Transcript: Philosophy VIDEO 2.1 – Let's Play Cards, And Think About Confirmation Bias

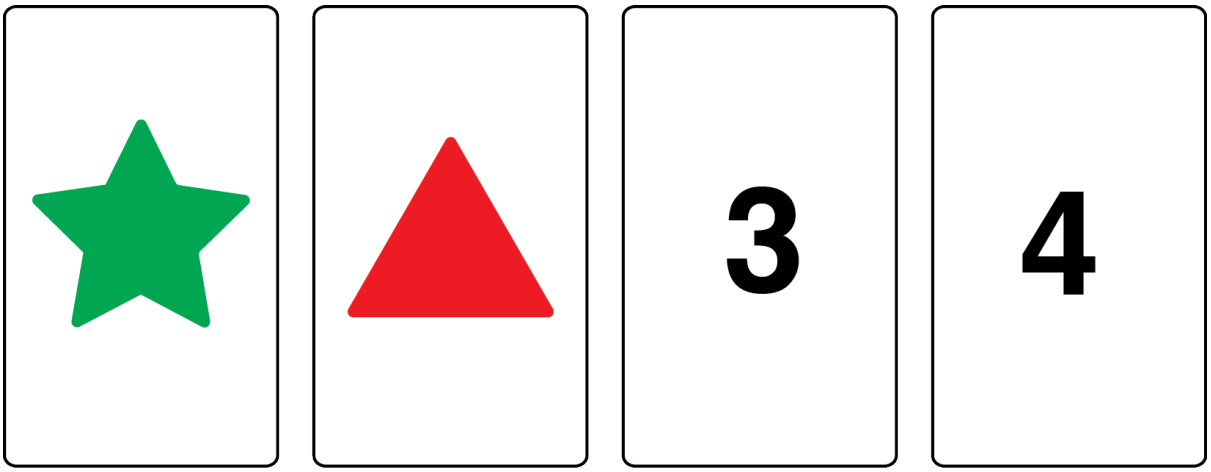
OK, here's the game. I have a deck of cards. All the cards in my deck display a colored shape on one side and a number on the other.

Colored Shape on one side, number on the other. This is a story-problem given. To repeat: all cards have a colored shape on one side, a number on the other.

Now, I am going to state a rule, and I am going to lay out four cards from the deck. You have to figure out if the cards follow my rule. You are allowed to flip cards, as needed, to do that. As you might guess, the challenge is: make as few flips as possible. That is, make all the flips that you need—no more—to verify that the rule holds true for these cards or does not. Let's begin.

Here's the rule:

If a card has a star on one side, then it has an even number on the other side.



Green star Red triangle 3 4

Pause the video and think about it. Which cards to do you need to flip to test it?

Answer: you have to flip the green star and the 3

Did you make a mistake? Sorry about that. But, if it's any consolation, most people get it wrong, really by an overwhelming margin. That's why I asked. I'm playing psychology games with you. Speaking of which:

This card game is an example of something called the Wason Selection Test, or Task. It was invented by a psychologist named Wason. It's psychologically interesting because the task should be easy. But it isn't. So what gives?

Let me talk you through the problem logically, then psychologically, then we'll try another round and see if it gets easier. (If you got it right the first time, good for you. You can just follow along with the explanation, feeling smug. Which is a great feeling! But you should still be able to learn about psychology, even if you did get it right.

The rule again:

If a card has a star on one side, then it has an even number on the other side.

The rule is a conditional. An If-then statement.

Formally, If P -> Q. Which, by the by, is equivalent to: -Q -> -P. That is,

If there is an odd number on one side, then a card does not have a star on the other.

That rule is logically equivalent to our rule. What do I mean equivalent? Just this: those two rules—or statements, call them what you will—are true or false, followed or broken, together. If one is true, the other is. If one is false, the other is.

Maybe you would like to learn a new word: **contrapositive**. That second rule is the **contrapositive** of the first.

[take a conditional (P -> Q) statement, negate both sides, flip their positions (-Q -> -P).]

Contrapostive means: I took both sides of a conditional IF-THEN statement, negated both, AND flipped their positions.

The word ‘contrapositive’ doesn’t get out much. I’m a professional philosopher. I can’t remember the last time I used it before making this video.

Still, the fact that a conditional shares truth-value with its contrapositive is conceptually handy. Logical equivalence is often handy for truth-test purposes. And, this is important, it hurts our brains a bit, right? That those two rules are effectively the same rule does not just jump right out at you. (Unless you are truly unusually good at logic.)

It doesn’t help that my original rule is not intuitive in the least. Why should stars have anything to do with numbers? No reason. So forget cards for a second. How about this:

If an animal is a mammal, then it is warm-blooded.

