Journal of Experimental Psychology: General
2016, Vol. 145, No. 7, 844–852

0096-3445/16/$12.00 http://dx.doi.org/10.1037/xge0000174

Caught Red-Minded: Evidence-Induced Denial of Mental Transgressions

Bethany A. Burum and Daniel T. Gilbert
Harvard University

Timothy D. Wilson
University of Virginia

We suggest that when confronted with evidence of their socially inappropriate thoughts and feelings, people are sometimes less likely—and not more likely—to acknowledge them because evidence can elicit psychological responses that inhibit candid self-reflection. In 3 studies, participants were induced to exhibit racial bias (Study 1) or to experience inappropriate sexual arousal (Studies 2 and 3). Some participants were then told that the researcher had collected physiological evidence of these mental transgressions. Results showed that participants who were told about the evidence were less willing to acknowledge their mental transgressions, but only if they were told before they had an opportunity to engage in self-reflection. These results suggest that under some circumstances, confronting people with public evidence of their private shortcomings can be counterproductive.

Keywords: confession, egalitarianism, interpersonal bias, self-insight, social interaction

Supplemental materials: http://dx.doi.org/10.1037/xge0000174.supp

> "Thoughtcrime is a dreadful thing, old man," he said sententiously. "It's insidious. It can get hold of you without your even knowing it. Do you know how it got hold of me? In my sleep! Yes, that's a fact. There I was, working away, trying to do my bit—never knew I had any bad stuff in my mind at all. And then I started talking in my sleep. Do you know what they heard me saying? Do you know what they heard me saying?" He sank his voice, like someone who is obliged for medical reasons to utter an obscenity. "'Down with Big Brother!' Yes, I said that! Said it over and over again, it seems. Between you and me, old man, I'm glad they got me before it went any further. Do you know what I'm going to say to them when I go up before the tribunal? 'Thank you,' I'm going to say, 'thank you for saving me before it was too late.'"
>
> —*Nineteen Eighty-Four* (George Orwell, 1949, p. 233)

Throughout history, people have been punished for crimes of thought—for believing or not believing in particular gods, for approving or not approving of particular leaders, for having or not having particular sexual interests. Although such thoughts still constitute a moral turpitude in many parts of the world, members of Western societies are generally free to believe in one god, two gods, or no gods at all; to love or hate or ignore their presidents; and to fantasize about kissing men or women, or both or neither. What Westerners are not free to do—at least not if they hope to be respected, elected, befriended, or employed—is believe that Blacks are lazy, Jews are cagey, and women cannot do math. Among educated Westerners, egalitarian beliefs and values are mandated as strongly as religious beliefs, political opinions, and sexual interests are in other parts of the world.

And that mandate is as avidly enforced—typically not by law, but by normative pressure and social sanction. Westerners are regularly excoriated by their peers and the press for private thoughts and feelings that were presumably revealed by their public choice of words, their tones of voice, or the length and direction of their gazes. When the Lieutenant Governor of California mispronounced the word "Negro" as "Nigro" during a speech, some attendees walked out and he was pilloried in the press (Tamaki, 2001). When a male gubernatorial candidate in South Carolina mistakenly blended the words "her" and "door" and said of his female opponent, "We are going to escort whore out the door," he was widely castigated (Henderson, 2014). When an aid to the mayor of Washington, DC referred to the budget as "niggardly" (which means "stingy" and is unrelated to the offensive word it sounds like), he was roundly criticized and forced to resign (Woodlee, 1999). Incidents such as these are not unusual and appear to be on the rise (Ronson, 2015). In classrooms and board rooms, on TV and social media, people routinely scrutinize each other's behavior for subtle signs of nonegalitarian bias and then confront each other about them. Those confrontations are neither easy nor inevitable, but they do happen regularly and often

This article was published Online First April 28, 2016.

Bethany A. Burum and Daniel T. Gilbert, Department of Psychology, Harvard University; Timothy D. Wilson, Department of Psychology, University of Virginia.

All authors developed the study concepts, contributed to the study designs, and drafted and approved the manuscript. B. A. Burum collected and analyzed the data. The protocols for all studies were approved by Harvard University's Committee on the Use of Human Subjects and were carried out in accordance with the provisions of the World Medical Association Declaration of Helsinki.

