Priming

1. Never Leave Your Room
2. Bogus Pipeline, Bona Fide Pipeline
3. The Implicit Association Test
4. Fight Biases, or Route Around Them?

Never Leave Your Room

Related to: Priming and Contamination

Psychologists define "priming" as the ability of a stimulus to activate the brain in such a way as to affect responses to later stimuli. If that doesn't sound sufficiently ominous, feel free to re-word it as "any random thing that happens to you can hijack your judgment and personality for the next few minutes."

For example, let's say you walk into a room and notice a briefcase in the corner. Your brain is now the proud owner of the activated concept "briefcase". It is "primed" to think about briefcases, and by extension about offices, business, competition, and ambition. For the next few minutes, you will shift ever so slightly towards perceiving all social interactions as competitive, and towards behaving competitively yourself. These slight shifts will be large enough to be measured by, for example, how much money you offer during the Ultimatum Game. If that sounds too much like some sort of weird New Age sympathetic magic to believe, all I can say is Kay, Wheeler, Bargh, and Ross, 2004.[1]

We've been discussing the costs and benefits of Santa Claus recently. Well, here's one benefit: show Dutch children an image of St. Nicholas' hat, and they'll be more likely to share candy with others. Why? The researchers hypothesize that the hat activates the concept of St. Nicholas, and St. Nicholas activates an idealized concept of sharing and giving. The child is now primed to view sharing positively. Of course, the same effect can be used for evil. In the same study, kids shown the Toys 'R' Us logo refused to share their precious candy with anyone.

But this effect is limited to a few psych laboratories, right? It hasn't done anything like, you know, determine the outcome of a bunch of major elections?

I am aware of two good studies on the effect of priming in politics. In the first, subjects were subliminally[2] primed with either alphanumeric combinations that recalled the 9/11 WTC attacks (ie "911" or "WTC"), or random alphanumeric combinations. Then they were asked to rate the Bush administration's policies. Those who saw the random strings rated Bush at an unenthusiastic 42% (2.1/5). Those who were primed to be thinking about the War on Terror gave him an astounding 75% (3.75/5). This dramatic a change, even though none of them could consciously recall seeing terrorism-related stimuli.

In the second study, scientists analyzed data from the 2000 election in Arizona, and found that polling location had a moderate effect on voting results. That is, people who voted in a school were more likely to support education-friendly policies, people who voted in a church were more likely to support socially conservative policies, et cetera. The effect seems to have shifted results by about three percentage points. Think about all the elections that were won or lost by less than three percent...

Objection: correlation is not causation! Religious people probably live closer to churches, and are more likely to know where their local church is, and so on. So the scientists performed an impressive battery of regression analyses and adjustments on their data. Same response.

Objection: maybe their adjustments weren't good enough! The same scientists then called voters into their laboratory, showed them pictures of buildings, and asked them to cast a mock vote on the education initiatives. Voters who saw pictures of schools were more likely to vote yes on the pro-education initiatives than voters who saw control buildings.

What techniques do these studies suggest for rationalists? I'm tempted to say the optimal technique is to never leave your room, but there are still a few less extreme things you can do. First, avoid exposure to any salient stimuli in the few minutes before making an important decision. Everyone knows about the 9-11 terrorist attacks, but the War on Terror only hijacked the decision-making process when the subjects were exposed to the related stimuli directly before performing the rating task[3].

Second, try to make decisions in a neutral environment and then stick to them. The easiest way to avoid having your vote hijacked by the location of your polling place is to decide how to vote while you're at home, and then stick to that decision unless you have some amazing revelation on your way to the voting booth. Instead of never leaving your room, you can make decisions in your room and then carry them out later in the stimulus-laden world.

I can't help but think of the long tradition of master rationalists "blanking their mind" to make an important decision. Jeffreyssai's brain "carefully put in idle" as he descends to a bare white room to stage his crisis of faith. Anasûrimbor Kellhus withdrawing into himself and entering a probability trance before he finds the Shortest Path. Your grandmother telling you to "sleep on it" before you make an important life choice.

Whether or not you try anything as formal as that, waiting a few minutes in a stimulus-free environment before a big decision might be a good idea.

Footnotes

1: I bet that sympathetic magic probably does have strong placebo-type effects for exactly these reasons, though.