Intuitive. True.

That logically equals:

if an animal is not warm-blooded, then it isn’t a mammal.

The contrapositive.

This statement is also obviously true, I hope you agree. But you might scratch your head over my claim that they are ‘equal’. ‘Equal’ meaning: they literally say the same?

Suppose someone asked you: “What is a mammal?” You might answer: “If an animal is a mammal, then it is warm-blooded”. That isn’t exactly a definition, but it feels definition-like. It’s informative and to the point.

By contrast, if someone asked, “What is a mammal?” And you answered: “If an animal is not warm-blooded, then it is not a mammal” you would sound like some sort of Martian. It sounds backwards and not to the point.

And yet: the statements are equivalent at least in the following sense. There is no way their truth-values diverge. If one is true, the other is. If one is false, then other is. Pause the video and think about it if you have

to.

On the other hand there are cases in which the equivalence relation is more intuitive. Suppose someone asks: “What does ‘healthy’ mean?”

You might reply: If an animal is healthy then it is neither sick nor injured.

Or you might say:

If an animal is either sick or injured, then it is not healthy.

Maybe the first answer is a bit clearer, but they both feel OK as an introduction to the concept of health. Right?

A lot of interesting semantic issues take off from this point. But, for present purposes, let me just say: if you were a creature who found it intuitive that conditionals always share truth values with their contrapositives, you would probably find the Wason Selection Task easier. But we humans … are mostly not that creature.

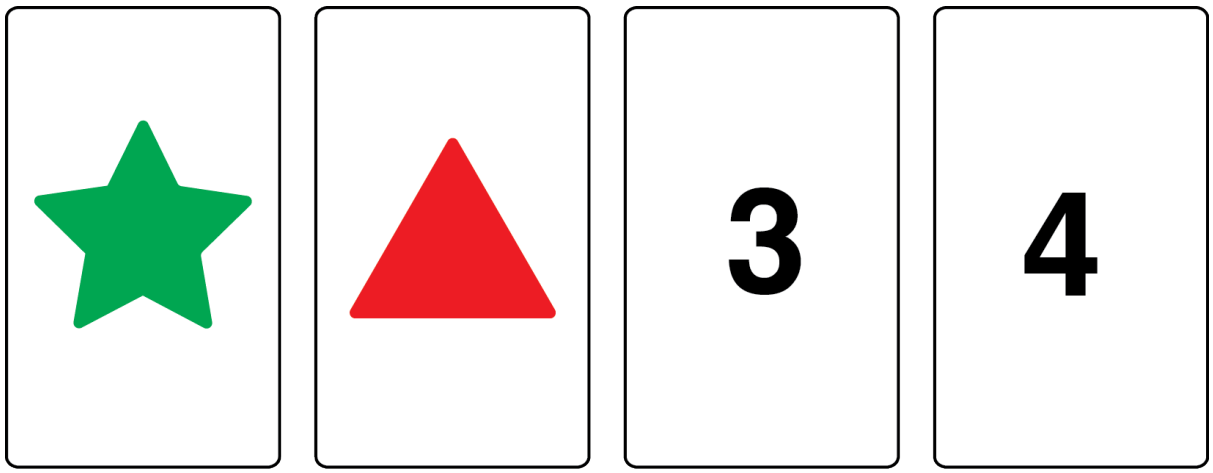
Let’s delve more directly into psychology. And pay attention. It’s going to start getting more interesting than just baby logic. I’m going to teach you new words you should totally be using all the time.

Our rule, once again:

If a card has a star on one side, it has an even number on the other side.

[slide18 cards]

And our cards are, once again: Green star/Red triangle/3/4



The mistake people most commonly make is to waste time flipping the 4 to see if there is a star on the other side.

The reason that's a rookie mistake is that there is no way that flip is going to be informative.

Why not? Just think it through.

There's no way for any result here to BREAK our rule. Either there's a star on the other side, in which case the rule is good. Or there isn't a star, in which case the rule is also good. Our rule doesn't say anything about what goes with non-star shapes, after all.

Why are people attracted to a wasted flip of a move?

Well, the mind is a funny thing. But I think the following is a helpful generalization. (Not a universal truth. But a likely circumstance.)

If you flipped the 4, you are probably falling prey to 'confirmation bias', a term uncoincidentally coined by that self-same psychologist, Peter Wason.

What's 'confirmation bias'?

It's the tendency to look first—look more—for evidence that confirms or supports our beliefs, rather than to for disconfirming evidence that undermines our beliefs.

