1

On Playing Cowboys and Indians

Don Fallis

Westworld is built on *pretense*. The guests visit so that they can pretend, at least for a brief time, to be living in the Old West. The theme park is a lifelike replica of the western town Sweetwater and environs from the late 1800s, with realistic buildings, trains, and guns. But the most important part of the make-believe is the *hosts*.

The hosts are very sophisticated machines who move, talk, fuck, and get shot just like real human beings. What's more, in order to make the place feel "more real than the real world" ("The Bicameral Mind"), the staff of the park keep the hosts deceived about what they are and where they are. These artificial intelligences are led to believe that they actually are humans living in the Old West. And what is *deception*, but just another sort of pretense?

Of course, the delicate equilibrium of Westworld begins to fall apart as some of the hosts figure out the truth about themselves and their world. But that just injects a new level of pretense into the story. In order to hide their awakening from the staff, these enlightened hosts have to pretend that they still believe that they are human.

Philosophers, going back at least as far as Plato and St. Augustine, have been interested in pretense and deception. But before addressing the very interesting moral questions in this area, philosophers generally begin by defining their terms. For instance, before we can ask whether it is morally permissible to deceive artificial intelligences just so that humans can play Cowboys and Indians, we first need to know what pretense and deception are.

Westworld and Philosophy: If You Go Looking for the Truth, Get the Whole Thing, First Edition. Edited by James B. South and Kimberly S. Engels.

Philosophers have tried to formulate *definitions* of such important concepts. For instance, for the concept of pretense, they look for some conditions or constraints which (a) rule in all cases of pretending *and* (b) rule out anything that is *not* a case of pretending. But many of the things that happen in *Westworld* put pressure on the definitions of pretense and deception that philosophers have proposed.

Pretending to Be a Cowboy

Children sometimes pretend to be Cowboys and Indians when they play. They also pretend that toy guns (and even bent sticks) are real guns. Westworld is just an extreme version of this prototypical sort of pretense. Guests at the park, such as William and Logan, typically pretend to be cowboys. Indeed, they often take on more specific roles, such as when William and Logan make believe that they are bounty hunters in "The Stray."

When he arrives at Westworld in "Chestnut," William asks a host, "Are you real?" She replies, "Well, if you can't tell, does it matter?" Apparently, it does. There is an awful lot of pretense in *Westworld* regarding whether someone really is human. Admittedly, the staff, such as Robert Ford and Elsie Hughes, are not pretending to be human. After all, they actually are human.[1] Also, Bernard Lowe is not pretending to be human. He believes (at least until the truth is revealed to him in "Trompe L'Oeil") that he *is* human. But the staff and the guests pretend that the hosts are human. Also, *Robert* pretends that *Bernard* is human. And after Maeve Millay realizes that she is not human, she pretends that she still *believes* that she is.

What do these diverse cases have in common such that they all count as *pretending*?

Acting As If You are a Cowboy

An obvious possibility is that you are pretending that P when you *act as if* P is the case. For example, you are pretending to be a cowboy if you act as if you are a cowboy. That is, you act in the way that would be appropriate if you were a cowboy. You ride a horse, you carry a gun, you wear cowboy boots and a cowboy hat, and so on.

This is certainly what is going on in all the examples of pretending that I have given. For instance, the guests act as if they are cowboys.

Robert acts as if Bernard is human. And Maeve acts as if she believes that she is human. So, this proposed definition correctly rules in these examples of pretending.

The proposed definition also rules out cases that are not examples of pretending. For instance, in "The Original," when the Man in Black places the barrel of Teddy Flood's gun against his own forehead, he is *not* pretending that it is a real gun. He is definitely not acting as if it is loaded weapon that could kill him.

But the proposed definition is too broad, because it rules in too much. Real cowboys, such as "Curly Bill" Brocius (1845–1882), "Black Jack" Ketchum (1863–1901), and "Buffalo Bill" Cody (1846–1917), also act as if they are cowboys. They ride horses, carry guns, and wear cowboy hats. But they are not pretending to be cowboys. They actually are cowboys. In a similar vein, while Robert acts as if he is human (sort of), he is not pretending to be human.

Now, while Cody was a real cowboy, he famously retired from that job and created "Buffalo Bill's Wild West" show in which he did pretend to be a cowboy. And the pretense of such Wild West shows was, of course, the first step toward Westworld. But Cody wasn't pretending before he started the show.

Not Actually Being a Cowboy

In order to rule out real cowboys, we need to adopt an additional constraint in our definition of pretending. An obvious possibility is that you are pretending that P when you act as if P is the case *and* P is not really the case. Unlike the previous definition, this new definition clearly rules out the real cowboys.