2: Priming is one of the phenomena behind all the hype about subliminal advertising and other subliminal effects. The bad news is that it's real: a picture of popcorn flashed subliminally on a movie screen *can* make you think of popcorn. The good news is that it's not particularly dangerous: your thoughts of popcorn aren't any stronger or any different than they'd be if you just saw a normal picture of popcorn.

3: The obvious objection is that if you're evaluating George Bush, it would be very strange if you didn't think of the 9-11 terror attacks yourself in the course of the evaluation. I haven't seen any research addressing this possibility, but maybe hearing an external reference to it outside the context of your own thought processes is a stronger activation than the one you would get by coming up with the idea yourself.

Bogus Pipeline, Bona Fide Pipeline

Related to: Never Leave Your Room

Perhaps you are a psychologist, and you wish to do a study on racism. Maybe you want to know whether racists drink more coffee than non-racists. Sounds easy. Find a group of people and ask them how racist they are, then ask them how much coffee they drink.

Problem: everyone in your study says they're completely non-racist and some of their best friends are black and all races are equally part of this vast multicolored tapestry we call humanity. Maybe some of them are stretching the truth here a bit. Until you figure out which ones, you're never going to find out anything interesting about coffee.

So you build a foreboding looking machine out of gleaming steel, covered with wires and blinking lights. You sit your subjects down in front of the machine, connect them to its electrodes, and say as convincingly as possible that it is a lie detector and they must speak the truth. Your subjects look doubtful. Didn't they hear on TV that lie detectors don't really work? They'll stick to their vehement assertions of tolerance until you get a more impressive-looking machine, thank you.

You get smarter. Before your experiment, you make the subjects fill in a survey, which you secretly copy while they're not looking. Then you bring them in front of the gleaming metal lie detector, and dare them to try to thwart it. Every time they give an answer different from the one on the survey, you frown and tell them that the machine has detected their fabrication. When the subject is suitably impressed, you start asking them about racism.

The subjects start grudgingly admitting they have some racist attitudes. You have invented the Bogus Pipeline.

The Bogus Pipeline is quite powerful. Since its invention in the 70s, several different studies demonstrate that its victims will give significantly less self-enhancing answers to a wide variety of questions than will subjects not connected to the machinery. In cases where facts can be checked, Pipeline subjects' answers tend to be more factually correct than normal subjects'.

In one of the more interesting Bogus Pipeline experiments, Millham and Kellogg wanted to know how much of a person's average self-enhancement is due to self-deception biases, and how much is due to simple lying. They asked people some questions about themselves under normal and Pipeline conditions, using the Marlowe-Crowne scale. This scale really deserves a post of its own, but the short version is that it asks you some loaded questions, and if you take them as an opportunity to say nice things about yourself, you get marked down as a self-enhancer. There was a correlation of .68 between Marlowe-Crowne scores in normal and Pipeline conditions. If we accept that no one deliberately lies under the Pipeline, that means we now know how much self-enhancement is, on average, self-deception rather than deliberate falsehood (tendency towards deliberate falsehoods correlated .37 with Marlowe-Crowne.[1])

Interesting stuff. But you still don't know whether racists drink more coffee! Your Bogus Pipeline only eliminates part of the self-enhancement in your subjects' answers. If you want to solve the coffee question once and for all, you can't count on a fake mind-reading device. You need a real mind-reading device. And in the mid 90s, psychology finally developed one.

The Bona Fide Pipeline is far less impressive-looking than the Bogus Pipeline. Though the Bogus Pipeline tries as hard as it can to scream "mind-reading device", the Bona Fide Pipeline has a vested interest in preventing its victims from realizing their minds are being read. It is a simple computer terminal.

The Pipeline uses a complicated process to disguise itself as an ordinary study on distraction or face recognition or somesuch, but the active ingredient is this: the subjects play a game where they must hit one key (perhaps "A") if the screen displays a good word (for example "wonderful"), and a different key (perhaps "L") if the screen displays a bad word (for example "ugly").