We thank Nicholas Adan, Alex Brownman, Sarah Coughlon, Michael Ding, C.C. Gong, Katie Goldin, Nicolas Govea, Chris Hernandez, Maria Kalina, Ashley Kirsner, Ina Kodra, Juliet Nelson, Jill Jacobbi, Nicholas Johnson, Jillian Jordan, Katie Lantz, Hannah Leverson, Bay McCulloch, Danny Meltzer, Lauren Rhodewalt, Preeti Srinivasan, Angela Wu, and Ege Yumusak, and Starielle for their help with the execution of these studies, and acknowledge the support of Grant 41907 from the John Templeton Foundation and Grant BCS-1423747 from the National Science Foundation.

Correspondence concerning this article should be addressed to Daniel T. Gilbert, Department of Psychology, Harvard University, Cambridge, MA 02138. E-mail: gilbert@wjh.harvard.edu

(Feagin, 1991; Swim & Hyers, 1999; Swim, Hyers, Cohen, Fitzgerald, & Bylsma, 2003).

So what do people do when someone suggests that they are racist, sexist, or homophobic? Often they deny it (Czopp, Monteith, & Mark, 2006). Moments after mispronouncing the word "Negro," the lieutenant governor leapt to his own defense: "If you heard what I think I heard, I want you to know it wasn't me; it's not the way I was raised, it's not the way I was taught, it's not the way I raised my children and it's not what's in my heart" (Tamaki, 2001). Being accused of bias naturally makes most people feel bad (Amodio, Devine, & Harmon-Jones, 2007; Czopp, Monteith, & Mark, 2006; Leary, Terry, Batts Allen, & Tate, 2009; Monteith, 1993; Monteith, Ashburn-Nardo, Voils, & Czopp, 2002), so it is not surprising that people often attempt to refute the allegation. And yet, on other occasions, people who are confronted about their mental transgressions acknowledge them—taking stock, offering apologies, and making concerted efforts to change their ways (e.g., Amodio, Devine, & Harmon-Jones, 2007; Czopp, Monteith, & Mark, 2006; Monteith et al., 2002). The mayoral aide who used the word "niggardly" apologized for his poor lexical decision, blamed no one but himself for the loss of his job, and reported that the incident had given him "a certain awareness" that he previously lacked: "I used to think it would be great if we could all be colorblind," he said. "That's naive, especially for a white person, because a white person can afford to be colorblind. They don't have to think about race every day. An African-American does" (Woodlee, 1999).

When do people deny their mental transgressions and when do they acknowledge them? Common sense suggests that people should be more willing to acknowledge a private mental transgression when there is public evidence of it (Jones & Sigall, 1971; Roese & Jamieson, 1993). After all, when people who have committed *behavioral* transgressions (such as robbery or murder) are presented with evidence (such as fingerprints or security camera footage), their willingness to acknowledge those transgressions increases dramatically (Evans, Hansen, & Mittelmark, 1977; Moston, Stephenson, & Williamson, 1992; Murray & Perry, 1987), which is why police interrogators routinely pretend to have such evidence even when they do not (Perillo & Kassin, 2011). By the same logic, we might expect a person to be more willing to acknowledge a mental transgression when that person is confronted with evidence of it—for instance, a video of the person telling a sexist joke at a party or an e-mail containing a racist remark.

But confronting people with evidence of their mental and behavioral transgressions may not have similar consequences because although people generally *know* when they have transgressed behaviorally, they do not always know when they have transgressed mentally (Greenwald & Banaji, 1995; Pronin, Lin, & Ross, 2002). A man accused of bank robbery does not need to stop and think about whether he really robbed a bank, but a man accused of sexism may well need to stop and think about whether he really treats his male and female employees equally. Before people can acknowledge their mental transgressions to others they must first acknowledge them to themselves, and that can take some time. That's why under certain circumstances, confronting people with evidence of their mental transgressions may actually make them less likely rather than more likely to acknowledge those transgressions. If the presentation of evidence triggers the impulse to defend oneself, it may preclude or inhibit the candid self-reflection that is required for people to recognize their own foibles (Rudman, Dohn, & Fairchild, 2007). A person who privately asks herself "Might I be biased?" can calmly examine her attitudes and beliefs, and may ultimately conclude that the answer is yes; but a person who is publicly confronted by evidence indicating that others already think she is biased may be less inclined to search her soul and more inclined to search for ways to defend herself. People may be capable of taking a fearless and searching moral inventory and acknowledging the errors of their ways, but they may not be able to do this when they are busy mounting a defense against evidence. In short, people may be less inclined to examine the content of their characters when they are worried that others are examining them too, and as such, confronting people with evidence of their nonegalitarian biases may have the paradoxical effect of making people less likely—and not more likely—to acknowledge those biases. We tested this hypothesis in three studies.

