war of all against all," and everyone ends up being in a state of "continual fear, and danger of violent death." When we decide to enter into a social contract, we give up these natural rights. The seeds of this idea go back at least to Plato. In the *Republic*, the character Glaucon suggests the origin of justice itself goes back to a kind of social contract.



In the state of nature, we might imagine that the best life would be the life of a superhuman, someone invulnerable to attack with the freedom to do whatever they wished. The problem is that this ideal scenario is beyond reach. What we have to do is settle for second best. That's easier said than done, though. If we care only about our own interests, why would we surrender our natural rights? And why would we trust anyone else to do so?

The Prisoner's Dilemma

Imagine that you and your friend Daniel robbed a bank together. You were desperate and it served your mutual interests. The problem is that you got caught. At the police station, you're isolated from each other. You're being interrogated in one cell, Daniel in another. You both care more about your own welfare and freedom than you care about each other.

But there's not enough hard evidence to file charges. The prosecutor's only chance is to get you to confess. But you don't know that. So, the detectives try to get each of you to turn on the other.

This thought experiment is the famous prisoner's dilemma, so named by the mathematician Albert Tucker, who popularized it. What would you do in this scenario? Here are the options you're given:

Option 1: Daniel confesses, but you don't. In that case, Daniel goes free and you get 10 years in prison.

Option 2: You confess, but Daniel doesn't. In that case, you go free and Daniel gets 10 years in prison.

Option 3: Both you and Daniel confess. In that case, you each get 7 years in prison.

Option 4: Both you and Daniel keep your mouths shut. In that case, you each get 6 months in prison.

If you and Daniel could talk with each other, you'd both realize that option 4 is your best bet for both of you. But you can't talk, so what should you do? Even if you each think you can trust the other not to confess, don't you both have to look out for yourselves first?

The dilemma is that it seems like you'd each, individually speaking, be better off confessing to get a deal, no matter what the other person does—and yet if you both end up confessing, you both end up in a bad situation.

The prisoner's dilemma has sparked a firestorm of debate in game theory, economics, philosophy, psychology, and other disciplines for over 70 years now. One major way to understand the puzzle it presents is that it highlights how hard it is to get self-interested individuals to cooperate with each other for their own common good, even when they're making rational decisions.

In this scenario, we have an approximation of the state of nature. Hobbes thinks that in the state of nature, we can't trust each other. What we need is some third-party mechanism for ensuring cooperation, something that would provide a kind of assurance that we can in fact trust other people to keep their word. What we need is a social contract, Hobbes argues—a general agreement among everyone to set aside our natural rights and to adhere to the judgment of a third party, the "sovereign." For Hobbes, the sovereign could be an individual, as in a monarchy, or it could be a collection of individuals, as in an aristocracy, or it could be all the citizens acting together, as in a democracy. As Hobbes sees it, the sovereign has authority over everyone and looks out for everyone's collective interest.

Mengzi's Sprouts of Goodness

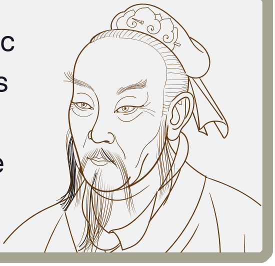
The Confucian philosopher Mengzi, better known as Mencius, flourished about a century before Xunzi and about a century after Kongzi, better known as Confucius. Mengzi famously has an optimistic view of human nature.

According to Mengzi, everyone has a natural sympathy for others, which means that we're fundamentally good at our core. To show this, Mengzi considers a famous thought experiment. Suppose you're walking in a field, and you see a young child on the verge of falling into a dark, deep well. How would you feel?

You'd probably feel a sudden sense of alarm as well as compassion or sympathy for the child. This feeling would arise naturally because feeling this way is part of what it means to be human. Mengzi thinks that feelings like this are the beginnings of altruism; they are the sprouts of goodness that we must cultivate to become truly good people.

But what can we say about someone who feels nothing at all when they see the child? Should we then just assume that everyone will act in their own self-interest? Is it necessarily a bad thing?

Kongzi himself never wrote a systemic treatise about his own views, whereas Mengzi did. Because of this, Mengzi, not Kongzi, is often recognized as the greatest Confucian philosopher.



Rawls's Veil of Ignorance

One philosopher says no—as long as we think about it the right way. In his landmark 1971 book *A Theory of Justice*, John Rawls asks us to imagine another thought experiment—perhaps the most famous thought experiment in political philosophy, variously called the veil of ignorance or the original position. The original position relies on our rational self-interest to consider what sorts of social policies and structures we would agree to. But there's a catch: We would do all of this thinking from "behind a veil of ignorance."

