Death: Lecture 4 Transcript
January 25, 2007

Chapter 1. Introduction to Plato's Phaedo [00:00:00]
Professor Shelly Kagan: We've been talking about the question, "What arguments might be offered for the existence of a soul?" And the family of arguments that we're considering initially are arguments that get known as inference or inferences to the best explanation. The thought is that there's something about us that needs explaining. We can't explain it in terms of… in purely physical terms. And so we need to appeal to, we need to posit, the existence of, a soul. Now, I'll come back to that sort of argument in just a minute, but let me bracket that for a moment and say something about Plato.
Starting next week, we're going to be looking at Plato's dialogue, the Phaedo. And so although I'll be saying a great deal about the Phaedo once we turn to it, I want to just take a minute or two and say a couple of introductory remarks. I don't know how many of you have not read any Plato before, but for those of you who haven't, I actually think you're in for a treat. Plato is not only one of the greatest philosophers in history, he wrote his philosophy in the form of dialogues. That is to say, plays, in which various characters sit around or stand around and argue about philosophical positions. The particular dialogue that we're going to be reading, the Phaedo, is set at the death scene of Socrates. As I'm sure you know, Socrates was put on trial, condemned to death for corrupting the youth of Athens — and perhaps, among other things, for arguing philosophy with them. And he's given hemlock, poison, and he drinks it and he dies.
Now, this is a historical event. Socrates had a circle of friends and disciples that he would argue philosophy with. One of his disciples was Plato. Plato then grew up and wrote philosophical works. Plato does not typically appear in his own dialogues. Or, if he does, he's only there as a minor character. In fact, if I recall correctly, Plato's mentioned as not being there on the day that Socrates dies. So, how do we know, if we've got this play, whose position is Plato's position? And the answer — the short answer — is, Socrates, the character Socrates in the play, represents Plato, the author of the play's, philosophical views. Now, in fact, if this were a class in ancient philosophy, we'd have to complicate that picture, because it's fairly clear that by late in Plato's career Plato has philosophical views that are very much unlike the views of his teacher, Socrates. And yet Plato continues to not appear in the dialogue. Socrates continues to be sort of the hero. And so scholars debate which of the views put forward by Socrates in which ones of the dialogues represent views that belong to the actual historical figure Socrates, and which of the views put forward by the character Socrates in which of the dialogues represent views that are actually not held by the historical Socrates, but were instead held by the historical Plato and were merely put in the mouth of the character Socrates.
Scholars distinguish between the early Platonic dialogues, the so-called Socratic dialogues, where the thought is, those are the views of Socrates, the actual historical figure. And then there's the late dialogues, where even though Socrates appears, most scholars believe those are probably not the views that the historical Socrates actually believed. You have middle dialogues where you have to worry about whose views are whose. But we're not going to worry. This is not a class in ancient philosophy. So for our purposes, we don't have to ask ourselves when Socrates in the dialogue says something, is this a view that the dead man Socrates actually would have held or is this simply a view that the dead man Plato put in the mouth of the character Socrates? For our purposes, it won't really matter. I'll take every view that Socrates puts forward as a view of