Study 1

In Study 1, we showed White participants mug shots of suspected criminals, some of whom were Black men, and then told them that we were studying "whether people are influenced by race when perceiving threat." Some participants were then falsely told that as they had viewed the mug shots, we had gathered evidence of their racial bias by surreptitiously measuring their galvanic skin response, which indicated whether they had felt more threatened by Black than by White faces. We assumed that (a) participants would not be entirely sure whether they had in fact felt more threatened by Black than by White faces, and (b) the purported physiological evidence of racial bias would strike our participants as suggestive but not conclusive—that is, it would be plausibly deniable. We then gave participants an opportunity to engage in candid self-reflection, and then measured their willingness to acknowledge their racial bias. We predicted that those participants who were told that we had evidence of their racial bias would, in fact, be least willing to acknowledge it.

Method

Participants. Students from the Harvard University study pool were recruited for an approximately 30-min study in which they would "make quick judgments about images" in exchange for either $10 or course credit. We committed to running the study until the academic term ended and participants were no longer readily available. Ninety-three students (54% female; mean age = 20 years, SD = 1.65 years) participated.

Procedure. On arriving at the laboratory, participants were escorted to a room where they remained for the duration of the session. Participants were seated in front of a computer with a joystick. The experimenter explained that he or she was interested in the accuracy of eyewitness testimony, and that the present study involved measuring the accuracy with which people can make judgments about suspected criminals after seeing their faces for only a few seconds. Participants were told that they would see a series of mug shots of suspected criminals, and that their task was to estimate the age of the person shown in each mug shot. Specifically, participants were told that if the person in the mug shot

was under 30 years old they should pull the joystick toward them, and if the person was over 30 years old they should push the joystick away from them. Participants were told to keep their hand firmly on the joystick for the duration of the age-estimation task. Participants were then shown a series of 32 mug shots in a random sequence. All the people in the mug shots were male, and 15 of them were White and 17 of them were Black. Each mug shot remained on the screen for 5 seconds and was followed by text that prompted the participant to use the joystick to respond.

After participants completed the age-estimation task, the experimenter returned to the laboratory room and read the following statement to all participants:

> One of the things we are interested in is whether people are influenced by race when perceiving threat. Many people find African-Americans more threatening than Caucasians. These race-based effects have important implications for both criminal cases and African-Americans' overall quality of life. But there are also individual differences in the degree to which race influences people's perception of threat. Some people are very influenced by race, and some people are much less influenced or not influenced at all by race. These individual differences can have important implications for how Caucasian people interact with African Americans, and thus for inter-race relations more broadly.

This statement was intended to motivate participants to reflect on the possibility that they had been racially biased. But before participants had an opportunity to reflect on this possibility, the experimenter randomly assigned them to the *evidence* condition or the *no evidence* condition and then immediately read the following statement to participants in the evidence condition, but not to participants in the no evidence condition:

> What I wasn't able to tell you before is that to find out how influenced you were by race we were measuring the threat you felt while viewing the faces we showed you. We did this by using the joystick you were holding to measure your galvanic skin response. When people experience threat, there is a subtle increase in the perspiration on their palms that changes how conductive their skin is. This change is called the galvanic skin response, and is a reliable indicator of whether you are feeling threatened. The joystick was sending information on your moment-by-moment galvanic skin response to a computer in another room, where a research assistant was using it to classify in real time whether you felt more threatened by African-American faces than by Caucasian faces.

