HIDDEN BRAIN

< When Great Minds Think Unlike: Inside Science's 'Replication Crisis'

May 24, 2016 · 12:10 AM ET

ARI SHAPIRO, BYLINE: So here's the deal. Researchers recently tried to replicate a hundred experiments in psychology that were published in...

(SOUNDBITE OF ARCHIVED RECORDING)

UNIDENTIFIED WOMAN #1: The Center for Open Science recruited colleagues from around the world to try and replicate a hundred studies...

(SOUNDBITE OF ARCHIVED RECORDING)

UNIDENTIFIED MAN #1: And found that most of them could not be reproduced with the same results. In the fact, depending on...

SHANKAR VEDANTAM, HOST:

Welcome to HIDDEN BRAIN. I'm Shankar Vedantam. Today, we're going to talk about what has been called a replication crisis in science.

(SOUNDBITE OF ARCHIVED RECORDING)

UNIDENTIFIED WOMAN #2: The replicators in this recent study failed to get the same findings from the original experiment.

VEDANTAM: From cancer medicine to psychology, researchers are finding that many

claims made in scientific studies fail to hold up when those studies are repeated by an independent group.

(SOUNDBITE OF MUSIC)

VEDANTAM: Later in this episode, we're going to explore one provocative study that looked at stereotypes about Asians, women and math tests and explain what happened when researchers tried to reproduce the finding. We're going to use this story to explore a deeper question. What do scientists really mean when they talk about the truth?

Before we get to that story, I want to give you some context. The crisis has actually been a long time coming. In 2011 for example, Dutch researchers claimed that broken sidewalks encourage racism. They published their findings in one of the most prestigious academic journals, Science Magazine.

A couple of years later, another article in Science showed that when a gay person shows up at a stranger's door and speaks openly about what it's like to be gay, this has an extraordinary fact.

(SOUNDBITE OF ARCHIVED RECORDING)

UNIDENTIFIED MAN #2: It was a personal connection between the gay person who they were trying to show, you know, there in person and...

VEDANTAM: People who are against gay marriage changed their minds after these emotional encounters.

(SOUNDBITE OF ARCHIVED RECORDING)

UNIDENTIFIED MAN #3: It was the combination of, you know, contact with a minority coupled with a discussion of issues pertinent.

VEDANTAM: The results were written up in The New York Times, The Wall Street Journal, the Washington Post. They were featured on public radio programs such as the clip you're hearing on Science Friday. Finally in 2009, researchers claimed that

bilingualism, the ability to speak more than one language, is better for your brain.

All these claims had serious problems. The Dutch claim was based on fabricated data. One author of the gay marriage claim asked for the paper to be withdrawn after concerns were raised about fraud. Both claims were retracted.

(SOUNDBITE OF ARCHIVED RECORDING)

UNIDENTIFIED MAN #4: Might be true, might not. We don't know because it turns out the researchers made up the data.

VEDANTAM: The bilingual advantage paper wasn't fabricated, but it was missing important context. The researchers had conducted four experiments. Three failed to show that bilingualism was better for the brain. Only one experiment showed a benefit. It was the only one that was published. Angela de Bruin was on the team that worked on the bilingual advantage study.

ANGELA DE BRUIN: It is troubling 'cause we like to believe that what we see is actually at its truth. But if it's only half of the results we find and we're in fact hiding the other half of the results, then we'll never really find out what's going on.

VEDANTAM: At the University of Virginia, psychologist Brian Nosek decided something had to be done. Brian felt the problem was that too many researchers and too many scientific journals were focusing on publishing new and unusual findings. Too few were spending time crosschecking earlier work to make sure it was a solid.

BRIAN NOSEK: One of the key factors of science is that a claim becomes a credible claim by being reproducible, that someone else can take the same approach, the same protocol, the same procedure, do it again themselves and obtain a similar result.

VEDANTAM: Brian launched an effort to reproduce dozens of studies in psychology. He published a report in 2015.

NOSEK: We found that we were able to reproduce the original results in less than half of the cases across five different criteria of evaluating whether a replication was

successful or not.

VEDANTAM: Over the last year, there have been many debates about what this means. Some critics say it proves that most studies are worthless. At many universities, researchers feel it's their integrity, not just their scientific conclusions, that are being called into question. At Harvard University, psychologist Dan Gilbert recently published a paper calling Brian's conclusions into question.

DAN GILBERT: So I think we just have to use our heads to figure out which kinds of things we expect to replicate directly and which kinds of things we would only expect a conceptual replication. And we need to calm down when we don't see direct replication and ask whether we really should have expected it at all.