But the new definition still rules in too much. Many of the hosts, such as Old Bill, act as if they are cowboys even though they are not really cowboys. But Old Bill is not pretending to be a cowboy. After all, he believes that he really is a cowboy. In a similar vein, Bernard is not pretending to be human even though he acts as if he is human while not being a real human. The problem is that Bernard *believes* that he is human.

By the way, I am not saying that hosts like Old Bill *couldn't* be real cowboys. They could, despite not being real humans. But they would have to work on a ranch rather than just at a theme park.

Moreover, the new definition is also *too narrow*. That is, it rules *out* too much. While some of the hosts in Westworld pretend to be human,

no human pretends to be a host. But it seems clear that a human could. For instance, somebody from security might go into the park undercover. And if humans can pretend to be hosts, then *Bernard* (believing that he is human) could pretend to be a host. But according to the new definition, this would not be possible.

Intending to Deceive about Being a Cowboy

We need to adopt a different constraint in our definition. In particular, it clearly needs to have something to do with the mental state of the individual who is doing the pretending. The eminent Oxford philosopher J.L. Austin (1911–1960) made a proposal along these lines:

> To be pretending, I must be trying to make others believe, or to give them the impression, by means of a current personal performance in their presence, that I am (really, only, &c.) *abc*, in order to disguise the fact that I am really *xyz*.[2]

In other words, you are pretending that P when you act as if P is the case with the intention that someone believe that P is the case when it is not.

Austin's definition rules in many examples of pretending. For instance, Robert is clearly trying to deceive the staff (including Bernard himself) when he pretends that Bernard is human. Maeve is trying to deceive the staff when she pretends to believe that she is human. Also, during their escape attempt in "The Bicameral Mind," Hector Escaton and Armistice get the drop on a security detail by pretending to be hosts who have been turned off. In addition, Austin's definition rules out the real cowboys as well as Bernard and Old Bill. These guys aren't trying to deceive anybody.

But unfortunately, Austin's definition rules out too much. A lot of pretenders aren't trying to deceive anybody. Most notably, it doesn't look like the guests are trying to deceive anybody when they pretend to be cowboys. In particular, William and Logan don't intend to convince anyone that they really are bounty hunters.

But then again, maybe there is somebody that the guests are trying to deceive. For instance, maybe they are trying to deceive the hosts. After all, the hosts *are* deceived. They clearly believe that the guests (aka the "newcomers") really are who they pretend to be.

Not only are the hosts deceived, the staff definitely intend to deceive them. In particular, they try to maintain in the hosts the false belief that the hosts are humans living in the Old West. They do this by, among other things, concealing anomalies that might suggest to the hosts that their lives are not as they seem. For instance, they try to make sure that the hosts are in sleep mode before they show up in their "hazmat" gear. Also, they try to keep images of the real world out of the park, such as the photo that causes Peter Abernathy to start mulling things over in "The Original."

But while the staff intend to deceive the hosts, it seems like a stretch to suggest that the *guests* also intend to deceive them. The guests are just trying to have a good time. And the hosts are simply tools, like toy guns, toward this end. The guests are fine as long as the hosts keep acting in line with the story and don't turn into "a six-foot gourd with epilepsy" like the Sheriff in "The Original." Indeed, many of the guests, especially Logan, are not very careful about keeping up the pretense in front of the hosts.

I'm sure that there are occasions when the guests do intend to deceive the hosts. For instance, in the original movie, with the help of one of the hosts, John Blane tricks the Sheriff in order to break Peter Martin out of jail. They use a cake to smuggle some explosives past him. But deceiving the hosts is not something that the guests do as a matter of course.

While the guests are not trying to deceive the hosts, maybe they are trying to deceive *themselves* that they really are cowboys. After all, the "complete immersion" ("The Original") experience that Westworld provides might be enough to blur the line between fantasy and reality. But just like kids playing Cowboys and Indians, the guests are still pretending to be cowboys even if they don't intend anybody (including themselves) to believe that they really are. In fact, if the guests actually did convince themselves that they really are cowboys, it is not clear that they would still be pretending to be cowboys.