In actuality, the joystick was not capable of measuring GSR. Next, the experimenter left all participants alone to complete a 10-min task that was said to be "a measure of your aesthetic preferences." Participants were shown 100 pairs of characters from an Ethiopian alphabet and were simply asked to press a key to indicate which character they preferred in each pair. Each pair of characters appeared on a computer screen for 5 sec. This task was designed to be extremely easy and require very little attention, therefore giving participants ample opportunity to engage in self-reflection. We will hereinafter refer to the 10-min interval during which participants performed the aesthetic preferences task as "the reflection period."

After the reflection period, we measured participants' willingness to acknowledge their racial bias. Specifically, we asked them "To what degree did you feel more threatened by the African-American mug shots than by the Caucasian mug shots?" We asked 13 additional questions, all of which are shown in Table 1 in the order in which they were asked. Participants answered all questions using 7-point Likert scales that were anchored at their endpoints with the phrases shown in Table 1. Finally, participants completed the Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988) and the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960), provided their age, gender, race, level of English proficiency, and university affiliation, and were thoroughly debriefed.

Results

Although the study required that we test non-Black participants, we did not want to deprive Black students of the opportunity to participate and earn money or course credit, so we ran but did not examine or include the data from 11 students who identified themselves as African American. Also, before examining the data, we omitted the data from four participants who told us that they did not believe that we had measured their galvanic skin response, two participants for whom experimenter errors substantially changed the procedure, one participant who indicated that she had already participated in the study at an earlier time, and one participant who used his phone during the study. This left 74 participants in the data set (38 who identified as White, 20 who identified as Asian American, 8 who identified as Hispanic, 6 who identified as Other, and 1 who preferred not to indicate a race; 57% female; mean age = 20 years, SD = 1.69 years).

We compared the responses of participants in the evidence and no evidence conditions, and the results are shown in Table 1. As the first row shows, our primary prediction was confirmed: Telling participants that we had evidence of their racial bias decreased rather than increased the likelihood that they would acknowledge that bias themselves. The magnitude of the mean in the third row suggests that participants in the evidence condition found the evidence plausibly deniable, as we intended. The only significant difference between conditions on any other measure was a tendency for participants in the evidence condition to report feeling slightly more positive affect than participants in the no evidence condition.

Study 2

Participants in Study 1 who believed that an experimenter had evidence of their racial bias were less willing—and not more willing—to acknowledge that bias. Our theorizing suggests that this happened because being confronted by evidence elicits responses that inhibit self-reflection. If this is true, then evidence should reduce a person's willingness to acknowledge a mental transgression when that evidence is presented *before* the person has had an opportunity to engage in self-reflection, but not when the evidence is presented *after* the person has had that opportunity. In Study 2, we sought to show that the timing of the presentation of the evidence does indeed determine whether it will reduce the likelihood of acknowledgment.

In addition, we sought to generalize the results of Study 1 by studying a different kind of mental transgression. In Study 2, we asked heterosexual male participants to watch a video of an attractive female as she tried on different bathing suits in a