We wouldn't know anything about our "place in society." That is, we wouldn't know anything about things like our social status, economic class, natural abilities, intelligence, personalities, gender, race, ethnicity, or our particular moral convictions. We wouldn't even know our names or who our parents are.

Rawls thinks we'd all pick a society with this basic structure: "All social values … are to be distributed equally unless an unequal distribution of any, or all, of these values is to everyone's advantage." This basic structure, Rawls argues, would provide everyone with social values that he calls "social primary goods," goods like "liberty and opportunity, income and wealth, and the bases of self-respect."

The power of the thought experiment that drives Rawls's *A Theory of Justice* is that we can test different views about what would be just and fair against widely shared intuitions, like, for instance, the intuition that justice is related to equality and that treating people justly means respecting their freedom and independence. The original position allows us to test ideas about justice against our intuitions, modify them, and test them again until we arrive at a "reflective equilibrium," a state where our intuitive sense of what's justice fits with what the theory tells us.

Philosophers like Carole Pateman and Charles Mills have pointed out that these idealized scenarios gloss over deep inequalities that are woven into the ways we tend to think about rights, contracts, and justice. The whole point of the veil of ignorance is to prevent anyone in the original position from adopting any particular point of view, but that imaginary blindfold also has the effect of blocking us off from the very real prejudices we actually have. And if those prejudices infect our intuitions about what's just, then, well, the reflective equilibrium Rawls seeks will be just another way to go wrong.

Reading

- Hobbes, Thomas. *Leviathan*. Reissued ed. Oxford World's Classics. Oxford University Press, 1998.
- Kuhn, Steven. "Prisoner's Dilemma." *The Stanford Encyclopedia of Philosophy*. Winter 2019 ed., edited by Edward N. Zalta. https://plato.stanford.edu/archives/win2019/entries/prisoner-dilemma/.
- *Mengzi: With Selections from Traditional Commentaries*. Translated by Bryan Van Norden. Indianapolis: Hackett Publishing, 2008.
- Rawls, John. *A Theory of Justice*. Revised ed. Cambridge: Harvard University Press, 1999.
- *Xunzi: Basic Writings*. Translated by Burton Watson. New York: Columbia University Press, 2003.

Table of Contents

THOUGHT EXPERIMENTS AS A WAY OF LIFE

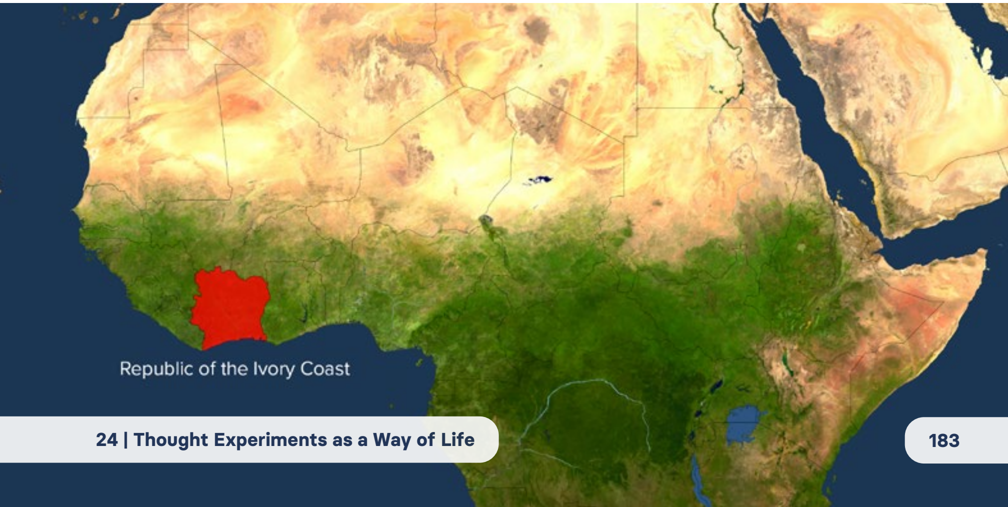
In this final lecture, you will explore dilemma tales and circle back to the general topic this course opened with: ethics. The hope is that you've come to see the value that thought experiments can have in getting us to think differently, and how they're useful in a whole range of areas. They can challenge us, shake us from adhering to the tried-and-true, and maybe even transform how we think about ourselves and the world.

