Edge

To arrive at the edge of the world's knowledge, seek out the most complex and sophisticated minds, put them in a room together, and have them ask each other the questions they are asking themselves.

https://www.edge.org/adversarial-collaboration-daniel-kahneman
Printed On Wed April 17th 2024

EdgeCast

[EDITOR'S NOTE: In marking this year's 25th anniversary of *Edge*, we are presenting original lectures from eminent scientists and other thinkers in the empirical world who are changing the way we think about science and our place in the world.]

"People don't change their minds."

—Daniel Kahneman

DANIEL KAHNEMAN is the Eugene Higgins Professor of Psychology Emeritus, Princeton University, author of *Thinking, Fast and Slow*, and co-author (with Cass R. Sunstein and Olivier Sibony) of *Noise*. He is the winner of the 2013 Presidential Medal of Honor, and the recipient of the 2002 Nobel Prize in Economic Sciences. **Daniel Kahneman's *Edge* Bio Page**

Adversarial Collaboration

32:48

Introduction
by John Brockman

Daniel Kahneman, while known for his work with Amos Tversky in the 1970s on judgment and decision-making, hopes that part of his scientific legacy will be the practice of adversarial collaboration, which he initiated in 2001.

Introduced as "a substitute for the format of critique-reply-rejoinder in which debates are currently conducted in the social sciences," Kahneman championed adversarial collaboration as "a good-faith effort to conduct debates by carrying out joint research."

He notes that the flowering of adversarial collaborations in various scientific disciplines is an indication that we are at the beginnings of a radical change in the way that science is done in psychology and other disciplines. "And," he says, "it's a major improvement."

—JB

ADVERSARIAL COLLABORATION

My first experience of an adversarial collaboration was about 40 years ago. My wife, Anne Treisman, and I were studying a new paradigm involving apparent motion and priming. It's a nice effect. There's a lot of work on it. And quite a few studies since have followed up on this work. Anne and I had many ideas, and we designed a large number of experiments, most of which succeeded. There was only one trouble. We didn't agree on the nature of the phenomenon, and we had different stories about the role of attention in the effect. The difference didn't prevent us from planning and interpreting useful experiments, but we found it difficult to construct a coherent theory.

I observed a phenomenon that I called the "15 IQ point benefit."

Then came February and the invitation to participate in the meeting of the Psychonomic Society in November. I suggested to Anne that we take two consecutive 25-minute slots, which would give us a whole hour. She said, "but we don't agree," which I answered by

"don't you believe in the scientific method? We'll run experiments to determine who of us is right. There's plenty of time before November."

So we started a cycle of critical experiments. I would design a study that I believed Anne could not explain, get her to agree that the result I expected would contradict her theory, and we would run that study. The results would come out as I had predicted, and there, I observed a phenomenon that I called the "15 IQ point benefit."

Within minutes of seeing the results, Anne would find a plausible explanation of why they were entirely compatible with her view. Anne was always very clever, but at those moments she would become quite extraordinary, able to come up with arguments that surprised and silenced me. Then I would go back to the drawing board, design another experiment, and it would all happen again. I was the aggressor in those games until Anne became exasperated and designed a critical test of my view. I agreed to the challenge. The results came in as Anne had expected, but it was my turn to get my 15 points, and I rejected the rejection of my theory. That was the end of that particular game.

We went on to give two good talks in November, somehow finessing the disagreement, but it took us eight more years before we published a paper summarizing these experiments, which got a substantial amount of attention. My faith in my naive version of the scientific method never recovered, and I want to comment on two parts of that story, the 15-point effect, and the fact that no minds were changed.

The extra wrinkle is hard to find—if it were easy, this would not be a serious critical test.

Why is it that we may agree in advance that a particular result is a fair test of our theory, then see so much more when the result is known? Why can't we anticipate our response to results that we do not expect to materialize?

The psychology of this is straightforward. The normal flow of reasoning is forward from what you believe to a possible consequence. When someone proposes a serious critical test, you cannot get from your theory to the result without adding an extra wrinkle to the theory. The extra wrinkle is hard to find—if it were easy, this would not be a serious critical test. On the other hand, the result probably follows from the adversary's theory. The lazy solution is to concede provisionally.

The situation changes completely when the result is known. It is a constraint and working backward to a slightly wrinkled theory is much easier. It's not the case that people refuse to admit that they had been wrong. From their perspective they were only wrong in failing to see that the experiment didn't prove anything. This is where the extra 15 IQ points come from. You can explain surprises that you could not anticipate.