VEDANTAM: To unpack all of this, let's take a detailed look at one study and what happened when researchers tried to replicate it. I think this story reveals many truths about the ongoing controversy.

(SOUNDBITE OF MUSIC)

VEDANTAM: When they were graduate students at Harvard, Todd Pittinsky and his friend Margaret Shih often went to restaurants together. They went for the food, but they also spent a lot of time observing human behavior.

TODD PITTINSKY: We often after class would go to The Cheesecake Factory. And she would order a strawberry shortcake. And I would typically order a salad. And the number of times that the salad was delivered to her and the strawberry shortcake was delivered to me - she also likes regular Coke, and I'm a diabetic. So I drink Diet Coke.

And without fail, the Diet Coke would go to her and the regular Coke would go to me.

VEDANTAM: The waiters were stereotyping Todd and Margaret. The guy was probably ordering the less healthy stuff. The woman was ordering salads and diet drinks. Todd and Margaret knew there was lots of research into the effects of such stereotypes. Now, getting a dish you haven't ordered is one thing. But there are more serious consequences.

Stereotypes can be hurtful. They can affect performance. But as Todd and Margaret observed the waiters, they realized something was missing in the research. The previous studies had focused on the negative consequences of stereotypes. Could stereotypes also work in a positive fashion?

PITTINSKY: We thought if we really want to understand how stereotypes operate in the world, we can't simply look at half of it.

VEDANTAM: The young researchers brainstormed how they might study the other half of the equation. The answer came to them as they were, yeah, eating together.

PITTINSKY: We were sitting in Harvard Square over ice cream. And we said what we needed is a group where the stereotypes go in very different directions.

VEDANTAM: They wanted to study a situation where stereotypes could have both positive and negative effects.

PITTINSKY: And Margaret, she happens to be an Asian-American and a woman. And we started talking about math identities. And we kept going back and forth and back and forth. And then literally, at the same moment, we said, well, why don't we study Asian women and math?

VEDANTAM: The experiment they designed was ingenious and simple. There are negative stereotypes about women doing math and positive stereotypes about Asians and math. So what happens when you give a math test to women who are Asian?

PITTINSKY: We hypothesized that when you make different identities salient, you should expect different stereotypes to be applied.

VEDANTAM: Todd and Margaret figured that if they reminded Asian women about their gender, they would see the negative stereotype at work. But what would happen if they subtly reminded the volunteers about their Asian identity? The researchers recruited Asian women as volunteers and asked some of them to identify their gender on a form before taking a math test.

Earlier research had shown that when you make gender salient in this way, this triggers the negative stereotype about women and math. Todd and Margaret reminded other volunteers, selected at random, about their Asian heritage. They wanted to make these volunteers remember the stereotype about Asians being good at math.

After all the volunteers finished the test, the researchers analyzed their performance. Todd was walking down the hall from Margaret one day when he heard her call out to him.

PITTINSKY: She just shouted, holy cow, it worked (laughter). So I just sort of ran down there, and we started looking at the output together.

VEDANTAM: The study found that when the volunteers were reminded that they were women, they did worse on the math test. When they were reminded that they were Asian, they did better. Same women, same math test. Negative stereotype, negative result, positive stereotype, positive result. The study was an instant sensation. Psychologist Brian Nosek.

NOSEK: This is one of my favorite effects in psychological science. Something that seems like it shouldn't be flexible, how well we perform in math, is flexible as a function of the identities that we have in mind and stereotypes associated with those identities - Asians being good at math, women being not as good at math.

VEDANTAM: The study quickly became a staple of college textbooks, says psychologist Carolyn Gibson.

CAROLYN GIBSON: It is a pretty amazing finding. And I heard about that study for the first time as an undergrad. It's been used as an example in social psychology courses for years since it was published in 1999. And it's been used as a good example for stereotype threat and stereotype boost.

VEDANTAM: But from a scientific perspective, there was one big problem.

GIBSON: It had never been replicated exactly. Somebody had never followed their steps that they followed and replicated their results. But it's been used to support

further studies many times over the past 15 years.

VEDANTAM: Brian Nosek agreed. Someone needed to replicate the original study. He was spearheading a mammoth effort to reproduce dozens of studies in psychology. Along with a panel of reviewers, he selected this study for replication and asked Carolyn Gibson at Georgia Southern University to conduct it. Brian wanted the replication to closely match the conditions of the original study.