But before it gives you the word, it shows you a picture of a white person or a black person. Remember priming? That picture of a black person is going to prime your brain's concept of "black person" and any concepts you associate with "black person". If you have racist attitudes, "bad" is one concept you associate with "black person". You're going to have a very easy time recognizing "ugly" as a bad word, because your "bad" concept is already activated. But you're going to have a harder time recognizing "wonderful" as a good concept, because your brain is already skewed in the opposite direction. It's not impossible, it's just going to take a few hundred more milliseconds. Each of which the Bona Fide Pipeline is recording and processing. At the end, it spits out a score telling you that you took an average of three hundred milliseconds longer to recognize good words when primed with black people's pictures than white people's pictures.

Does this actually work? The original study (Fazio et al, 1995) tested both whites and blacks, and found the whites were more likely to be prejudiced against blacks than the blacks were, which makes sense. In the same study, a black experimenter conversed with the subjects for a while, and rated the quality of the interaction by a typically rigorous rubric. This fuzzy unscientific measure of racist behavior correlated well with the Pipeline's data for the individuals involved. A study by Jackson (1997) find that people who score high on prejudice by Pipeline measures on average give lower scores to an essay written by a student known to be black.

The Bona Fide Pipeline has lately been superseded by its younger, sexier, Harvard-educated cousin, the IAT. More on that, the associated controversy, and the relevance to rationality tomorrow.

Footnotes:

1: I doubt that deceptions can be separated cleanly into self-deception and deliberate falsehood like this. More likely there are many different shades of grey, and the Bogus Pipeline captures some but not all of them.

The Implicit Association Test

Continuation of: Bogus Pipeline, Bona Fide Pipeline
Related to: The Cluster Structure of Thingspace

If you've never taken the **Implicit Association Test** before, try it now.

Any will do. The one on race is the "classic", but the one on gender and careers is a bit easier to watch "in action", since the effect is so clear.

The overwhelming feeling I get when taking an Implicit Association Test is that of feeling my cognitive algorithms at work. All this time talking about thingspace and bias and categorization, and all of a sudden I have this feeling to attach the words to...

...which could be completely self-delusional. What is the evidence? Does the Implicit Association Test work?

Let the defense speak first[1]. The Implicit Association Test correctly picks up control associations. An IAT about attitudes towards insects and flowers found generally positive attitudes to the flowers and generally negative attitudes to the insects ($p = .001$), just as anyone with their head screwed on properly would expect. People's self-reports were also positively correlated with their IAT results (ie, someone who reported loving flowers and hating insects more than average also had a stronger than average IAT) although these correlations did not meet the 95% significance criterion. The study was repeated with a different subject (musical instruments vs. weapons) and similar results were obtained.

In the next study, the experimenters recruited Japanese-Americans and Korean-Americans. Japan has been threatening, invading, or oppressing Korea for large chunks of the past five hundred years, and there's no love lost between the two countries. This time, the Japanese-Americans were able to quickly match Japanese names to "good" stimuli and Korean names to "bad" stimuli, but took much longer to perform the opposite matching. The Korean-Americans had precisely the opposite problem, $p < .0001$. People's self-reports were also positively correlated with their IAT results (ie, a Korean who expressed especially negative feelings towards the Japanese on average also had a stronger than average IAT result) to a significant level.

There's been some evidence that the IAT is pretty robust. Most trivial matters like position of items don't much much of a difference. People who were asked to convincingly fake an IAT effect couldn't do it. If the same person takes the test twice, there's a correlation ofabout .6 between the two attempts[2]. There's a correlation of .55 between the Bona Fide Pipeline and the IAT (the IAT wins all competitions between the two; it produces twice as big an effect size). There's about a .24 correlation between explicit attitude and IAT score, which is significant at the 90% but not the 95% level; removing certain tests where people seem especially likely to lie on their explicit attitude takes it up to 95. When the two conflict, the IAT occasionally wins. In one study, subjects were asked to evaluate male and female applicants for a job. Their observed bias against women correlated more strongly with their scores on a gender bias IAT than with their own self-report (in other experiments in the same study, explicit self-report was a better predictor. The experimenters concluded both methods were valuable in different areas)

Now comes the prosecution. A common critique of the test is that the same individual often gets two completely different scores taking the same test twice. As far as re-test reliability goes, .6 correlation is pretty good from a theoretical point of view, but more than enough to be frequently embarrassing. It must be admitted: this test, while giving consistent results for populations, is of less use for individuals wondering how much bias they personally have.