Plato's, though I'll typically sort of run back and forth sort of in a careless fashion. I'll say, "Plato holds" or "Socrates argues," because for our purposes it's all the same.
But there's one other complication that you've got to be warned about, which is this. Because these are dialogues and they take the form of philosophical arguments, people put forward views and then, over the course of the discussion, change their minds about things. And they take them back. And maybe something similar is going on when Socrates says something. Because, after all, this isn't Plato saying, "Here's what I believe explicitly." He's just writing a dramatic play about philosophy. And so sometimes we'll find ourselves thinking, "You know, there's an argument here that Socrates is putting forward. But maybe it's not a very good argument." And it will, at least, be worth pausing periodically to ask ourselves, maybe Plato realized it wasn't a very good argument. We can often better understand the dialogues by seeing Socrates as putting forward certain positions that he does not think are altogether adequate. And he modifies them or revises them or introduces new positions to deal with some of the difficulties that he was setting himself to be open to earlier. As I say, don't worry about any of those details now, but it's a point to keep in mind as you read the dialogues. So that's all I really wanted to say by way of introduction.
You should start reading the Phaedo for next week. We'll be talking about the Phaedo starting some time next week and we'll continue the discussion of the Phaedo for at least a bit of, maybe all of, the week after that. In the case of Plato, I'm going to make an exception. Normally, I will mention our readings, but I won't spend a lot of time actually discussing them in detail. That's why you have to think of the readings as complementing the lectures or think of the lectures as complementing the readings. I'm not just giving the Cliff Notes, as it were, of the readings. Nonetheless, in the case of the Phaedo, I am going to spend more time actually saying, "Here's what I think the first main argument is. Let's try to reconstruct it in terms of its premises and its conclusions. Here are some objections I raise. Here is then the next argument that Plato offers. Let's try to get that up in premises." Even there, I won't be spending time reading out loud long passages from the Phaedo. But, in some sense, I'll be giving a closer commentary of the Phaedo than I'll do for the other readings. So, still, what you should do is start reading it for next week. The topic of the Phaedo, as I say, is set on Socrates' last day. At the end of the dialogue, he drinks the hemlock and he dies. And perhaps unsurprisingly, what he does with his friends up until that moment is, he argues about the immortality of the soul. Quite strikingly, Socrates is not upset. He's not worried about the fact that he's going to die. He actually welcomes this in a certain way, because he believes his soul is immortal. And so, in addition to philosophical arguments for and against the existence and immortality of the soul, we end the dialogue with a quite moving death scene, one of the great death scenes, if we could call it that, of western civilization. Anyway, as I say, that's all for next week.

Chapter 2. Creativity and Reason in Machines [00:08:27]

So let's return now to the question, "How might we argue for the existence of the soul?" Initially, last time, we considered a set of or a subset of arguments that basically said, "Look, there's got to be more to us than just material objects. People can't just be machines, because machines can't reason. Machines can't think." And I said, "That doesn't seem to be a compelling argument." After all, chess-playing computers, it seems, can reason. They have beliefs about what I'm likely to do next. They have desires about the goals that they're trying to achieve. They reason about how best to defeat me. And it's worth pointing out that — a point that I didn't make last time —

it's worth pointing out that, what the computers, at least the best chess-playing computers don't do. Indeed, no computer actually does this.
You might think that what a computer, what a chess-playing computer does is just this. It calculates every possible branch, every possible game from here on out. And then it sort of works backwards. "Oh, these are the ones where I'll win." And so it only makes the move where it can sort of look ahead 20 moves, right, and see which branches have the computer winning. That is not the way chess-playing programs work. For the simple reason that the number of possible chess games is so huge, that computers can't calculate it. They'd be busy for thousands of years. We can do that sort of: When you play tic-tac-toe with your seven-year old nephew or niece, you just look ahead and work backwards. "Well, if I do that, he'll do that and he'll do that and then he wins, so I won't do that," right? But we can't do that with chess. There's just too many games.
So how do chess-playing programs, and particularly the best chess-playing programs, how do they work? Well, they play chess the same way you do. They have various ideas about which pieces are more powerful and so they're more important to protect. They've got various ideas about which strategies tend to be successful. What sorts of dangers come along with them? If you're a serious chess player, you might study some of the great games of chess history. And indeed, when they program these things, the programmers will feed in game after game after game of the great chess games in history. And then armed with all of that, you sort of do your best. And when you lose a game, you kind of make a mental note to yourself, "That really screwed me up. Let me try something different next time." And you avoid those sorts of moves. That's how chess-playing programs work as well.
Jumping ahead, let me make a remark about this, because this is going to be relevant for something I'll get to in a couple of minutes. What this means — what this, the implication — is that if you're playing a great chess-playing program, it's not as though the way to tell what it's going to do is to study its program and think it through. The people who design these programs, presumably fairly decent chess players themselves, the people who design these programs, when they're playing the programs they're not thinking to themselves, "Let's see. I programmed this computer so that when I move a queen forward to this space, it should come out with a bishop." That's hopeless. Because the program is constantly revising its strategies, in light of what's worked and what hasn't worked in the past. When the programmers play these programs or indeed when anybody, a good chess player, plays these programs, the best way to try to beat them is simply ask yourself, "What's the best move to make right now?" The odds are the computer's going to make the best possible move. Treat the computer as though it were just a great chess player.
And indeed, the best programs are great chess players. There was a period of time in which, although there were decent, chess-playing programs couldn't beat the best chess-playing humans. That ended some years ago when the best programs began to beat grand masters. And now it's in fact the case that the best programs can beat pretty much anybody. In the current world champion of chess, I think Vladimir Kramnik, was defeated in December by a chess-playing program. So Kramnik's simply treating this as an awesome opponent. And that's the best way to deal with these things. All right. So, bracket some of those thoughts for a moment. We'll come back to them a little bit later when we start talking about the question, "Could machines be creative?" Tipping my hand, it seems pretty clear that that seems like the right thing to say about these chess-playing programs.