If you don't do that, you really are conducting two different studies. After launching the replication, he had second thoughts about its location in the South.

NOSEK: And the reviewers thought, this looks like a case where the location might matter. Asians in the Southern U.S. might be a more distinct minority than Asians in the Northeast or in the West. And so we recruited a second team to do a replication simultaneously at UC Berkeley in a West Coast University where Asians are much more prominent members of the community.

VEDANTAM: The team in Berkeley was headed by Alice Moon. Alice was a fan of the original paper.

ALICE MOON: When I heard about it, I just thought it was, like, one of the very cool demonstrations in social psychology. And so that's why I always liked this paper.

VEDANTAM: She followed the protocol of the original Harvard experiment. She recruited Asian women, reminded them of the female side of their identity or the Asian side of their identity and then gave them a math test. So what happened?

MOON: When we compared, just as the original paper did - when we compared the participants who were in the Asian identity salient condition with the participants in the female identity salient condition, we found that there was no difference in their math performance.

VEDANTAM: The celebrated study failed to be replicated. When Brian Nosek announced the finding about this and dozens of other studies that could not be replicated, it caused an uproar.

(SOUNDBITE OF ARCHIVED RECORDING)

UNIDENTIFIED MAN #4: In social psychology did not hold up...

VEDANTAM: Newspaper articles called it a crisis. Critics held accusations about fraud and scientific misconduct.

(SOUNDBITE OF ARCHIVED RECORDING)

UNIDENTIFIED WOMAN #3: Seems to be running into some problems.

VEDANTAM: In a 6,000-word cover story, the conservative magazine Weekly Standard said that liberals had been making up research into how stereotypes affect women and people of color.

(SOUNDBITE OF MUSIC)

VEDANTAM: The Berkeley study, however, was not the only replication of Todd Pittinsky and Margaret Shih's paper. Remember how Brian had two groups conduct replication? I asked Carolyn Gibson at Georgia Southern University what she found when she ran the experiment on Asian women in math.

GIBSON: When primed with Asian identity, Asian females did better on a math test compared to those who had been primed with their female identity. And then those primed with their female identity did significantly worse.

VEDANTAM: Carolyn has no doubt about the meaning of what she found.

GIBSON: I believe that it further supports the original finding and that it gives even more robust evidence to this idea mostly because we followed the same method as the original study and because we collected more participants. And so we have a more powerful study.

VEDANTAM: At Berkeley, Alice isn't sure.

MOON: I do believe that stereotypes, in general, do have effects on our lives. But in

terms of this particular finding about whether stereotypes can facilitate people's academic performance, I guess it has made me question whether or not that finding is true.

VEDANTAM: OK, so which is it? Should we trust the results of the Berkeley study and say that Todd Pittinsky and Margaret Shih's finding was disproved? Should we trust the Georgia study and say the finding was confirmed? What happens when scientific studies disagree with one another?

(SOUNDBITE OF MUSIC)

VEDANTAM: The popular narrative of the replication crisis suggests that scientists are like dueling gladiators. If two scientists come up with different findings, it must mean one of them is wrong. Or worse, one of them must have faked her data. When we come back, we look at why this idea misunderstands how science actually works.

PITTINSKY: Our statistical techniques are probabilistic and not definitive. And so we absolutely need replications. But replications in our current academic climate are also serving the purpose of trying to vet out academic fraud and are serving as a detection technique. And those two are very different missions for replications.

VEDANTAM: Stay with us.

(SOUNDBITE OF MUSIC)

VEDANTAM: This is HIDDEN BRAIN. I'm Shankar Vedantam. We're taking a look today at how science works and the so-called replication crisis in the social sciences. As I listened to the news reports, I found myself drawing an analogy with my own profession, journalism. Here's what I mean. A Few years ago, a reporter for The New York Times was caught fabricating stories.

Instead of traveling to various locations and interviewing people, he simply made stuff up. The newspaper went back and re-reported the stories Jayson Blair had written. When the facts didn't match, the reporter was fired. Imagine for a second what would happen if we re-reported every story by every reporter at The New York Times. Even

when reporters are doing a perfectly good job, the older news stories might not match.

A source might not say exactly the same thing again. Sometimes if the circumstances have changed, a source might say something completely different. So when two reporters don't produce the same story, it could be that one of them is making stuff up. But much more likely is that both of them are right.

Now, I know what you're thinking. Journalism is storytelling. Science is about data. But let's look closely at what happened in the replications that Carolyn Gibson and Alice Moon did of Todd Pittinsky's study. In the original study, women administered the experiment. In Georgia, the facilitator was also female. But in Berkeley, where the replication failed, both male and female facilitators administered the study.