Dilemma Tale

Suppose that a man is taking his sister, his wife, and his mother-in-law across a river in a canoe. The canoe capsizes, spilling everyone into the water. The only person who can swim is the man. However, the current is so strong that he can only save one person. Who should he save?

This is a vivid example of what the folklorist William Bascom calls a dilemma tale. Dilemma tales like this are found throughout Africa. This one comes from present-day Republic of the Ivory Coast, on the southern coast of West Africa. They are short narratives that present their audience with a puzzling scenario that ends with a difficult choice. Sometimes the narrator offers a solution to the problem, but usually the question remains open-ended. Typically, dilemma tales are shot through with moral, social, or legal undertones.

According to William Bascom and other scholars, dilemma tales have important social functions in the communities where they're told. They spark debate, teach about important social norms and cultural standards, and offer a chance to practice moral and even legal reasoning. They can also serve as entertainment.



Although they share features with morality tales and folktales, dilemma tales often have all the hallmarks of puzzling and austere thought experiments. And they are a wonderful example of what it would mean to live with thought experiments. Open-ended dilemma tales are invitations to wonder.

So, who should the man save? We've considered scenarios that involve strangers, but what happens when we must choose between people who are close to us in some way? Which relationships would you privilege over others?

Consider a slightly different dilemma tale. Suppose that a man was traveling with his mother and his fiancée. Their canoe was capsized by a sudden storm. The man was an excellent swimmer, but he could save only one woman. Which one should he choose?

WEIRD Societies

The psychologists Darlingtina Esiaka, Glenn Adams, and Annabella Osei-Tutu note that in scenarios like this, narrators of African dilemma tales typically suggest that the man should pick his mother over his fiancée, or his sister over his wife or his mother-in-law. The "conventional wisdom," they note, "in many African settings is that people should trust and prioritize obligations to kinship connections over mating connections." If this was not your reaction, it might be because you're WEIRD. This acronym identifies people who live in Western, educated, industrialized, rich, and democratic societies.

The anthropologist Joseph Henrich and the psychologists Steven Heine and Ara Norenzayan have suggested that WEIRD people make up as much as 80% of the participants in psychological studies, many of whom are college students. So, we're basing a good number of our conclusions on a group that comprises only about 12% of the world's population. We're not just unrepresentative of human beings as a whole; we're likely significant outliers.

Starting intuitions represent our initial judgments or reactions to thought experiments. Considered intuitions, by contrast, represent our initial judgments that have gone through and withstood critical scrutiny.

Esiaka, Adams, and Osei-Tutu note that conventional wisdom says that people from WEIRD societies “should trust and prioritize obligations to mating partners over kinship connections.”

Dilemma tales help us identify what we care about. They’re an initial point for thinking through starting intuitions to arrive at considered intuitions. It’s not enough to say that it seems right to us to save our mother over our spouse, or our spouse over our mother. We then want to figure out why. What should give us pause is that so many people throughout the world don’t share our WEIRD starting intuitions. Why is that? And whose intuition-driving values are best? Those are questions thought experiments can surely help us explore, even though they can’t answer them definitely.

Moral Responsibility

Moral psychologist Jonathan Haidt suggests that our ethical intuitions arise from various “innate and universally available psychological systems,” which serve as the foundation for moral judgments. Different cultures, societies, religions, and philosophies build their considered views about morality on this foundation. The five central foundations are

Jonathan Haidt

- care and harm, which is associated with virtues like kindness;
- fairness and cheating, which is associated with justice, rights, and autonomy;
- loyalty and betrayal, associated with patriotism and collective interest;
- authority and subversion, associated with respect for tradition and social hierarchies; and
- sanctity and degradation, which is associated with disgust and contamination.

Consider how you'd feel if you let someone borrow something and they ruined it right in front of your eyes. You'd be upset, but there would be more to it than that. You'd resent being treated that way. Resentment plays a big role in recent thinking about moral responsibility. In general, resentment is just a negative emotional reaction to being treated poorly and unfairly—being mistreated and disrespected.

One way of thinking about moral responsibility, tracing back to an influential 1962 paper by the philosopher P. F. Strawson, is that feeling resentment toward someone is a sign that they're responsible for their actions. Sometimes our resentment is misplaced. We might initially resent something but then find out that we were somehow wrong or misunderstood. Perhaps feeling resentment toward that person doesn't makes sense—maybe they're suffering from dementia, for instance, and they're no longer themselves in an important sense. In all these cases, the person might not be morally responsible for what we initially resented them for.