The power of reasons is an illusion. The belief will not change when the reasons are defeated. The causality is reversed. People believe the reasons because they believe in the conclusion.

I was also impressed by the fact that Anne and I didn't change our minds. I had read Kuhn and Lakatos about the robustness of paradigms, but I didn't expect that minor theories would also be impervious to evidence. In fact, the stubborn persistence of challenged beliefs is much more general. To a good first approximation, people simply don't change their minds about anything that matters.

Let's start from the main domains where we know people don't change their minds—

politics or religion. When you ask people, why do you believe what you believe? They answer by giving reasons for their beliefs. Subjectively, we experience that reasons are prior to the beliefs that can be deduced from them. But we know that the power of reasons is an illusion. The belief will not change when the reasons are defeated. The causality is reversed. People believe the reasons because they believe in the conclusion.

In politics and in religion, the main driver is social. We believe what the people we love and trust believe. This is not a conscious decision to conform by hiding one's true beliefs. It's the truth. This is how we believe. Indeed, beliefs persevere even without any social pressure.

Classic studies by the late Stanford social psychologist Lee Ross established the phenomenon of belief perseverance. In those experiments, you first provide people with evidence that supports a particular belief. For example, you may give people the task of guessing which suicide notes are genuine, then provide feedback about accuracy. People draw inferences from what they're told. Those who have been given positive feedback, score themselves much higher on empathy than people who have been given negative feedback.

Then you discredit the feedback by telling people there was a mix-up and you test their beliefs about their empathy. The outcome? The elimination of the evidence does not eliminate the beliefs that were inferred from it. People who have raised their opinions of how empathetic they are, maintain their new belief, and the same is true if people have been convinced that they're not very good at guessing other people's feelings.

If a large and diverse body of published evidence supports a conclusion, you must believe in it. You are not allowed to believe only results that seem plausible to you.

I will now share a personal experience of belief perseverance that I cannot shake. Ten years ago, when I was young and foolish, I published *Thinking, Fast and Slow*. An important chapter in that book was concerned with behavioral priming. For example, the famous study in which people who have been made to think of old age walk more slowly than they normally would. I felt at the time that the evidence was important to my view. And I credited the author of that particular study, John Bargh, with being a significant influence on my work.

The spectacular results that I described in that chapter were mentioned in most reviews of my book and came up more often than anything else in the letters that I got about it. I was completely committed to believing those findings, and in arguing about them I made what I thought was a good point about science, that belief in results is not optional, that if a large and diverse body of published evidence supports a conclusion, you *must* believe in it. You are not allowed to believe only results that seem plausible to you.

Well, I was wrong, at least in that case. The studies of behavioral priming that I had cited in the chapter were largely discredited in the famous replication crisis of psychology. I'll come back later to what the authors of those studies made of this. As you may now expect they didn't change their mind, but I did. And I publicly retracted the chapter.

However, it turns out that I only changed my mind about the evidence. My view of how the mind works didn't change at all. The evidence is gone, but the beliefs are still standing. Indeed, I cannot think of a single important opinion that I have changed as a result of losing my faith in the studies of behavioral priming, although they seemed quite

important to me at the time.

Adversarial collaboration is an alternative to what I call "angry science."

The main thing I want to talk about here is adversarial collaboration, a notion that I introduced. The idea is that people who don't agree on a scientific idea commit themselves to work together towards a joint truth, either by experimentation or by discussion. Of course, what I've said about the phenomena of belief perseverance and the 15-point effect set limits to what can be achieved by adversarial collaboration. We can expect that even successful collaborations will end with few minds changing and we can expect adversaries to renege on their commitments to critical experiments. They will do this in good faith as beneficiaries of the 15-point increment.

If adversarial collaboration would rarely change minds, you may well wonder what's the good of it. There are several solid answers to this question. One of the answers is that adversarial collaboration is an alternative to what I call "angry science," which is my name for the way controversies are normally conducted.

Around the time that Anne and I were debating object reviewing, our joint work, my work with Amos Tversky on judgment and choice began to attract attention. And as it might be expected, not all that attention was friendly. I was exposed to the nasty world of critiques replies and rejoinders, and to the exhausting experience of trying to write a sensible review of articles that I thought infuriatingly unfair. Controversy is a terrible way to advance science. It's normally conducted as a contest, where the aim is to embarrass —sarcasm for beginners, and advanced sarcasm. Those things can go on forever. Gerd Gigerenzer published his first critique of our work thirty-seven years ago, and he's still not done with me.