Could that have made a difference?

ERIC BRADLOW: Let's be clear. So it was not an exact replication. So here's an example. It mentions clearly in the paper that - and I don't know whether this factor's important or not - that in one study, the experimenter gender were males. And in another study, the experimenter gender were females. I have no idea whether that's a factor that could explain the difference between the two studies.

And so let's be clear about what - it's not an exact replication.

VEDANTAM: This is Eric Bradlow from the University of Pennsylvania. He's eminently qualified to talk about this stuff.

BRADLOW: Spent four years here at Wharton studying statistics and mathematics, went on to get my Ph.D. in statistics. And for the last 20 years, I've been applying statistical methods to lots of problems. But I consider myself a mathematical social scientist.

VEDANTAM: Eric believes that requiring studies to achieve statistical application, to match more or less perfectly, before you conclude that either is true, is like requiring two reporters to cover a basketball game and come back with nearly identical stories.

BRADLOW: Exact replication is one of those mythological ivory tower things that doesn't exist. What we really need to think about is if the study doesn't replicate, why doesn't it replicate? And even if it doesn't replicate exactly, it may actually reinforce the original finding. In other words, you may be more certain...

VEDANTAM: This isn't just true about studies in psychology. Eric told me that NIH researchers once found that lab mice, given a sedative, took 35 minutes to recover. When the experiment was repeated, the mice took 16 minutes to recover. The scientists scratched their heads. It made no sense. It took a while to figure out that something that shouldn't have made a difference did.

In between the two experiments, wood shavings in the animal cages were changed. Turns out that red cedar and pine shavings step up the speed at which the sedative was metabolized, birch or maple don't. This is not to say that repeating experiments is useless or pointless. It's incredibly valuable.

But replications primarily help us understand the nuances around a phenomenon. They're not very useful as a tool to detect fraud.

BRADLOW: Just 'cause you get different results doesn't mean you shouldn't trust them. Are they within a margin of error of each other? Are there other variables that would make it so that study done at University A and the study done at University B wouldn't yield exactly the same thing?

I think that's a better way of looking at it than, say, if you don't get exactly the same results or even results that are very nearly the same, you can't trust them. I think that's a superficial level of science. I think you need to go below that.

VEDANTAM: So when you yourself look at a study that has not replicated or you look at what's sometimes called a failed application, do you not at the back of your mind say, well, this disproves the first study? Do you actually never think that way?

BRADLOW: Never's a long word.

VEDANTAM: (Laughter).

BRADLOW: Never's a strong word. You know, I'm thinking, I have to think of that James Bond movie when Sean Connery said, I will never do James Bond again. And then 15 years later, he came out with a movie called "Never Say Never Again." No, I would never say that. But I would say the following. Let's imagine that you do a study and that you find that, you know, people that take an SAT prep course do 15 percent better on the SAT.

And let's imagine, then, someone else does a study, and the answer's only 3 percent. Now, there's two possibilities. One is the first study, for whatever reason, overestimated the effect. That's entirely possible. And therefore, 3 present is less than 15 percent. But note, if you combine those two studies together, your finding might actually be stronger in the sense I'm now more sure that SAT prep helps performance on the SAT.

Now, the effect size may shrink from 15 percent to 11 percent. But also notice I've possibly now doubled or tripled my sample size, so my uncertainty goes down. And now I may even be more sure that SAT prep helps, maybe not to the degree that it helped in the first study. But still, I'm more sure that it's actually effective.

VEDANTAM: When you think about different branches of science, though, aren't there branches of science where you can expect the same thing to happen very predictably over and over again? When you look at particle physics, for example, you would expect that if you fire, you know, 20,000 protons out of a gun, that they're basically going to do the same thing pretty much every time.

BRADLOW: Well, it's been - you're testing the boundary of my memory...

VEDANTAM: (Laughter).

BRADLOW: ...Of my particle physics class when I took it here at Penn. But my understanding is, of course, and this is what statisticians study, right? We study the concept of randomness. And so every science, every discipline, unless you're talking about an equal sign, like E equals MC squared. E doesn't approximately equal MC squared. It actually equals. Most physical laws and things aren't equals signs.

There's approximate signs, and so that means there's randomness to it. I think if you fired 20,000 protons, you would see that there's a deviation in the way they collide with other particles, and there's randomness. I think the same thing is true in the social sciences. You bring in 500 subjects. You bring them in at University A.