Carl Shulman would be heartbroken if I didn't mention Philip Tetlock, so here goes. This is from Would Jesse Jackson Fail the Implicit Association Test?, by Tetlock and Arkes (2004):

> Measures of implicit prejudice are based on associations between race-related stimuli and valenced words. Reaction time (RT) data have been characterized as showing implicit prejudice when White names or faces are associated with positive concepts and African-American names or faces with negative concepts, compared to the reverse pairings. We offer three objections to the inferential leap from the comparative RT of different associations to the attribution of implicit prejudice: (a) The data may reflect shared cultural stereotypes rather than personal animus, (b) the affective negativity attributed to participants may be due to cognitions and emotions that are not necessarily prejudiced, and (c) the patterns of judgment deemed to be indicative of prejudice pass tests deemed to be diagnostic of rational behavior.

In other words, there are a bunch of legitimate reasons people might get negative IAT scores. Any connection whatsoever between black people and negative affect will do. It could be the connection that black people generally have low status in our society. It could be that a person knows of all the prejudices against black people without believing them. It could be that a person has perfectly rational negative feelings about black people because of their higher poverty rate, higher crime rate, and so on. Or it could be somethng as simple as that, for whites, black people are the out-group.

...this actually isn't much of a prosecution at all. I consider myself a moderate believer in the IAT, and I think it all sounds pretty reasonable.

What most IAT detractors I've read want to make exquisitely clear is that you can't hand someone an IAT, find an anti-black bias, and say "Aha! He's a racist! Shame on him!"[3]

I think this is pretty obvious[4]. You can hold beliefs on more than one level. A person may believe there is a dragon in his garage, yet not expect an experiment to detect it. A skeptic may disbelieve in ghosts, but be afraid of haunted houses. A stroke victim may deny an arm is hers while admitting it is attached to her body. And it's supposed to be news that you can give black people some sort of vague negative connotation on a nonconscious level without being Ku Klux Klan material?

There is a certain segment of society which interprets the sun rising in the morning as evidence of racism. It is not surprising that this segment of society also interprets the IAT as evidence for racism. I myself think racism is a bad word. Not in the way "shit" is a bad word, but in the way "wiggin" is a bad word. It divides experience in a perverse way, drawing a boundary such that Adolf Hitler ends up in the same category as the guy who feels a pang of guilty fear late at night when he sees a big muscular black guy walking towards him[5]. Taboo the word "racism", "prejudice", and any other anti-applause-light[6], and a lot of the IAT debate loses its meaning.

Which is good, because I think the IAT is about **much** more than who is or isn't racist. The IAT is a tool for measuring distances in thingspace.

Thingspace, remember, is the sort of space in which we draw categories[7]. "Chair" is a useful category because it describes a cluster of things that are close together in concept-space in a certain way: stools, rocking chairs, office chairs, desk chairs, et cetera. "Furniture" is another useful word because it describes another cluster, one that includes the chair cluster and other concepts nearby. Quok, where a "quok" is defined as either a chair or Vladimir Lenin, is a useless category, because Lenin isn't anywhere near all the other members.

Speaking of communists, remember back when East and West Germany got reunited? And remember a little further back, when North and South Vietnam got reunited too? Those reunifications, no matter how you feel about them politically, were natural links between culturally and historically similar regions. But imagine trying to unite East Germany with South Vietnam, and West Germany with North Vietnam. The resulting countries would be ungovernable and collapse in a matter of weeks.

If you associate white people with good things, and black people with bad things, then forming the categories "white and good" and "black and bad" is like reuniting East and West Germany. You're drawing a natural border around a compact area of the map. But being forced into the categories "white and bad" and "black and good" is about as natural as trying to merge East Germany and South Vietnam into the new country "Southeast Vietnermany". You're drawing an arbitrary boundary around two completely unrelated parts of the map and then begging in vain for the disgruntled inhabitants to cooperate with each other.

If you provoke a war between the reunified Germany and Southeast Vietnermany, and watch which side coordinates its forces better, you get the Implicit Association Test.

Why would we want to measure distance in thingspace? Loads of reasons. Take a set of pictures of famous cult leaders, mix them with a set of pictures of famous scientists, and test Less Wrong readers' reaction times associating a picture of Eliezer Yudkowksy's face with either set[8]. If it's easier to place him with the scientists, or there's no difference, that's some evidence we haven't become a cult yet. If it's easier to place him with the cult leaders, we should start worrying.