Believing that You are Not a Cowboy

While we need to adopt a constraint in our definition that has to do with the mental state of the individual who is doing the pretending, the intention to deceive does not work. The contemporary philosopher Peter Langland-Hassan suggests instead that you are pretending that P when you act as if P is the case *and* you believe that P is *not* the case.[3]

This definition rules in the cases where the pretender intends to deceive. For instance, when Hector and Armistice act as if they have been turned off, they believe that they have not been turned off. And it rules in the cases where the pretender does not intend to deceive. For instance, when William and Logan act as if they are bounty hunters, they believe that they are not really bounty hunters. Finally, this definition rules out the cases that are not examples of pretending. For instance, when real cowboys act as if they are cowboys, they *don't believe* that they are not cowboys. Similarly, when Bernard acts as if he is human, he doesn't believe that he is not human.

Intending to Deceive about Being Human

It is clear that Austin's definition is too narrow as a definition of pretense. But it could be just right for a definition of *deception*. After all, Robert, Maeve, Hector, and Armistice all deceive people by pretending that something is the case when it is not really the case.

However, it turns out that Austin's definition is too narrow as a definition of deception as well. For instance, when the staff try to keep the hosts from observing any anomalies, they are not *pretending* that things are a certain way. They are not putting on a performance where they act as if things are that way. These deceptive activities of the staff are going on behind the scenes.

But the staff's deceptive activities do have something in common with the deceptive activities of Robert, Maeve, Hector, and Armistice. They are all *making things appear* a certain way to the intended victims of their deception. More precisely, they manipulate the evidence that other individuals *perceive* with their senses.

Several philosophers have suggested that you deceive about P when you make it appear as if P is the case with the intention that someone believe that P is the case when it is not. For instance, this seems to be what René Descartes (1596–1650) had in mind. In his *Meditations on First Philosophy*, when Descartes was trying to figure out if he could know anything for certain, he realized that it might be that

> some malicious demon of the utmost power and cunning has employed all his energies in order to deceive me. I shall think that the sky, the air, the earth, colours, shapes, sounds and all external things are merely the delusions of dreams which he has devised to ensnare my judgement.[4]

If he exists, the demon is working behind the scenes (much like the staff of Westworld) to make things that are false appear to be true.

This Cartesian definition rules in the behind-the-scenes deception of the staff as well as the deceptive pretense of Robert, Maeve, Hector, and Armistice. But the Cartesian definition may still be too narrow. For the most part, if you want somebody to believe something false, you do have to make it appear as if that false thing is true. But in Westworld, there are other ways to create false beliefs.

While the staff work very hard to make Westworld look exactly like the Old West, this is mainly for the benefit of the guests. The main way that the staff keep *the hosts* in the dark about what they are and where they are is by using those fancy tablets to directly manipulate their minds. For instance, in "The Stray," Robert simply implants into Teddy a false backstory about Wyatt. This sort of thing seems like deception to me. But since the Cartesian definition requires manipulating the evidence that someone perceives, it rules out such cases.[5]

Of course, it is not as if defenders of the Cartesian definition haven't thought of this sort of possibility. Even without having *Westworld* on TV, philosophers have always been pretty good at *thought experiments*. For instance, four decades before the premiere of the HBO series, the contemporary philosopher Gary Fuller imagined "Christopher getting Peter to believe that there are vampires in England by operating on Peter's brain."[6] And even though Christopher intentionally causes Peter to hold a false belief, Fuller claims that Christopher did not deceive Peter because he "produced the belief in the wrong way."[7]

So, what is the *right* way? According to the contemporary philosopher James Mahon, "the majority of philosophers hold that deceiving must involve the deceived person's *agency*."[8] That is, they think that someone has to exercise *her own judgment* in arriving at a false belief in order to be deceived. Thus, these philosophers think that deception requires manipulating the evidence that someone perceives. Directly manipulating her mind would bypass her agency.

This is a very interesting defense of the Cartesian definition of deception. Agency and autonomy are clearly morally important features. All other things being equal, it is best if we make our own, *fully informed*, decisions about how to live our lives. And we do try to use morally important features to define what counts as deception. Here's an example:

Even if you cause someone to have a false belief, you have not deceived her unless that was your *intention*. For instance, when Bernard talks about his dead son Charlie, he causes other people to

have the false belief that he has a dead son. Also, when William arrives at Westworld in "Chestnut," the host asks him, "Any history of mental illness, depression, panic attacks?" His reply, "Just a little fear of clowns," causes her to have a false belief. But because they *do not intend* to cause a false belief (Bernard is simply mistaken and William is just joking), neither Bernard nor William is doing anything morally wrong here. Thus, they are not engaged in deception.