When our resentment isn't misplaced, however, then, according to Strawson, we can say that the person is morally responsible. The idea here is that resentment-worthiness tracks moral responsibility. The standard view is that children aren't resentment-worthy in this sense and, therefore, they aren't morally responsible. Of course, things get complicated when kids get older.

Imagine an elementary school bully, who makes fun of other kids. He's a skilled manipulator. He's often framed another kid for something he's done, and he's frequently invited other kids to his house to play, only to claim, once they've shown up, that they were never invited in the first place. This bully is mean-spirited. Suppose his attention has been focused recently on a kid named Taylor. It's causing Taylor a lot of stress, social anxiety, and emotional pain.

Can a kid fittingly resent another kid in the same grade? The philosopher Samuel Reis-Dennis thinks so. His view is that what makes it fitting for a child to resent another child is that children are social peers who are typically on equal social footing. Bullies do bad things with an ill will that create unjust power imbalances.

That last part, about unjust power imbalances, is why children can fittingly resent other kids but adults can't fittingly resent children. Under normal circumstances, kids don't have social power over adults. If we think that resentment-worthiness reliably tracks being responsible, then this also suggests that children can reasonably hold other children responsible in ways adults can't.

The power of these thought experiments is that they force us to examine our everyday practices and attitudes in much more depth than we otherwise would, and to push us to clarify what we really think. And that's what living with thought experiments is all about.

Jun's Rabbit and Hunter

The final thought experiment is from the contemporary Chan Buddhist Guo Jun's *Essential Chan Buddhism*. Imagine that you're sitting peacefully under a tree when a rabbit darts by you. It turns for a moment and looks at you, and then runs away into the forest. A minute later, a hunter appears and asks you if you've seen a rabbit. Should you lie to the hunter to save the rabbit?

Guo Jun

Although this is a low-stakes scenario, it still presents a powerful puzzle. Lying might seem like the right choice, especially if you value the life of the rabbit. But in the Buddhist tradition, lying is wrong. The moral precepts of Buddhism include right speech, which is about speaking kindly to others, avoiding gossip, and telling the truth, or not lying. Buddhists also think that killing is wrong. Behind both ideas is the view that suffering, in all its forms, is bad, and that what we should do, morally speaking, is try our best to eliminate suffering to the extent we can.

But this means that you should also consider the hunter and his needs. He might be hungry and might have a family to feed. He's also in a bad position. This is because the Buddhist tradition also emphasizes the importance of right livelihood, which is the idea that people should embody their ethical commitments in the work they do. Hunting is a profession that perpetuates suffering in the world through harm, even if it also alleviates suffering by providing food.

Students often say that they would lie to the hunter because they value the life of the rabbit more than telling the truth to some stranger. Few grapple with the real dilemma the scenario poses. They follow their intuitions one way or the other, and then rationalize their choice.

The idea that we can use language or reason to arrive at definite conclusions about difficult matters is something the Chan Buddhist tradition is suspicious of. Chan Buddhism and later Japanese Zen Buddhism—especially the Rinzai sect—are famous for puzzling statements that resist intellectual analysis.

Jun's thought experiment is a little like this. We can think it through, but in the end, being clever or merely justifying our initial intuitions won't work, because the Buddhists can't easily appeal to principles to solve the problem. It takes some creativity and the kind of insight that comes with taking the scenario seriously on its own terms.

Guo Jun offers a solution. "The Chan spirit," he tells us, "is to ask the hunter to sit down and have tea. You ask about his life and his family. … You have a nice conversation and share a few laughs. And when he's ready to start hunting again, you give him a sweet potato."

Jun's suggestion is transformative. It turns out that you can live with integrity even when faced with a dilemma that initially seems irresolvable unless you sacrifice one set of values for another. You don't have to be a Buddhist to appreciate Jun's dilemma or his resolution. We all care about living meaningful, authentic lives. We all care about upholding our moral principles and staying true to our values. We all value the truth and life. We just don't all agree about how things divide up or what matters most to us, of course. And this is where thought experiments can be used to great effect.

Reading

- Bascom, William. *African Dilemma Tales*. Reprint ed. Mouton Publishers, 2011.
- Esiaka, Darlingtina, Glenn Adams, and Annabella Osei-Tutu. "Dilemma Tales as African Knowledge Practice: An Example from Research on Obligations of Support." *Frontiers in Psychology* 11 (2020): 1–9.
- Jun, Guo. *Essential Chan Buddhism: The Character and Spirit of Chinese Zen*. Monkfish Book Publishing, 2013.
- Reis-Dennis, Samuel. "Rank Offence: The Ecological Theory of Resentment." *Mind* 130, no. 520 (2021): 1233–1251.
- Strawson, Peter. "Freedom and Resentment." *Proceedings of the British Academy* 48 (1962): 187–211.