The feature that makes most critiques intellectually useless is a focus on the weakest argument of the adversary. It is common for critics to include a summary caricature of the target position, refute the weakest argument in that caricature, and declare the total destruction of the adversary's position. It's rare for anyone to concede anything in replies and rejoinders. Doing angry science is a demeaning experience. I've always felt diminished by the sense of losing my objectivity when I get into the mode of scoring points in a debate. I hated it so much that I adopted a policy that Amos Tverksy thought irresponsible: I do not respond to hostile papers. And if a submitted manuscript makes me angry, I do not review it.

Old people don't really kick themselves. Their regret is wistful, almost pleasant. It's not emotionally intense.

Because I did not like published critiques, it was natural for me to suggest a collaboration when I found myself disagreeing with a piece of research. All the more so because the adversaries were friends. It's a simple story. I had published a claim that people are more likely to kick themselves about something they did than about something they did not do. For example, if John invested in company A, and knows that if he had invested in B, he would've made a hundred thousand dollars—more compared to Tom who held stocks in company B and sold them to buy stock in A.

They're in the same objective situation. But one of them did something, sold his stocks and the other didn't do something. He didn't buy the better stock. It's very clear that in that case, one of them feels more regret. At least I thought it was clear and there were data indicating that. But Tom Gilovich and Vicki Medvec published the finding that old

people spend a lot more time regretting the things they did not do than the things they did.

I had an answer, my 15 IQ points. I said that old people don't really kick themselves. Their regret is wistful, almost pleasant. It's not emotionally intense. We ran an experiment, and everyone was wrong. It turned out that delayed regret is mostly wistful, but it can be intense. I developed a protocol for what I named adversarial collaboration, unaware of a similar collaboration that had been carried out by Latham and Locke in 1988 and published as sort of an experiment in collaboration.

The next collaboration was much harder. It involved Ralph Hertwig, who was Gigerenzer's student and frequent collaborator, and is now his successor as the Director of the Max Plank Institute in Berlin. Here, we needed an arbiter that we both trusted, and we chose Barbara Mellers, now at Penn. That episode took a long time, and it was sometimes quite unpleasant because both Ralph and I really wanted to win and didn't entirely trust each other. We published a protocol for adversarial collaboration that reflected our experiences, including how to deal with arguments about the precise predictions to which we had committed ourselves. It turns out that a lot of notetaking is necessary. The protocol insisted on the mediator's responsibility for record keeping.

In an ideal world, scholars would feel obliged to accept an offer at adversarial collaboration. Editors would require adversaries to collaborate prior to, or instead of, writing independent exchanges.

The key statement in the protocol requires participants to accept in advance that the initial study will be inconclusive. Allow each site to propose an additional experiment, to exploit the font of hindsight wisdom that commonly becomes available when unwelcome results are obtained. And we ended up, Ralph and I, with considerable mutual respect. We concluded our paper on an upbeat note: "Despite our mishaps, we hope the approach catches on. In an ideal world, scholars would feel obliged to accept an offer at adversarial collaboration. Editors would require adversaries to collaborate prior to, or instead of, writing independent exchanges. Scientific meetings would devote time for scholars engaged in adversarial collaboration to present their joint findings. In short, adversarial collaboration would become the norm, not the exception."

On a personal level, Ralph and I have had very cordial relations ever since.

We theorists are not fully aware of the extent to which the experiments we plan and carry out are biased to favor our theoretical point of view.

I had another adversarial collaboration with a group of British economists. And a generalization seemed to emerge, which at the time I found disappointing. All three empirical collaborations in which I participated ended up with messy and incoherent results. I now believe that this outcome is both very likely and desirable. In every case, both sides made wrong predictions that were not confirmed. I think that's good, and I need to explain why it is good. I believe that we theorists are not fully aware of the extent to which the experiments we plan and carry out are biased to favor our theoretical point of view. I'm not alluding to a file drawer problem, to people hiding research that they don't like. The bias enters at the design stage. When you consider possible experiments, you apply your intuition to select those that are likely to support your view.

In an adversarial collaboration, the other side is pushing for experiments whose results are likely to be embarrassing to you, because your theory doesn't rule them out. Now, you

don't have to subscribe to a view that science only advances by refuting wrong ideas to accept that exposing the weaknesses of a theory is useful. And in a world in which neither adversary is likely to concede, it may be optimal for both of them to be wrong.