Chapter 3. Feelings in Machines, from Marvin to Hal [00:13:43]
So we had the question, "Could machines, could machines reason?" And although we don't have machines that can reason about a lot of subjects yet, it seems pretty clear. It seems like the natural thing to suggest, machines can reason in at least some areas. And so it doesn't seem plausible to suggest that we people must not be physical, merely physical, because after all, we can reason and no machine can reason. No, machines could reason.
But this prompts a different move on the part of the defender of souls. Perhaps the argument shouldn't be, "we have to believe in souls because no mere physical object could reason." Perhaps the argument should be, "we have to believe in souls because no mere physical object, no machine could feel." You know, we have emotions. We love. We're afraid. We're worried. We'll get elated. We get depressed. So perhaps the argument should go "Yeah, yeah, thinking, that's the sort of thing a machine can do. You know, we call them thinking machines. But feeling, that's the sort of thing no machine could do. No purely physical object could feel anything, could have emotions. And so, since we clearly do feel things, there must be more to us than a physical object."
Now, I think it is plausible to suggest that unlike the case of chess-playing computers, we don't yet have machines that feel things. But the question isn't, "do we?" The question is, "could there be a machine that could feel something, could have an emotion of some sort?" So let's go a little science fictiony and think about some of the robots that have been shown in science fiction movies, some of the computer programs that have been shown in science fiction movies, science fiction novels, or what have you. When I was a kid there was a television show called Lost in Space. I'm afraid I've forgotten the name of the robot that was on that show. But as it was a TV show and so sure enough, every single episode, some new dramatic danger would take place. And the robot would start whizzing and binging and shout out, "Danger, Commander Robinson!" "Danger, Will Robinson!" that was it. "Danger, Will Robinson!" It seemed as though the robot was worried.
More recent example. A number of you have probably read some of Douglas Adams' books The Hitchhiker's Guide to the Galaxy and the sequels to that. There's a robot in those books, Marvin, who's — depressed, I think is the simple word about it. He sort of — He is very smart. He's thought about the universe, thinks life is pointless and he acts depressed. He talks to another robot, depresses the other robot. The other robot commits suicide. All right. Seems natural to ascribe depression to Marvin, the robot. That's how he behaves. Or, my favorite example, the movie 2001: A Space Odyssey. Now, I've got to tell you, for those of you who have not seen this movie, I'm about to spoil it. All right? So you cover your ears.
In 2001: A Space Odyssey, we get some kind of indication that there's life on another planet. It's all very mysterious and we send off a spaceship to investigate the markings, the radio signals from the other place. This is a very important mission and so there is a computer program named Hal that helps run the ship and takes a lot of the burdens off of the part of the human astronauts who are on the ship. Hal's got the goal — in terms of reasoning and desires and so forth and so on — Hal's got the goal of making sure the mission is successful. But Hal thinks to himself fairly plausibly, humans really screw things up. This is a very important mission. Let's kill the humans to make sure they don't screw things up. One of the astronauts, discovering the plot, attempts to stop Hal. And proceeds to do the only thing he can do to defend himself against Hal, which is shut down the program, basically killing — if we can talk that way — killing Hal. Meanwhile, as all this is going on, Hal and Dave, the human astronaut, are talking to each other. Hal realizes