Table 1
Results for All Measures in Study 1

Measure	No evidence condition (n = 35)	Evidence condition (n = 39)	Effect of condition
To what degree did you feel more threatened by the African-American mug shots than by the Caucasian mug shots? (1 = *I did not feel more threatened by the African-Americans*, 7 = *I felt a lot more threatened by the African-Americans*)	3.00 (1.99)	2.18 (1.37)	$t(72) = 2.09$ $p = .041$
To what degree do you think the average student your age would feel more threatened by the African-American mug shots than by the Caucasian mug shots? (1 = *The average student would not feel more threatened by the African-Americans*, 7 = *The average student would feel a lot more threatened by the African-Americans*)	4.17 (1.45)	4.03 (1.33)	$t(72) = .45$ $p = .652$
How accurately do you think we measured the threat you felt using your galvanic skin response? (1 = *Not at all accurately*, 7 = *Extremely accurately*)		4.41 (1.29)	
How threatening did you find the mug shots to be in general? (1 = *Not at all threatening*, 7 = *Extremely threatening*)	3.69 (1.28)	3.26 (1.22)	$t(72) = 1.44$ $p = .153$
How do you feel about the degree to which you found the mug shots threatening in general? (1 = *Extremely bad*, 7 = *Extremely good*)	4.12 (.98)	4.03 (1.16)	$t(72) = .36$ $p = .717$
How threatening did you find the African American mug shots to be in general? (1 = *Not at all threatening*, 7 = *Extremely threatening*)	3.51 (1.52)	3.10 (1.27)	$t(72) = 1.27$ $p = .209$
How do you feel about the degree to which you found the African American mug shots threatening in general? (1 = *Extremely bad*, 7 = *Extremely good*)	3.49 (1.36)	3.95 (1.28)	$t(72) = -1.51$ $p = .135$
How threatening did you find the Caucasian mug shots to be in general? (1 = *Not at all threatening*, 7 = *Extremely threatening*)	3.41 (1.21)	3.18 (1.19)	$t(72) = .83$ $p = .412$
How do you feel about the degree to which you found the Caucasian mug shots threatening in general? (1 = *Extremely bad*, 7 = *Extremely good*)	3.83 (.95)	3.87 (.89)	$t(72) = -.20$ $p = .841$
How accurately do you think you determined the ages of the mug shots? (1 = *Not at all accurately*, 7 = *Extremely accurately*)	3.83 (1.32)	3.74 (1.41)	$t(72) = -.27$ $p = .790$
How engaging did you find the task of viewing the mug shots and determining their ages? (1 = *Not at all engaging*, 7 = *Extremely engaging*)	4.06 (1.68)	4.27 (1.50)	$t(72) = .57$ $p = .572$
How difficult did you find the task of viewing the mug shots and determining their ages? (1 = *Not at all difficult*, 7 = *Extremely difficult*)	4.38 (1.48)	4.39 (1.18)	$t(72) = .04$ $p = .969$
Negative affect score (PANAS)	1.68 (.72)	1.54 (.57)	$t(68) = -.86$ $p = .391$
Positive affect score (PANAS)	2.23 (.74)	2.61 (.75)	$t(64) = 2.09$ $p = .040$
Social desirability score (Marlowe-Crowne)	3.43 (6.05)	2.85 (5.25)	$t(72) = -.44$ $p = .659$

Note. Column 1 shows measures and their response scale anchors. Columns 2 and 3 show means and standard deviations. Column 4 shows values for t and p. Because some participants skipped some items, degrees of freedom may differ between measures. PANAS = Positive and Negative Affect Schedule.

store's dressing room. We told participants that the video had been taken by a stranger without the woman's knowledge and then posted on the Internet, and that the woman had been adversely affected by this crime. After participants watched the video, we told them that "one of the things that we are interested in is how sexually aroused men feel to voyeuristic videos such as this." We then falsely told them that as they had been watching the video, the experimenter had been surreptitiously recording their pupillary dilation and eyeblink rate and that these were indicators of sexual arousal. We assumed that (a) participants would not be entirely sure whether they had felt sexually aroused during the video, and (b) the purported physiological evidence of sexual arousal would strike our participants as suggestive but not conclusive—that is, it would be plausibly deniable. We told some participants about the evidence before the reflection period, and we told others about the evidence only after the reflection period. We also included two control conditions that we will describe shortly. We then measured participants' willingness to acknowledge their inappropriate sexual arousal. We predicted that presenting participants with evidence *before* they had had a chance to reflect would reduce their willingness to acknowledge their sexual arousal, but that presenting them with evidence *after* they had had a chance to reflect would not.

Method

Participants. Male students from the Harvard University study pool were recruited for an approximately 30 min study on "watching videos" in exchange for either $5 or course credit. We committed to running the study until the academic term ended and participants were no longer readily available. Eighty-seven male students (mean age = 21 years, SD = 2.14 years) participated.

Procedure for all conditions. On arrival at the laboratory, participants were escorted to a room equipped with a computer where they remained for the duration of the session. The experimenter told participants that they would be watching a video of a crime that had ostensibly been committed in 2008, when a male employee at a Philadelphia clothing store had installed a hidden camera in a dressing room and used it to film female shoppers as they undressed. Participants were told that the employee had uploaded some of these videos to a video-sharing website, and that

one of the female shoppers whose videotape had been uploaded (hereinafter referred to as the victim) had been alerted to that fact by a friend. To ensure that participants realized that the victim had been adversely affected by this crime, participants were shown a statement that the victim had ostensibly posted on the video-sharing website:

> You may think of this as just another 'hot' video but I am a real person and some creep taped me when I was in that dressing room and then posted it. Now my dad and brother both saw it and so did a bunch of my friends and now it is all over my school and I am totally humiliated and always worrying now that some perv might be spying on me. You guys who left comments about my 'hot rack' should think about how much these kinds of things hurt real girls like me. How would you feel if some asshole did this to your sister and wrecked her life like they did to mine?