'Confirmation bias' is also known as 'myside bias', for reasons that are obvious. Whose side do I tend to take in an argument? My side, that's whose! I look for reasons why I am right, not reasons I am wrong.

But what does this have to do with our card rule?

The trouble with flipping the 4 is that it can only show the rule is right, not wrong. So obviously if I were very attached to the rule, very invested in finding it to be right, I might be attracted to the 4. But you aren't biased in favor of the rule, are you? 10 minutes ago you had never heard of it. It's not YOUR rule. You have no stake in it turning out right. Who cares?

That's plausible. Even so, here are two models of how confirmation bias sneaks in.

First, we only have one hypothesis here. One candidate rule. People like to have a side. If there's only one side visible, you will tend to make it your side. Strange but true, your weird brain may start to look for reasons why this rule is right, not wrong, from the second you hear it.

But I think the following is an even likelier model of how confirmation bias slips in. You see the cards. You hear the rule. Your mind—your brain—makes a snap judgment about what is *relevant* to solving the problem. The rule says stuff about stars and even numbers. So: cards with stars and cards with even numbers are probably relevant. And, now that you have a hypothesis, confirmation bias makes you look more for reasons why you are RIGHT about what's relevant. Maybe you say to yourself: If there's a star on the other side of that 4 card, that would confirm me! That feels like making sense.

Of course, that's wrong. We already worked that out.

Thinking contrapositively would probably help. It might counterbalance wrong instincts. But Think Contrapositively!—although it might be a funny t-shirt tagline—is not the true cure here.

The true key is **disconfirmation**. To play this game and win, we must resist any temptation to ask what card flips might confirm the rule. We should only ask: what flips might disconfirm it? Those are the only ones you care about. Why? Because it doesn't matter if some cards fit the rule. All that matters is if at least one doesn't fit it.

The only combo that disconfirms P -> Q will be P & - Q. That is, star plus odd. So: ignore every card that isn't showing either a star or an odd number. Thus, you only need to flip the star and the 3.

I know, I know, I kind of took the long way. But here's the short way: disconfirm. Try to break the rule.

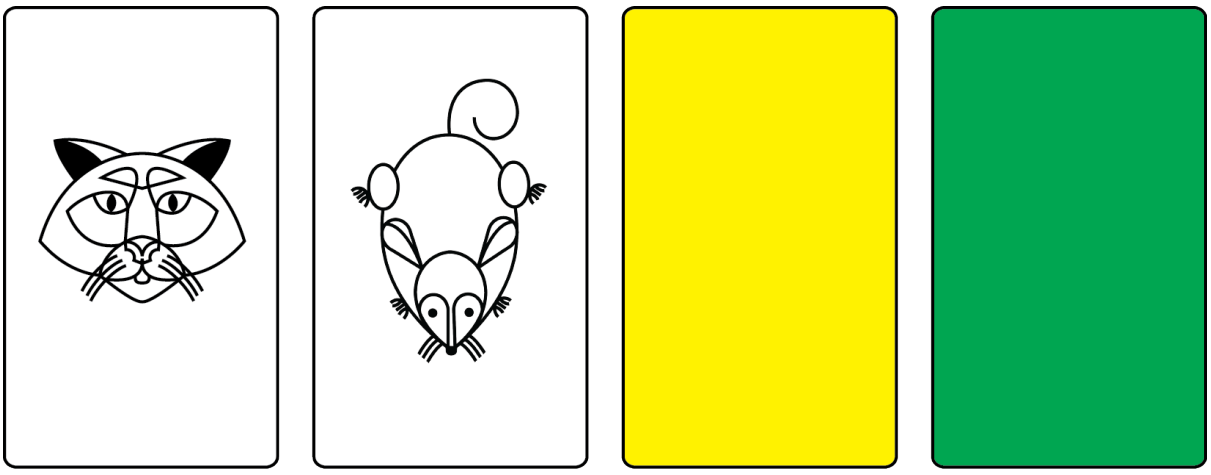
Let's try it again, shall we. Same game. I'm going to vary the deck, to keep you on your toes. This deck has cards with animals on one side, colors on the other.

But now you know the winning strategy. Don't confirm. Don't get distracted by stuff that would 'fit with the rule', as if that matters. It doesn't. Hear the rule. Pre-formulate any result that would break it. Look for the rule-breaker.

The rule:

If a card has a dog on one side, it is yellow on the other side.

The cards:



Cat mouse yellow green.