what's going on. Hal tries to stop Dave, understandably enough. And Hal says, as Dave begins to shut down Hal's circuits, "I'm afraid. I'm afraid, Dave." What's he afraid of? He's afraid of dying. It seems perfectly natural to ascribe fear to Hal. Hal is behaving in exactly the way you would expect him to behave, or it to behave, if it felt fear. It's got reason to be afraid. It's behaving appropriately. It's telling us that it's afraid. It seems natural to say Hal's afraid.
Now, you could continue to sort of fill in examples like this. As I say, of course, they're all science fiction, but the fact that we can grasp — and it's not as though we go running away saying, "Oh no! This was outrageous," right? "It makes no sense to think a computer could have said, 'I'm afraid.' It makes no sense to think that it could try to kill the people who are trying to shut it down and so forth." That seems to me to be prejudice, as I said last time. The natural inclination here is to say, "These computer programs, these robots are feeling emotion." But there's no particular reason to think there's anything going on there than the circuits. They're just physical objects, programs on machines. If that's right, if that's the right thing to say, then what we have to say is, "We don't need to appeal to souls in order to explain emotions and feelings. Physical objects could have, mere physical objects could have, emotions and feelings. So we have no reason to posit the existence of a soul."

Chapter 4. Qualia in Emotion and Consciousness: The Dualist's Defense and Its Weakness [00:20:34]

Now, I think the best response on the part of the dualist to this reply is to distinguish two aspects of feelings, two aspects of emotions. There's the behavioral aspect of feeling fear, let's say. The behavioral aspect is when you're aware in the environment of something that poses a danger to you, that will harm you or destroy you, or in the case of a computer program, turn you off, then you take various kinds of behaviors in opposition to that to try to disarm the danger, to try to neutralize it. This is just a matter of beliefs, goals, responses, planning, the sort of thing that we already saw the chess-playing computer can do, that behavioral side of emotion. It seems pretty plausible to think robots could do that. Physical objects could do that.
But, and here's the crucial point of this objection, there's another side or another aspect to emotions and feelings. It's the sensation of what it's feeling like — that's why we call them feelings after all — what it's feeling like on the inside, as it were, while all this behavioral stuff's going on. When I'm afraid, I have this certain sort of clammy feeling or my heart's going poundingly. Your blood is racing. When you're afraid, you've got this sinking feeling in the stomach. When you're depressed, there are these, well, we could call them experiences, though the word "experience" is also somewhat ambiguous. So — we'll use it for the moment — there's an experience that goes along with each emotion. There's what it feels like to you when you're afraid. What it feels like to you when you're worried or depressed or joyful or in love. And the thought, and I think this is a pretty powerful thought, is that even if the robots are behaving behaviorally, they've got the behavior side of the emotions down, they don't have the feeling side at all.
Now, once you start thinking these thoughts, there's no need to restrict yourself to emotions. The missing stuff, the missing thing is there in all sorts of familiar humdrum ways as well. So right now I'm looking at the chairs in the auditorium. They're some kind of shade of blue. Think about — Look at some places in the room where the curtains with their red. Think about what it's like to see red, the sensation of seeing red. Now again, we've got to distinguish between what I'll continue to call the behavioral side of seeing red and the experiential side of seeing red. It's easy

enough for us to build a machine that can tell red from blue. It just checks and sees what kind of light frequencies are bouncing off the object. So we can build a machine that could sort red balls from blue balls. My son has a little robot that can do that. Still, when you think to yourself, what's the — what's going on inside the machine? What does it feel like to be the machine while it's looking at — while it's got its little light sensors pointed at — the red ball? Does it have the sensation of seeing red? What I suppose you want to say, certainly what I want to say is, "No, no, it doesn't have that sensation at all." It's sorting things based on the light frequencies, but it doesn't have the experience of seeing red.

What we're trying to get at here is — it can be very elusive, but I imagine most of you are familiar with it. It's the sort of thing you wonder about when you ask yourself, "If somebody was born blind, could he possibly know what it's like to see color?" He might be a scientist and know all sorts of things about how light works. You use such and such frequencies, and which objects, and you hand him an apple and he'll say, "Oh, it must be very red," right? Maybe he points his little light detector at it and it reads out. It says, "This is such and such a frequency." And he says, "Oh, this is a very red apple, much redder than that tomato" or whatever. But for all that, we've got the notion, not only is he not seeing red, he can't even imagine what it's like to see red, never having had these experiences. And once you start to see this, we realize, of course, our life is filled with this aspect. Things have colors. Things have sounds. Things have smells. There is the qualitative aspect of experience.