Tomorrow: some more serious applications to rationality.

Footnotes:

1: Most of these results taken from this, this, and this study.

2: There's some evidence that priming can change your IAT score. For example, subjects shown a picture of a happy black family enjoying a picnic just before an IAT got lower bias scores than a control group who didn't see the picture. And before condemning the test too much for its tendency to give different scores on different occasions, remember back to your school days when you'd have to take endless quizzes on the same subject. Occasionally just by chance you'd get a spread of ten point or so, and if you were on the borderline between passing and failing, you might very well pass one test and fail another test on the exact same material. This doesn't

mean grade school tests don't really measure your knowledge, just that there's always a bit of noise. The IAT noise is greater, but not overwhelmingly so.

3: There's also a fear someone might use it for, say, evaluating applicants for a job. Due to its weakness as an individual measurement and the uncertainty about how well it predicts behavior, this would be a terrible idea.

4: Full disclosure: Despite strongly opposing prejudice on a conscious level and generally getting along well with minorities in my personal life, I get assessed as moderately biased on the racism IAT. I had some memorable bad experiences with certain black people in my formative years, so this doesn't much surprise me.

5: In fact, Jesse Jackson (note for non-Americans: a well-known black minister and politician who speaks out against racism) himself admits to occasionally having these pangs of guilty fear - hence the name of Tetlock's article.

6: I think Eliezer once coined a term for the opposite of "applause light", for things like "racism" and "scientism" invoked only so people can feel good about hating them, but I can't seem to find it. Can someone refresh my memory?

7: I was split on whether to use the term thing-space or concept-space here. Eliezer uses concept-space in a very particular way, but "good" and "black" seem much more concepts than things. I eventually went with thing-space, but I'm not happy about it.

8: This is a facetious example. It's possible in theory, but there would be so much to control for that any result would be practically meaningless.

Fight Biases, or Route Around Them?

Continuation of: The Implicit Association Test
Response to: 3 Levels of Rationality Verification

I've not yet seen it pointed out before that we use "bias" to mean two different things.

Sometimes we use "bias" to mean a hard-coded cognitive process that results in faulty beliefs. Take as examples the in-group bias, the recall bias, the bad guy bias, and various other things discovered by Tversky and Kahneman.

Other times, we use "bias" to mean a specific faulty belief generated by such a process, especially one that itself results in other faulty beliefs. For example, Jews are sometimes accused of having a pro-Israel bias. By this we mean that they have a higher opinion of Israel than the evidence justifies; this is a specific belief created by the in-group bias. This belief may itself generate other faulty beliefs; for example, they may have a more negative opinion of Palestinians than the evidence justifies. It is both the effect of a bias, and the cause of other biases.

Let's be clear about this "more than the evidence justifies" bit. Hating Hitler doesn't mean you're biased against Hitler. Likewise, having a belief about a particular ethnic group doesn't mean you're biased for or against them. My Asian friends hate it when people sheepishly admit in a guilty whisper that they've heard Asians are good at academics. Asians *are* good at academics. Just say "55% chance an average Asian has a GPA above the American population mean" and leave it at that. This is one of Tetlock's critiques of the Implicit Association Test, and it's a good one. I'd probably link Asians to high achievement on an IAT, but it wouldn't be a bias or anything to get upset about.

And let's also be clear about this faulty belief thing. You don't have to *believe* something for it to be a belief; consider again the skeptic who flees the haunted house. She claims she doesn't belief in ghosts, and she's telling the truth one hundred percent. She's *still* going to be influenced by her belief in ghosts. She's not secretly supernaturalist any more than someone who gets "strongly biased" on the IAT is secretly racist. But she needs to know she's still going to run screaming from haunted houses, and IAT-takers should be aware they're still probably going to discriminate against black people in some tiny imperceptible way.

Okay, back to the example. So the President appoints Isaac, a synagogue-going Jew, as the new Middle East peace envoy. Due to some amazing breakthrough in the region, both the Israelis and Palestinians agree to accept whatever plan Isaac develops. Isaac's only job is to decide what long-term plan is best for both sides. And he's a good man: he has an honest desire to choose the maximum-utility solution.