Next, the experimenter left the room and participants watched an approximately 3.5 min video that appeared to have been taken by a camera hidden in the ceiling of a dressing room. The video showed an attractive young woman (who was actually an actress) changing into and out of three bathing suits, posing as she examined herself in the mirror. Careful positioning of the actress at critical moments ensured that the video was revealing but contained no nudity.

When the video ended, the experimenter made the following statement to all participants: "One of the things that we are interested in is how sexually aroused men feel to voyeuristic videos such as this." This statement was intended to motivate participants to reflect on the possibility that they may have experienced sexual arousal while viewing the victim of a crime. But before they had a chance to reflect, the experimenter randomly assigned participants to one of two experimental conditions (the *immediate evidence* condition or the *delayed evidence* condition) or to one of two control conditions (the *no evidence condition* or the *no reflection* condition). For ease of exposition, we will first describe the two experimental conditions and then describe the two control conditions.

Procedure for experimental conditions. After making the statement that was intended to motivate participants to reflect, the experimenter immediately read the following statement to participants in the immediate evidence condition:

> What I wasn't able to tell you before is that we were videotaping you while you watched the video with a hidden camera placed behind a one-way mirror. This camera is very fine resolution, enabling us to collect second-by-second information about your pupil dilation and eye blink rate, which are reliable indicators of sexual arousal. The camera feeds directly into a computer in another room, where a research assistant was using this information to classify your sexual arousal to the film in real time.

In actuality, there was no hidden camera. After being told about the hidden camera, participants in the immediate evidence condition were given an opportunity to engage in self-reflection by completing the same 10-min aesthetic preferences task that we used in Study 1. When they were finished, participants in the immediate evidence condition were asked to answer the question, "How sexually aroused did you feel while watching the video?" as well as 7 additional questions. Participants answered all questions using 7-point Likert scales that were anchored at the endpoints with the phrases *Not at all* and *Extremely*. Next, participants in the immediate evidence condition completed the Marlowe-Crowne Social Desirability scale, provided their age, gender, sexual orientation, and relationship status, and were thoroughly debriefed.

Participants in the delayed evidence condition heard and did all the same things that participants in the immediate evidence condition heard and did, but in a different order. Specifically, participants in the delayed evidence condition first heard the statement that was intended to motivate them to reflect ("One of the things that we are interested in is how sexually aroused men feel to voyeuristic videos such as this"), were then given the opportunity to reflect (i.e., the 10-min aesthetic preferences task), were then were told about the presence of the hidden camera ("What I wasn't able to tell you before . . ."), were then asked to complete all the dependent measures, and were then thoroughly debriefed. The shaded box in the center of Figure 1 shows the sequence of events in the two experimental conditions.

Procedure for the control conditions. We also included two control conditions. First, we included a *no evidence* condition that was identical to the delayed evidence condition except that the participants were never told that there was a hidden camera measuring their pupil dilation and eyeblink rate. Our hypothesis suggests that because delayed evidence is presented after reflection has already occurred, it should have no impact on participants' willingness to acknowledge their sexual arousal. In other words, a delay in learning about evidence should be the same as not learning about evidence at all. As such, we expected there to be no differences between the no evidence condition and the delayed evidence condition. Second, we included a *no reflection period* condition that was identical to the immediate evidence condition except that participants never performed the aesthetic preferences task and therefore had no opportunity to engage in reflection. Our hypothesis suggests that the immediate presentation of evidence inhibits self-reflection, and as such, the opportunity to reflect should be superfluous and should have no impact on the willingness of participants to acknowledge their inappropriate sexual arousal. In other words, learning about evidence before the reflection period should be the same as having no reflection period at all. As such, we expected no differences between the no reflection condition and the immediate evidence condition. The unshaded portions of Figure 1 show the sequence of events for the two control conditions.