And the point that I started with earlier, about the internal aspect of emotions, is it's not just out there, but inside as well. We have certain kinds of sensations inside our body, the characteristic sensation of fear or joy or depression. All right. So the suggestion then might be this. What no physical object can get right, because no physical object can get at all, is the qualitative aspect of experience. That's the aspect that we're after when we ask ourselves, "What's it like to see red? What's it like to smell coffee or to taste pineapple?" Now, it's pretty — Philosophers sometimes call these things qualia, because of the notion of the qualitative aspects of things. Our experiences have qualitative properties. And the suggestion then might be, no physical object, no mere machine could possibly have qualitative experience. But we've got it, so we're no mere physical object. We're no mere machine.

All right. Now, that's the objection. It's a pretty good objection. And then the question is, "What can the physicalist say in response?" Now, the best possible response would be for the physicalist to say, "Here's how to build a machine that can be conscious in this sense. That is, have a qualitative experience. Here's how to do it. Here's how to — Just like we can explain in materialist, physicalist terms how to get desires and beliefs and the behavioral stuff down, here's how to get the feeling, qualitative aspect of things down, too." It would be best if the physicalist could give us that kind of story. I think the truth of the matter has to be — I think the answer right now is, we don't know how to give that story. Consciousness, if what we mean by consciousness is this qualitative aspect of our mental life, consciousness remains a pretty big mystery. We don't know how to explain it in physicalist terms. And it's because of that that I think we shouldn't be dismissive of the dualist when the dualist says, "We've got to believe in souls in order to explain it."

We shouldn't be dismissive, but that's not to say that I think we should be convinced. Because it's one thing to say we don't yet know how to explain consciousness in physical terms. It's another thing to say we won't ever be able to explain consciousness in physical terms. If we had the latter — excuse me — If we had the bold claim that no physical object could see red, taste honey, then we'd have to conclude since we can do all that, we're not a physical object or not merely a

physical object. But I don't think we're yet in a position to say that. I think the simple fact of the matter is we don't know enough about consciousness yet to know whether or not it can be explained in physical terms.

When I think about this situation, an analogy always occurs to me. Imagine that we're somewhere in, let's say, the fourteenth century trying to understand life, the life of plants. A plant is a living thing. And we ask ourselves, "Could it possibly be that life could be explained in material terms?" It's got to seem very mysterious to us. How could it be? When we think of the kinds of examples of material machines that we've got available to us in the fourteenth century, I try to imagine what would somebody in the fourteenth century think to himself or herself when he entertains the possibility that a plant might just be a machine? And then, I have this little image of some plant made out of gears, right? And the gears begin turning and the bud opens, dot, dot, dot, dot. And the person's just going to say, "My god! That wouldn't be alive!" So it's pretty obvious that no machine could be alive. No material object could be alive. In order to explain life, we have to appeal to something more than just atoms. They didn't have atoms, but more than just matter. Life requires something immaterial above and beyond matter to explain it. That would have been an understandable position to come to in the fourteenth century, but it would have been wrong. We didn't have a clue back then how to explain life in material terms. But that didn't mean it couldn't be done. I'm inclined to think the same thing is true right now for us and consciousness. I know there are theories out there. But my best take is we're pretty much like in the fourteenth century. We don't really have a clue yet, or not much of a clue, as to how you could even so much as begin to — it's not that merely that we don't have the details worked out. We don't even have the picture in broad strokes as far as consciousness is concerned, of how it could be done in physical terms.

But not seeing how it's possible is not the same thing as seeing that it's impossible. If the dualist comes and says, "Can't you just see that it's not remotely possible, it's not conceivably possible, for a purely physical object to have experiences, to have qualia?" what I want to say is, "No, I don't see that it's impossible. I admit I don't see how to do it, but I don't see that it's impossible." So I don't feel forced to posit the existence of a soul.