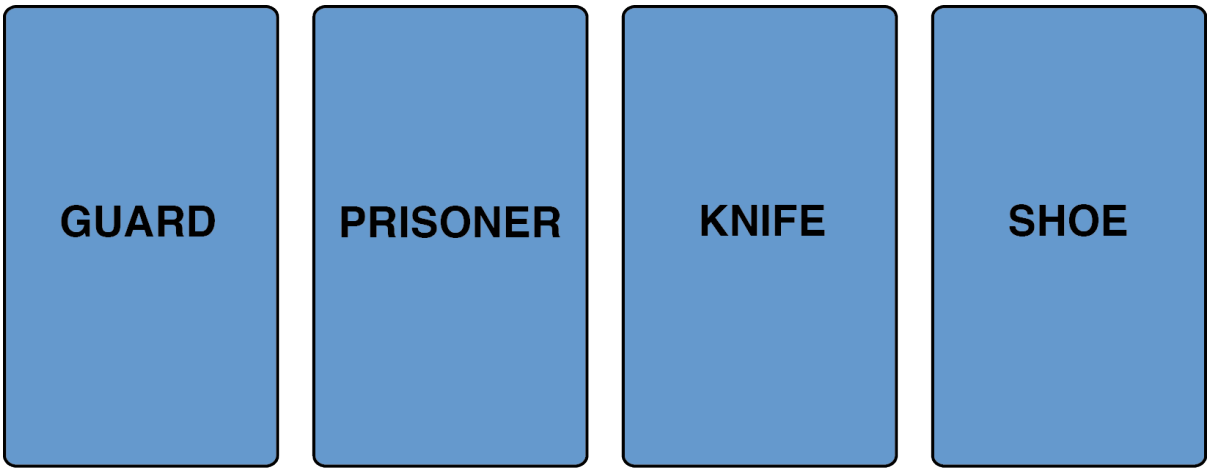
Did you get it?

You only have to flip the green one. I explained the first one problem at length. I won't go through and beat this second one to death. If you had trouble, think about it until you get it.

One more game. Now you are - oh, let's say - a prison guard whose whole job is make sure that the following rule is followed:

if someone is a prisoner, then that person is only allowed to have items that are not weapons.

No cards this time - real life! serious business! - but we may represent it as a card game. Each card has a person on one side, on the other an item they are holding. Flip all and only the cards you need, to make sure you are doing your job as a guard.



I'll bet you got it right. Prisoner and knife. You should search the prisoner and check that the knife isn't held by a prisoner. You can safely ignore your fellow guard and any shoes in the vicinity. (I'm assuming shoes aren't weapons.) The reason this task was easy was not because I taught you that supervaluable trick about seeking disconfirmations—although that trick works here, too. Nope. I'm sure you didn't need it. The scene made social sense to you, and we humans are very good at sizing up the social scene.

Let me teach you another new word: **heuristic**. That just means: rule of thumb. Or bias, if you prefer. Bias sounds bad, but bias isn't always bad. In many circumstances you are biased in favor of extra safety, for example. In many circumstances, that's a very good thing. (Of course, sometimes it can make you paranoid, sure.) Heuristic may be a better term than bias because it's more neutral-sounding.

Anyway, as a card-carrying human being—I assume that's you—certain

cognitive heuristics that may be precisely what allow you to navigate complex social settings effortlessly, may predispose you to be bad at certain logic games. Isn't that weird? But it's true. I'll bet it is. Mountains of psych test data say so.

This is a big thought and we'll come back to it. There's a fine line between the mind working optimally, in most cases, and the mind messing up, in some cases.

But, just for now: what does it have to do with questions? Card games and contrapositives, confirmation and counter-examples. Isn't this supposed to be a question module?

I'm glad you asked that question. It brings us to our next video.

I'll just give you a taste right here at the end

Remember what I asked in the previous video? What good is Q? Why are we making you take this module?

Let's rephrase that. Since we are making you take this module, the following conditional had better be true, as a rule:

If you take Q -> You get smarter.

That better be it. Otherwise, what's the point?

I could do the next bit with a fresh deck of cards. Lightbulbs are the universal symbol for sudden smarts, so we could use that. On one side, these cards either have a lightbulb or—some other common household item. On the other side there are letters. Q, of course, is the letter we care about. Our plan for the course could be stated as a rule:

if Q on one side, then lightbulb on the other.

We make you smart.

I could ask you which cards you need to flip to test the rule. And you now know that the trick is to try to disconfirm. Don't look for combos that fit. Look for combos that don't. **Q and not-smarter**. That's where the disconfirmation action is going to be.

But what did I actually do in the previous video? I sought confirmation. I imagined happy scenarios in which Q works, not thinking about ways in which it might fail. In short, I spent a whole video falling victim to confirmation bias.

Oh, no! Apparently, I'm bad at logic and weak-minded. In the next video I'll fix it. In order to ensure that Q succeeds, I'm going to think about how, maybe, it fails.