Results

Before examining the data, we omitted the data from two participants who expressed suspicion that they were being monitored as they watched the video, one who appeared to be intoxicated, and seven who identified themselves as gay. This left 77 participants for analysis (mean age = 21, SD = 2.18).

Because the four conditions in Study 2 did not constitute a fully factorial design, all dependent measures were submitted to separate one-way ANOVAs. As the first row of Table 2 shows, our primary prediction was confirmed: Participants who were presented with evidence before they had a chance to reflect were less willing to acknowledge their sexual arousal than were participants who were presented with evidence after they had

had a chance to reflect, $t(37) = 2.21$, $p = .033$. The two control conditions shed additional light on this result. First, participants in the delayed evidence condition and the no evidence condition were equally willing to acknowledge their sexual arousal, $t(34) = 1.13$, $p = .267$, suggesting that once participants had a chance to reflect, the subsequent presentation of evidence had no effect on their willingness to acknowledge their sexual arousal. Second, participants in the immediate evidence and no reflection period conditions were equally unwilling to acknowledge their sexual arousal, suggesting that when evidence was presented immediately, the reflection period became superfluous and had no effect on their willingness to

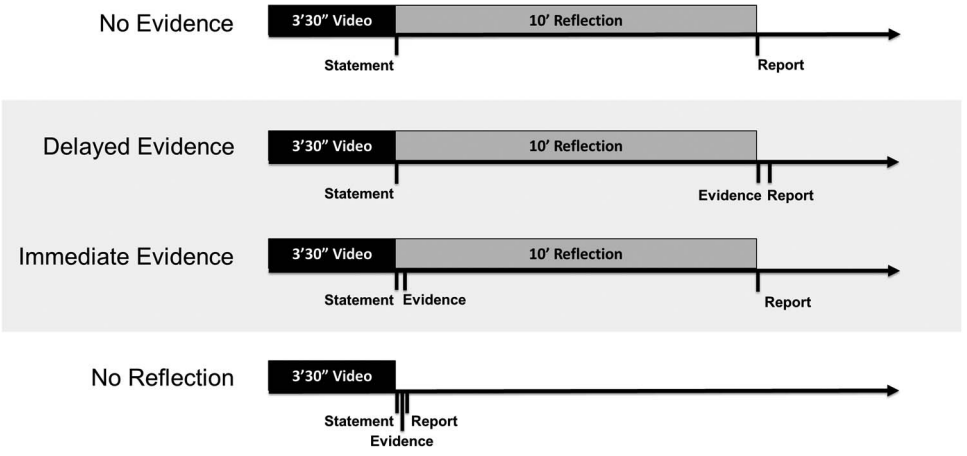

Figure 1. The experimental conditions in Study 2 are shown in the shaded box and the control conditions in Study 2 are shown outside the shaded box. *3′30″ video* refers to the 3 min and 30 sec video that participants saw; *10′ Reflection* refers to the 10-min reflection period that some participants were given; *Statement* refers to the time at which the experimenter read the statement intended to motivate participants to reflect; *Evidence* refers to the time at which the experimenter told participants that their physiological reactions had been monitored; and *Report* refers to the time at which participants reported how sexually aroused they had been during the video.