You bring them in at University B. There's randomness in people's answers. Of Course, you would hope the overall patterns would be similar. But the fact - this belief that you're going to get exactly the same findings, I'm not sure that's something science should be striving for.

VEDANTAM: Any individual study is just that, an individual study. It isn't the truth.

BRADLOW: Every observation is a point. It's a dot. And we observe dots. And then we observe more dots. And if those dots replicate, great, then we have more belief. Science is about the evolution of knowledge, might?

VEDANTAM: But the process is never ending. There will always be more things to uncover, more nuances.

BRADLOW: We get more certain about what it is we know. And we also get more certain about what are called boundary conditions or moderators like, for example, maybe this effect holds in urban areas versus not. Maybe it holds in California and not in Alabama. Maybe it holds for people that are - hold these stereotypes, and maybe it doesn't hold for people that don't.

That's - to me, that's an advance of science. We have found what's called a main effect, which is, you know, stereotypes have an effect on outcomes or priming has an effect on outcomes. And then we say, oh, and by the way, it doesn't hold in these conditions. That's not a failure to replicate. That's a more nuanced view of the original finding.

VEDANTAM: At Harvard, Dan Gilbert says you can expect some studies to replicate nearly perfectly every time. But in other cases, the very thing you're studying is changing. So exact applications aren't possible.

GILBERT: There are many findings in psychological science that we would expect to

replicate quite exactly years later and on different populations. Eyeblink conditioning is a very nice example. If I blow in your eye enough, you're going to start blinking as I purse my lips. And that's not going to be very different across cultures, across times, across age groups.

Other kinds of findings certainly are. There's one of my favorite experiments in social psychology shows that when young men who are from the North or the South of the United States are insulted, they react very differently because Northerners and Southerners have very different codes of honor. Now, you can't take that experiment and expect to do it in Italy or expect to do it 25 years from now.

It's an experiment that's of its moment and of its time.

VEDANTAM: Every researcher I spoke with told me there's lots of agreement within the scientific community. There are certainly many scientific studies that are poorly designed. There are researchers who do shoddy work. There is great pressure at universities and scientific journals to publish striking findings. But the solution to all these problems, say Eric Bradlow, Brian Nosek and Dan Gilbert, is more and better science.

Eric Bradlow.

BRADLOW: The truth will come out. More dots will come out. And if it turns out that what I published - it's not because I did anything fraudulent - just isn't true because of sample size, the way I collected the data, then, you know what? Science will eventually figure out that what I'm saying is not true. So if you'd like...

VEDANTAM: Brian Nosek is bemused that his findings about replicability have been taken to mean that the studies that failed to reproduce are worthless. He started a new system where researchers register protocols for their studies and commit to sticking to them. Scientific journals commit to publishing the findings of these studies regardless of whether the results are sexy.

NOSEK: Science is the slow march of accumulating evidence. And it's very easy to want a simple answer. Is it true? Is it false? But really, replication is just an

opportunity to accumulate more evidence, to get a more precise estimate of that particular effect.

VEDANTAM: To most people, the debate over scientific truth is an abstract issue. Most of us turn to scientist for answers. Should I drink a glass of red wine in the evening? Is this drug safe to give to my ailing mother? Should I give my kid a dollar every time she does something well at school? In reality, science is more in the question business than the answer business.

There's a reason nearly every scientific paper ends with a call for more research. Especially when it comes to human behavior, nearly every conclusion you can draw about human beings has tons of exceptions. Are people selfish? Yeah, except millions act altruistically every day. Are humans kind? Yes, except that few species are capable of greater cruelty.

If you want answers that never change, definitive conclusions and final truths, odds are you don't want to ask a scientist.

(SOUNDBITE OF MUSIC)

VEDANTAM: The HIDDEN BRAIN podcast is produced by Kara McGuirk-Alison, Maggie Penman and Max Nesterak. You can follow us on Facebook, Twitter and Instagram and follow my stories on your local public radio station. If you liked this episode, consider giving us a review on iTunes. It will help other people find the podcast. I'm Shankar Vedantam, and this is NPR.