I had two non-empirical collaborations that deserve mention. The first was with two people I had considered friends, who wrote a very aggressive comment on a paper I had published about the evaluation of experiences. When I read their piece, I suggested an adversary collaboration as an experiment, but they turned me down. One of them said he thought a controversy would be more interesting. I wasted a lot of time doing something I had decided never to do—writing a snide reply to their critique. On the day before my reply was due, I decided to write those people, those former friends, pointing out that our exchange of biting comments would hurt all our reputations. And I suggested a format for a joint piece to replace the reply-rejoinder.

We started by stating what we agreed on, then we presented conflicting views on a series of topics. The outcome was vastly better for everybody than the alternative, and we ended up on civil terms. In general, a common feature of all my experiences has been that the adversaries ended up on friendlier terms than they started.

It took us six or seven years to write a paper that was titled "A Failure to Disagree." To get that far, both of us had to overcome objections from our tribes. People didn't want us to collaborate, which was strange.

My most satisfying experience of adversarial collaboration was with Gary Klein, the intellectual leader of a community of applied psychologists who were pretty much united in their rejection of the work I have been involved with. Gary is best known for his work on expert intuition, for which he's greatly admired, in apparent opposition to the work that Amos and I did on the limitations of intuitive thinking. It was clear, or should have been, that we're both right. Intuition is sometimes marvelous and sometimes flawed. The question is when, what are the boundaries of the marvels and the flaws?

I invited Gary to explore the question of boundaries. This wasn't easy. It took us six or seven years to write a paper that was titled "A Failure to Disagree." To get that far, both of us had to overcome objections from our tribes. People didn't want us to collaborate, which was strange.

There were two positive outcomes of those six or seven years of hard work. We became very warm friends and we agreed on the boundaries. We specified three conditions for trusting expert intuitions. But perhaps the most interesting outcome was that we did not really agree. Gary remained an admirer of expert intuitions and I remained the critic, and we continued to differ in our basic beliefs and in our intuitive tastes. In particular, we differed on what we found funny and delightful. For Gary, it is when bureaucratic rules lead to stupid mistakes. For me, it is when smug and self-satisfied experts fall flat in on their faces.

While thinking about this topic recently, I tried to identify positions that I will not give up. I found that they come in two kinds: there are methodological preferences and there are tastes for theories.

I don't like to reduce cognitive processes to flow charts and to mathematical models. I have a strong preference for facts.

When I look at the experimental methods that I prefer, I find that I like experiments that rule out challenges to my theories. I accept as challenges only experiments that are done

my way. I prefer simple and natural situations, and I prefer a particular experimental design that's called between subjects. I don't take very seriously results obtained by other methods, because I reject them as not real. The only real results are those that come out of my preferred methods. I discovered that I'm unreasonably stubborn about these methodological preferences, which define a terrain in which I am comfortable.

Deeper yet, it turns out that I have preferences about theories and theoretical styles. I like phenomenology and Gestalt psychology more than I like analysis. I don't like to reduce cognitive processes to flow charts and to mathematical models. I have a strong preference for facts over theory and I like irony. These stylistic tastes matter because they determine the character of the work I do. I realized while thinking about this that it is not an accident that critiques of my work are often very similar to critiques of gestalt ideas. The critics have different tastes on fundamental issues, they don't like my style and don't find my results compelling. I discovered the extent to which we talk across differences that are unbridgeable because we disagree on our tastes. To my surprise, I found that my basic psychological tastes were established and explicit in my early twenties before I went to grad school.

The last anecdotes that I want to share with you are about the replication crisis, which turns out to be relevant to the topic of adversarial collaboration in several ways. I already mentioned that the early focus of the replication crisis was behavioral priming. Specifically, someone failed to replicate Bargh's iconic study of slow walking. I got actively engaged because I believed in Bargh's work. A former student had told me sometime earlier that she didn't believe the study and wanted to replicate it. And my answer to her was "don't try to replicate because you will fail. I would also fail, and yet I believe the results. These experiments demand a sort of artistic directorial skill that neither you nor I possess. But I believe that John Bargh does, and that he is capable of replicating his own results. And that's enough for me."

I now think that this was probably another error of judgment on my part, but at that time, when I heard there would be a workshop on the problem of priming, I volunteered because I wanted to argue the case that some social psychologists have skills that are unique but solid.