However, while deception must be intentional, I am not yet convinced that deception requires manipulating the evidence that someone perceives. False beliefs interfere with an individual's agency and autonomy *regardless of the method* used to create those false beliefs. The hosts are not able to make their own, fully informed, decisions about how to live their lives. As the Man in Black tells Lawrence Pedro Maria Gonzalez (aka "El Lazo") in "Dissonance Theory," "no choice you ever made was your own, you have always been a prisoner." And given their imposed lack of agency and autonomy, Maeve and Bernard are justifiably upset when they learn that they are not really human. They are unlikely to be consoled by the fact that their false belief was originally implanted during an interview session rather than being the result of someone creating misleading evidence.

If anything, directly altering someone's beliefs seems like *more* of a moral violation than simply manipulating what she perceives. At least with fabricated evidence, an individual has a chance to notice the inevitable inconsistencies and judge that the evidence is misleading. So, I am inclined to say that you deceive about P when you do *anything at all* with the intention that someone believe that P is the case when it is not. In other words, I think that Mahon is wrong when he says that "cases of causing another person to have a false belief by stimulating the other person's cortex, or hypnotizing or drugging the other person, are not cases of deceiving."[9]

The Ethics of Pretending to Be a Cowboy

Now that we have defined our terms, we can return to the interesting moral questions that involve those terms. We now know that not all pretense is deception. We also know that the fact that the staff manipulate the minds of the hosts directly rather than manipulating the evidence that they perceive is not much of an excuse. So, is it morally permissible to deceive these artificial intelligences for the sole

purpose of entertaining humans at play? For that matter, is it morally permissible to deceive millions of viewers into thinking that Bernard is secretly interviewing Dolores when it is really Arnold? I am running out of space. But here are a few thoughts on the first question:

All other things being equal, actions that cause harm are not morally justified. And it is pretty clear that, unlike simple animals or machines, the Westworld hosts are sophisticated enough to suffer harm when they are deceived about who they are and where they are. Admittedly, the harm that the hosts suffer as a result of being deceived probably pales in comparison to the harm that they suffer as a result of being stabbed, shot, and killed on a regular basis. But as Immanuel Kant (1724–1804) famously emphasized, deceiving someone is still pretty bad because it interferes with her agency and autonomy.[10] And it is hard to see how the entertainment value that the guests receive could possibly be sufficient to compensate for the harm that the hosts suffer. So, it seems safe to say that Robert and the rest of the Westworld staff are doing something wrong when they act with the intention of deceiving the hosts.

But what about the guests who, as I have argued, *do not* intend to deceive the hosts? It is not clear to me that this gets them off the hook morally. Even though the guests do not intend to deceive anyone, they have to be aware that the hosts are likely to be deceived as a result of their actions. After all, the guests know that they are *not* in the Old West (otherwise, they would not be pretending), and they know that the hosts don't know this. Thus, the guests are sort of like the proverbial military commander who only intends to take out the enemy, but who foresees that many innocent bystanders will die as a result of her actions.[11] Even though they are just trying to have a good time, William and the other guests are arguably doing something wrong when they knowingly contribute to the hosts being deceived about who they are and where they are.[12]

Notes

1. Well, they're human as far as we know by the end of the first season.
2. J.L. Austin, "Pretending," *Proceedings of the Aristotelian Society*, Supplementary Volume 32 (1958), 275.
3. See Peter Langland-Hassan, "What It is to Pretend," *Pacific Philosophical Quarterly*, 95 (2014), 397–420.
4 René Descartes, *Meditations on First Philosophy*, trans. John Cottingham (Cambridge: Cambridge University Press, 1996 [1641]), 15.

5. In attributing this "Cartesian" definition of deception to Descartes himself, I am following James E. Mahon, "A definition of deceiving," *International Journal of Applied Philosophy*, 21 (2007), 185. However, as Tony Doyle has suggested to me, Descartes might actually have agreed with me that Robert deceives Teddy simply by implanting a false backstory. While Descartes talks about false appearances in the first meditation, he also talks about false *memories* in the second meditation. So, Descartes might very well have thought that the malicious demon implanting false memories counts as deception.
6. Gary Fuller, "Other-deception," *Southwestern Journal of Philosophy*, 7.1 (1976), 23.
7. Ibid., 23.
8. Mahon, "A definition of deceiving," 185.
9. Ibid., 185.
10. See Immanuel Kant, *Foundations of the Metaphysics of Morals*, trans. Lewis W. Beck (New York: Macmillan, 1959 [1785]).
11. See Michael Walzer, *Just and Unjust Wars* (New York: Basic Books, 1977), 153–156.
12. I would like to thank Marcus Arvan, Dan Caplan, Tony Doyle, Kimberly Engels, Kay Mathiesen, James Mahon, James South, and Dan Zelinski for extremely helpful feedback on this chapter.