Image Credits

i	©inPhotograph.com
3	Getty Images
4	Getty Images
5	Getty Images
8	David Orban/Wikimedia Commons/CC BY 2.0
9	Berkley Center/Flickr/ Public Domain
12	Getty Images
13	Gage Skidmore/Wikimedia Commons/CC BY-SA 4.0
14	Getty Images
15	Getty Images
16	Getty Images
20	Getty Images
22	Getty Images
23	Wellcome Library, London/ CC BY 4.0
24	Jonah Anderson/Wikimedia Commons/CC BY-SA 4.0
28	Getty Images
29	Matthew W. Hutchins/ Wikimedia Commons/CC BY 2.0
30	Getty Images
31	Vjacheslav Rublevskiy/Wikimedia Commons/Public Domain
32	Nobody60/Wikimedia Commons/CC BY 3.0
33	Getty Images
37	Getty Images
39	Getty Images
45	Getty Images
48	Getty Images
51	Kborer/Wikimedia Commons/ CC BY-SA 4.0
52	Yann/Wikimedia Commons/ Public Domain
53	Getty Images
58	Getty Images
62	Getty Images
65	Getty Images
66	Getty Images
67	Library of Congress, Prints and Photographs Division
69	Getty Images
72	Silver Screen/Flickr/ Public Domain
73	Getty Images
75	The Teaching Company Collection
76	The Teaching Company Collection
77	Library of Congress, Prints and Photographs Division
80	Wellcome Library, London/ CC BY 4.0
80	Wellcome Library, London/ CC BY 4.0
81	Getty Images
82	Wellcome Library, London/ CC BY 4.0
83	Tullio Saba/Flickr/Public Domain
83	German Federal Archives/ Wikimedia Commons/CC BY-SA 3.0
89	Wellcome Library, London/ CC BY 4.0
90	Getty Images
92	Getty Images
93	Getty Images
94	Getty Images

95	Anandajoti Bhikkhu/Wikimedia Commons/CC BY 2.0
98	New York Public Library
99	New York Public Library
100	Getty Images
101	Getty Images
104	Getty Images
105	Getty Images
108	New York Public Library
111	Library of Congress, Prints and Photographs Division
112	Getty Images
113	Getty Images
115	Getty Images
116	Getty Images
117	Getty Images
120	Googol30/Wikimedia Commons/Public Domain
120	Silly rabbit/Wikimedia Commons/CC BY 3.0
121	Getty Images
122	Getty Images
122	Nagelt/Wikimedia Commons/CC BY-SA 4.0
124	Art Uncut/Wikimedia Commons/CC BY-SA 4.0
125	FranksValli/Wikimedia Commons/CC BY-SA 4.0
126	The Teaching Company Collection
129	Getty Images
129	Robespierre 7/Wikimedia Commons/CC BY-SA 4.0
131	Getty Images
133	Getty Images
137	Lliurepensactua/Wikimedia Commons/CC BY-SA 4.0
138	Getty Images
141	Chris McKenna/Wikimedia Commons/CC BY-SA 3.0
144	Basim/Wikimedia Commons/Public Domain
145	Getty Images
146	Getty Images
148	The Teaching Company Collection
149	Getty Images
153	Getty Images
159	Getty Images
160	Getty Images
161	Getty Images
162	Getty Images
163	Getty Images
164	Getty Images
165	Getty Images
167	DaveGettier/Wikimedia Commons/CC BY-SA 4.0
169	A Williamson/Wikimedia Commons/CC BY-SA 4.0
170	Steven HAuse/Wikimedia Commons/Public Domain
171	Getty Images
172	Getty Images
175	Getty Images
176	Marie-Lan Nguyen/Wikimedia Commons/Public Domain
177	Getty Images
179	Getty Images
180	Shutterstock
183	The Teaching Company Collection
185	Miller Center/Flickr/CC BY 2.0
186	Getty Images
187	Yoko santoso/Wikimedia Commons/CC BY-SA 3.0
188	Getty Images

Notes

Notes

Notes

Notes

THE GREAT COURSES®

Printed in the United States of America

4840 Westfields Boulevard, Suite 400

Chantilly, VA 20151-2299

USA

1-800-832-2412

www.thegreatcourses.com