However, during the two-day workshop that we had on that problem, I was impressed by the strength of the sentiment that was building up against priming. When I came home, I wrote a letter to a list of priming researchers whose names I got mostly from Bargh. In the letter, I warned the researchers about a looming train-wreck that would be especially damaging to their students on the job market. I identified myself as a general believer and reiterated my skepticism about replications by investigators who lack the special skills needed for successful priming research. That's what I believed then. I suggested an idea that I call "daisy chain" replications, where a group of labs that agree on the phenomenon and agree that behavioral priming is real get together. Each lab picks its favorite result. The result of lab A is replicated by lab B, the result of B replicated by C, and so on.

Social psychologists circled the wagons and developed a strong antipathy for the replicators. A President of the American Psychological Society called them "methodological terrorists."

One week later, the letter was leaked and published in *Nature* with an incendiary title: "Nobel Laureate tells social psychologists to clean up their act." I had naively failed to anticipate this outcome. Then all hell broke loose.

Believe it or not. I've been blamed for causing the replication crisis by attracting media attention to a minor problem. Some social psychologists have wondered about my motives for wanting to destroy social psychology by that letter, and I lost many friends.

The crisis provides ample evidence for the thesis that I'm developing today. People didn't change their minds. Social psychologists circled the wagons and developed a strong antipathy for the replicators. A President of the American Psychological Society called them "methodological terrorists," and another eminent psychologist suggested that people who have ideas of their own would not get involved in replications. There were essentially no takers for my suggestion that priming researchers should proactively replicate each other's work. This eventually convinced me that they did not have real confidence. They believed their findings were true, but they were not quite sure they could replicate them, and they didn't want to take the risk—another instance of belief perseverance.

Besides antagonizing social psychologists, I also managed to make myself unpopular among replicators when I published a paper on the etiquette of replication, which argued that replication should always be an adversarial collaboration. People argued that method sections should be sufficiently explicit to guarantee replicability without having to consult the author. I find this attitude shocking, just about as shocking as the defensiveness of priming researchers.

But none of this really matters. The crisis has been great for psychology. In terms of methodological progress, this has been the best decade in my lifetime. Standards have been tightened up, research is better, samples are larger. People pre-register their experimental plans and their plans for analysis. And behavioral priming research is effectively dead. Although the researchers never conceded, everyone now knows that it's not a wise move for a graduate student to bet their job prospects on a priming study. The fact that social psychologists didn't change their minds is immaterial.

A great flowering of adversarial collaboration followed the crisis. Along the way, several different advances were proposed in protocols for adversarial collaboration. There was one group that spent an entire week in a hotel to write a joint paper. There were protocols suggested for allowing adversarial collaboration between people who differ in rank and status. There were continued collaborations over several years. The combination of adversarial collaboration with pre-registration, which is becoming increasingly popular, is particularly useful.

The occasion for my thinking about this was that I was exposed to a massive adversarial collaboration. The Templeton Foundation is funding adversarial collaborations on the topic of consciousness. There are many theories of consciousness, and they are spending a total of twenty million dollars on five large adversarial collaborations. These projects are conducted with extremely careful protocols, including a requirement that the adversaries sign on to take seriously results that challenge their theory. I predict they will go back on this commitment, but that's the way it's done.

Most important, the research is collected by neutral laboratories. It's not collected by the adversaries themselves. I'm told that, in some cases, the theorists found giving up control quite disconcerting. There is a clear sense of a movement that is now spreading.

The replication crisis did not start in psychology. It started in medicine when a famous paper by Ioannidis claimed that most published research in medicine are false. Although the crisis did not start in psychology, the response to the crisis has been most impressive in my discipline. There has been a tightening of methodological standards, including "the

open science movement," which greatly increases the pre-commitments that researchers make before they do their research.

The path on which a theory is set eventually becomes obvious to everyone—except perhaps the theorist. And this is how adversarial collaborations can advance science without adversaries changing their mind.

They have to preregister their plans on a public site, and, in articles they publish later, they are obligated to focus on the results that they got from the parts of the experiment, or the analysis, that they had pre-registered. Results that they just discovered have a lower status are considered as hypotheses, not as established findings. This is a radical change in the way that science is done in psychology and in some other disciplines. And it's a major improvement.

Lakatos distinguishes two paths for theories that are challenged by unanticipated findings. One is progressive refinement; the other is defensive degeneration. The path on which a theory is set eventually becomes obvious to everyone—except perhaps the theorist. And this is how adversarial collaborations can advance science without adversaries changing their mind.

I want to end with a quote from Barb Mellers, who said, "do not change minds, just open a little wider." This is what dentists say when they're about to pull a tooth. And it is a good thing.

John Brockman, Editor and Publisher

Contact Info:editor@edge.org