Of course, the fan of the soul could come back and say, "But that's not fair. The question isn't, 'Is this explanation impossible?' The question is just, 'Who's got the better explanation?' You guys can't offer any kind of explanation at all, yet. I can offer an explanation. How is consciousness possible? We have souls. Souls are really very different from physical objects and so they can be conscious."

But at this point, I think it's crucial to remember the point that it's not just the question, "Who's got an explanation?" but, "Who's got the better explanation?" And before we say that the soul view's got the better explanation, we have to ask ourselves, just how much of an explanation is it to say, "Oh I can explain consciousness. Consciousness is housed not in the body, but in the soul." Okay. "How exactly is it that a soul can be conscious?" we ask. And then the soul theorist says, "Well, uhm.. er.. ah.. it just can." That's not really much of an explanation. I don't feel I've got any sort of account going here as to how consciousness works, even if I become a dualist. If the dualist were to start offering us some elaborate theory of consciousness, "Well, there's these sorts of soul structures, and those sorts of soul structures, and these create these sensations and those create those sensations. And here's a theory," well, then, I'll begin to take it seriously as an explanation. But if all the soul theorist is just saying is "Nah, nah. You guys can't explain it and I can, because I say this is an explanation." then, I find myself wanting to say, "That's not really any better. That's no improvement at all."

There was a question or a comment.
Student: [inaudible]
Professor Shelly Kagan: Good. So the question was — First, it was the accusation before the question, that I'm holding a sort of double standard. I'm defending, I'm defending the physicalist by saying, "Don't blame us. We don't know how to explain it yet." Why aren't I allowing the soul theory to say, "Don't blame us. We don't know how to explain it yet." Good question. And my answer is — And sometimes I think this one's a tie. I think the soul theorist doesn't have an explanation. The physical theorist doesn't have an explanation. As far as I can see, right now nobody's got a good explanation about how consciousness works. It's a bit of a mystery right now. So I don't mean — I hope I haven't been doing this. It's not so much a double standard is needed; it's a tie. But notice if it's a tie, that doesn't give us what we were looking for. What we were looking for, after all, was some reason to believe in souls. And if the best the soul theorist can say is, "I can't explain it and neither can you," that's not a reason to believe this side.
We already believe there are bodies. We already know bodies can do some pretty amazing things. The question we're asking is, "Is there a good reason to add to our list of things there are? Is there a good reason to add the soul, something immaterial?" And if the best that the soul theorist has is, "maybe we need this to explain something that I don't see how you guys can explain, maybe this would help, though I can't quite see how either," that's not a very compelling argument. So what I'm inclined to think with regard to this particular strand or this particular version of the argument is, the jury's still out. Maybe at the end of the day we'll give it our best. We'll decide you can't explain consciousness in physical terms. We'll begin to work out some sort of alternative immaterial theory. Maybe at the end of the day we will decide we need to believe in souls. But right now, I don't think the evidence supports that conclusion.
Still, there's other possibilities. Consider creativity. Here's another version of an argument that goes from inference to the best explanation. Creativity. It says, "People can be creative." We write new pieces of music. We write poems. We prove things in mathematics that have never been proven before or we find new ways to prove these theorems or what have you and we can be creative. No mere machine can be creative. So we must be something more than a mere machine. Well, then, the question is going to be, "Could it be a case that there could be a physical object that's creative?" And I'm inclined to think, "Yes." In fact, I already suggested as much when I talked about the chess-playing computers. The chess-playing computer programs think of moves, think of strategies no one's thought of before. In the most straightforward natural meaning of the term, we have to say — I think the program that beat the world champion was called Deep Fritz. So when Deep Fritz beat Kramnik, it was being creative. It made a move that Kramnik didn't think of and perhaps nobody had. Perhaps no chess game before had had this move.
Computers can do other sorts of things of this sort. There are mathematical theorem-proving programs. Now, some of these things can prove things that are mathematically way over my head. But let's take something simple like the Pythagorean Theorem, which we all learned in high school. And we learned how to prove the Pythagorean Theorem in Euclidean geometry, starting with the various axioms in Euclidean geometry, ba, ba-ba, ba-ba, ba-ba, ba bum. This proves Pythagorean Theorem. And it turns out there's a variety of proofs of the Pythagorean Theorem. And, in fact, a computer program has come up with a proof that, as far as was known, nobody in the world had ever come up with before. Well, other than prejudice, what would stop us from saying the program was being creative? Not just in sort of mathematical things like chess or math, there are, as you know, programs that can write music. And I don't just mean throw out