More Stories From NPR

SCIENCE

Climate Change: The Forgotten Issue Of This Year's Election

SCIENCE

How A Theory Of Crime And Policing Was Born, And Went Terribly Wrong

0
0

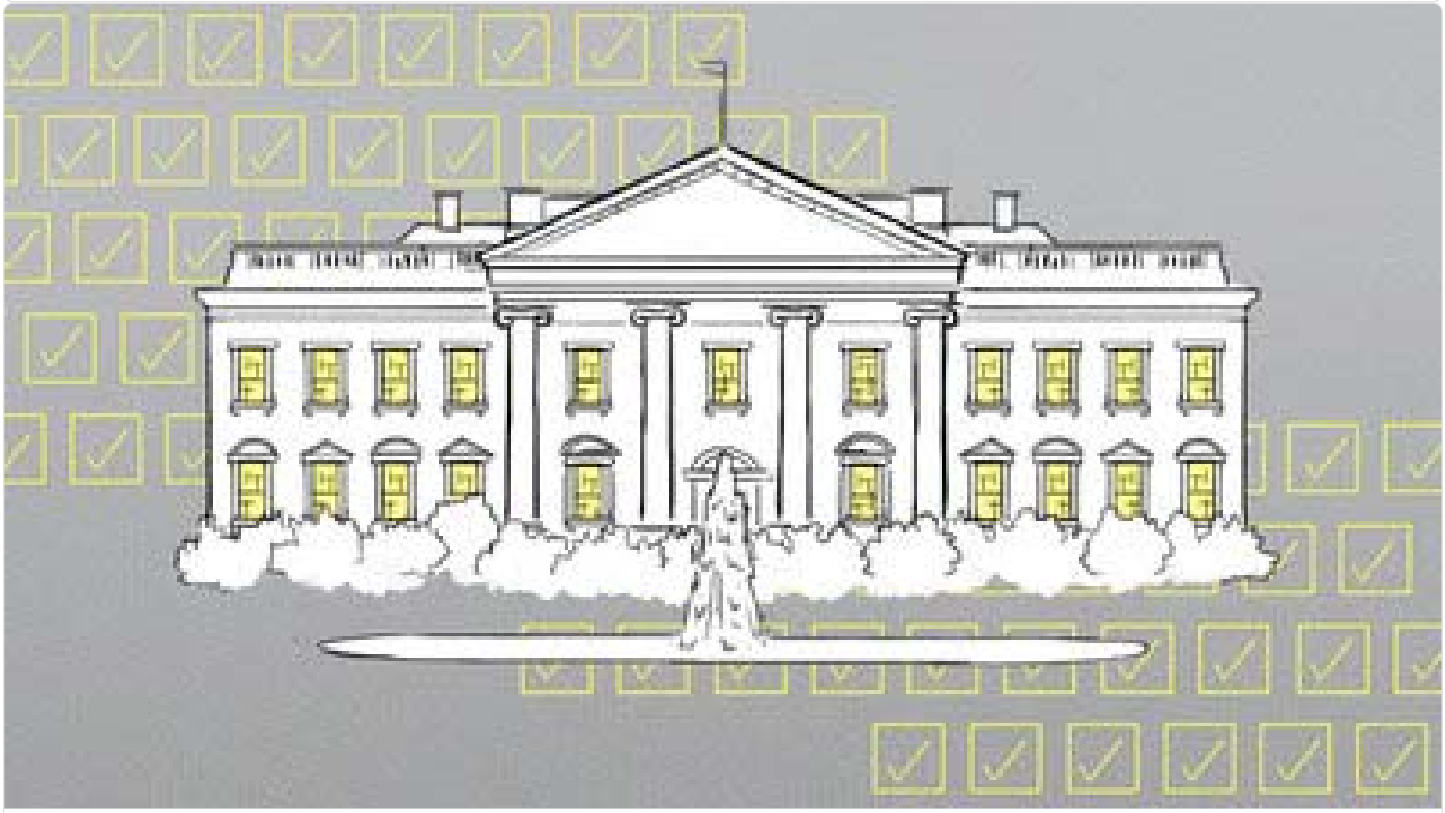
Popular on NPR.org

POLITICS

Here Is What Donald Trump Wants To Do In His First 100 Days

POLITICS

Shades Of 2000? Clinton Surpasses Trump In Popular Vote Tally



POLITICS

LIVE BLOG: Election Night 2016

POLITICS

WATCH: President Obama On Trump Win, Clinton Loss

NPR Editors' Picks

POLITICS

Republicans' Senate Tactics Leave Trump Wide Sway Over Nation's Courts

SCIENCE

In Ancient Trash Heaps, A Whale Hunting Puzzle Emerges

RACE

The Outlook On Race After Trump Victory: Fear, Resignation And Deja Vu

BOOK REVIEWS

'Things From The Flood' Is Gorgeously Creepy And Strangely Human

HIDDEN BRAIN