Isaac legitimately worries that he has a bias for the Israelis and against the Palestinians. How can he test the hypothesis? He can take a hypothetical souped-up version of the Implicit Association Test[1]. He finds that yes, he has a strong pro-Israel anti-Palestine bias. Now what does he do?

He can try to route around the bias. This is the approach implicitly endorsed by Overcoming Bias and by rationalism in general. He can take the Outside View and look at successful approaches in other world conflicts. He can use some objective metric to

calculate the utility of everything in Israel, and check to make sure neither group is getting an amount disproportionate to their numbers. He can open a prediction market on metrics of success, and implement whatever policies trades at the highest value. All of these will probably improve Isaac's solution a fair bit. But none of them are perfect. In the end, Isaac's the one who has to make a decision that will be underdetermined by all these clever methods, and Isaac is still biased against the Palestinians.

Or he can try to fight the bias.

Diversity workshops try to fight biases directly . These don't work, and that's no surprise. Diversity workshops are telling you, on a conscious level, that minorities really are great people, aren't they? Well, yes. On a conscious level, you already believe that. Isaac already knows, on a conscious level, that the Palestinians deserve a fair solution that protects their interests just as much as the Israelis do. A diversity workshop would be a flashy video in which a condescending narrator explains that point again and again.

We don't have a lot of literature on what does work here, but I predict a few things would help. Make some Palestinian friends, to build mental connections between Palestinians and positive feelings. Learn to distinguish between Palestinian faces. Read works of fiction with sympathetic Palestinian characters. I would say "live in Palestine" but by all accounts Palestine is a pretty grim place; he might do better to live in a Palestinian community in America for a while.

Those techniques aren't especially good, but I don't care. We know how to improve them. By making a group take the Implicit Association Test, applying a technique to them, giving them the test again, and seeing how their score changed, we gain the ability to test bias-fighting techniques. I wouldn't want to do this on one person, because the test only has moderate reliability at the individual level. But a group of a few dozen, all practicing the same technique, would be quite sufficient. If another group learns a different technique, we can compare their IAT score improvement and see which technique is better, or if different techniques are better in different circumstances.

Again, there's no reason why this method should be limited to racial biases. No matter how hard I try to evaluate policies on their merits rather than their politics, I am biased towards the US Democratic Party and I know it. This ought to be visible on an IAT, and there ought to be techniques to cure it. I don't know what they are, but I'd like to find them and start testing them.

What about the second method of overcoming bias, routing around it? The IAT is less directly valuable here, but it's not without a role.

In one of the IAT experiments, subjects evaluated essays written by black or white students. This is a fiendishly difficult task upon which to avoid bias. A sneaky researcher can deliberately select essays graded as superior by a blind observer and designate them "white essays", so anyone trying to take the easy way out by giving all essays the same grade can be caught immediately. I like this essay task. It's utterly open to any technique you want to use to reduce bias.

So give someone IATs until you find a group they're especially biased against - black people, Palestinians, Korean-Americans, frequentists; any will do. Then make them grade essays by the control group and the disliked group. Collect statistics correlating

IAT bias with essay grading bias. If a person using a special technique to route around mental bias can grade essays more accurately than other people with the same level of IAT bias, that person has routed around their bias successfully.

So: How do we tell if a technique for routing around bias works? Test whether people are better able to conduct a rating task than their IAT scores would predict. How do we test a technique for fighting bias directly? See if it lowers IAT scores. All terribly inconvenient because of the IAT's low effect size and reliability, but with a large enough sample size or enough test-retest cycles the thing could be done. And the psychologists who transformed the Bona Fide Pipeline into the IAT may yet transform the IAT into something even more powerful.

This, then, is one solution to schools proliferating without evidence. With enough research, it could be turned into one of the missing techniques of rationality verification.

Footnotes

1: Remember, the IAT is only moderately good at evaluating individuals, and has a bad habit of changing its mind each time someone takes it. Much of what is in this essay would work poorly (though probably still better than nothing) with a simple IAT. But having someone take the IAT ten times over ten days and averaging the results might give a more accurate picture (I don't know of any studies on this). And in any case the IAT is quite good at comparing groups of people with sample size >1. And I expect that souped-up versions of the IAT will be out within a few years; these tests have gotten better and better as time goes on.