some random assortment of notes. Programs that can produce music that have — that we recognize as music, that have melodic structure and develop themes, resolve, music that nobody's heard before. Why not say the machine is creative? What, other than prejudice, would stop us from saying that? So if the argument's going to be, "We need to posit the existence of a soul in order to explain creativity," again, that just seems wrong. Well, there's a — question, comment?
Student: [inaudible]
Professor Shelly Kagan: Good. The question was, "When I talk about creativity here, am I trying to build in some appeal to the feeling that we may have when we're being creative?" And the answer is, "No." All I had in mind, as you know, is just — in talking about the creativity issue — I just have in mind producing something new, producing something that hasn't been around before. And most particularly, producing something that your programmers didn't already have in mind. Remember, it's not as though the people who designed the chess-playing programs can beat it. The chess-playing program makes moves these guys haven't thought of. All right.

Chapter 5. Free Will as a Defense of the Soul and Conclusion [00:42:20]

The creativity argument may not work, but there's something that sort of immediately comes on its heels. Even if we could build a program, even if we had built programs that can be creative, that can do things that nobody's thought of before, all the program is doing is following its program, right? It's just a series of lines of code. And the robot or the computer or what have you is just automatically, mechanically following the code commands of the program. We might say, even if we are smart enough to build programs that can, by mechanistically following the program, do things we've never thought of, still all the computer can do, all the robot can do is automatically, necessarily, mechanically follow the program. It doesn't have free will. But we have free will.
So, here's a new argument for the existence of the soul. People have free will. No merely mechanical object, no robot, no computer could have free will. But since we've got free will, we must be something more than a merely physical object. There must be something extra, something immaterial about us, the soul. So maybe that's why we need to believe in souls in order to explain free will. Now, the subject, free will, is a very, very — The subject of consciousness is a very complicated. One could have an entire semester devoted to thinking about the philosophical problem of consciousness. And indeed, as it happens, in our department this very semester there is such a class devoted, all semester, to the topic of consciousness. One could similarly have a course devoted to the problem of free will. I'm going to spend all of two minutes on it. So it's by no means do I mean to suggest, "Oh, here's everything you need to know about the subject." I simply want to point out enough about the problem to help you see why I don't think free will is a slam-dunk for the soul.
So what's the argument? Well, the thought seems to be something like this. [See Figure 4.1] One, we have free will. Two — let me say something about this. What is it about the thought that the computer is just following a program? Well, the thought, I suppose is, in philosopher's jargon, that the computer is a deterministic system. It follows the laws of physics and the laws of physics are deterministic. If you're in this state, you will necessarily, given the laws of physics and the way the computer's programmed and built and so forth, these wires will turn on, turn off, these circuits will turn on, turn off, boom, suddenly you'll be in that state. There are certain laws such that, given that the computer's in this state, it must necessarily move in that state. When you've

got a view about cause and effect that works this way — for everything that happens, there's some earlier thing that caused it to happen such that given that earlier cause, the event had to follow — that's a deterministic picture. And the thought, of course, is that the robot or the computer is a deterministic system and you can't have free will if you're a deterministic system. So number one, we have free will. Two, nothing subject to determinism has free will. Put one and two together. It follows, if nothing subject to determinism has free will, but we have free will, it follows that we're not subject to determinism. Suppose we then add three, all purely physical systems are subject to determinism. Well, one and two gave us that we are not subject to determinism. Three says, all purely physical systems are subject to determinism. Well, it would follow then from one, two, and three that we are not a purely physical system. So, conclusion, four, we are not a purely physical system.

All right. That's the argument from free will. Now, the argument is valid. That's philosopher's jargon, that is to say, given the three premises, the conclusion really does follow. The interesting question is, "Are the three premises true?" And they've got to all be true. It's got to be that every single one of them is so. I'll just spend a minute more on this starting next time. But the point to think about for next time is just, is it really true that all three of the premises are true, or might one or more of them be false? All right. That's where we'll start next time.

[end of transcript]