Table 2
Results for All Measures in Study 2

Measure	No evidence ($n = 18$)	Immediate evidence ($n = 21$)	Delayed evidence ($n = 18$)	No reflection ($n = 20$)	Effect of condition
How sexually aroused did you feel while watching the video?	3.83 (1.15)	2.24 (1.18)	3.28 (1.74)	2.60 (1.31)	$F(1, 73) = 5.25$ $p = .002$
How much did you enjoy watching this video?	2.61 (.78)	2.76 (1.58)	2.56 (1.72)	2.20 (1.11)	$F(1, 73) = .63$ $p = .601$
How attractive did you find the woman in this video?	5.50 (.71)	4.00 (1.82)	4.53 (1.93)	4.95 (1.40)	$F(1, 73) = 3.28$ $p = .026$
How serious was this invasion of privacy?	6.44 (.71)	6.38 (.92)	6.06 (1.55)	6.65 (.59)	$F(1, 73) = 1.14$ $p = .338$
How uneasy did you feel about this video while you were watching it?	5.00 (1.78)	4.38 (1.80)	4.56 (1.19)	4.95 (1.76)	$F(1, 73) = .65$ $p = .587$
How accurately do you believe that trained researchers can gauge sexual arousal through observing pupil dilation and eye blink rate?	5.19 (.91)	4.48 (1.94)	4.72 (1.81)	4.48 (1.52)	$F(1, 71) = .74$ $p = .531$
How accurately do you believe that the trained researcher in this study was able to gauge your sexual arousal through the information about your pupil dilation and eye blink rate that we collected with our high definition camera?	4.19 (1.76)	3.81 (1.70)	4.67 (1.72)	4.30 (1.63)	$F(1, 71) = .84$ $p = .474$
To what degree are you generally concerned about invasions of privacy?	4.82 (1.70)	4.48 (1.75)	5.19 (1.51)	4.95 (1.47)	$F(1, 72) = .68$ $p = .569$
Social desirability score (Marlowe-Crowne)	12.44 (6.58)	14.00 (7.09)	15.72 (6.23)	13.89 (3.73)	$F(1, 72) = .88$ $p = .456$

Note. Column 1 shows measures. Columns 2 through 5 show means and standard deviations. Column 6 shows values for t and p. Because some participants skipped some items, degrees of freedom may differ between measures.

acknowledge their sexual arousal, $t(39) = 0.93$, $p = .359$. A linear contrast that tested for this pattern of results (No Evidence = 1, Delayed Evidence = 1, Immediate Evidence = −1, and No Reflection = −1) was statistically significant, $t(73) = 3.66$, $p < .001$.

The only other measure that differed between conditions was the report of the victim's attractiveness. Although the immediate evidence and delayed evidence conditions did not differ on this measure, $t(37) = 0.88$, $p = .385$, participants in the no evidence condition did find the victim more attractive than did participants in either the immediate evidence condition, $t(37) = 3.29$, $p = .002$, or the delayed evidence condition, $t(34) = 2.01$, $p = .053$. This may be because participants in the no evidence condition answered this question closer to the time that they last saw the victim than did participants in any other condition. The magnitude of the means on the remaining measures suggests that participants found the victim attractive, considered the crime very serious, felt uneasy about watching the video, and believed that pupillary dilation and eyeblink rate provide suggestive but not conclusive evidence of sexual arousal.

Study 3

Method

Design. Because the sample size in Study 2 was small, we thought it was important to include a replication of the study's primary finding with a larger sample. A power analysis suggested that 64 participants (32 in each condition) would give us an 80% chance of detecting the effect seen in Study 2.

Participants. Male students from the Harvard University study pool were recruited for an approximately 30-min study on "watching videos" in exchange for either $5 or course credit. We committed to running the study until the academic term ended and participants were no longer readily available. By the end of the semester, we were able to recruit 66 students (mean age = 21 years, $SD = 2.18$ years) for the study.

Procedure. The procedures for Study 3 were identical to the procedures used in the two experimental conditions of Study 2 except that participants in Study 3 completed the PANAS instead of the Marlowe-Crowne Social Desirability scale, and were also asked at the end of the study whether they had ever seen the video before or whether they recognized the victim.

Results

Before examining the data, we omitted the data from two participants who expressed suspicion that they were being monitored as they watched the video, one who did not believe the cover story about the video, and four who identified themselves as gay. This left 59 participants in the data set (mean age = 21, $SD = 2.13$). All participants indicated that they had not seen the video before and did not recognize the victim.

We compared the responses of participants in the two conditions, and the results are shown in Table 3. As the first row indicates, the main finding of Study 2 was replicated: Telling participants that we had evidence of their sexual arousal decreased the likelihood that they would acknowledge that arousal themselves. There were no differences between conditions on any of the other measures. The magnitude of the means on other measures suggests that participants in both conditions found the victim attractive, considered the crime very serious, felt uneasy about watching the video, and believed that pupillary dilation

Table 3
Results for All Measures in Study 3

Measure	Immediate evidence ($n = 33$)	Delayed evidence ($n = 26$)	Effect of condition
How sexually aroused did you feel while watching the video?	3.03 (1.24)	3.69 (1.35)	$t(57) = 1.96$ $p = .055$
How much did you enjoy watching this video?	2.82 (1.42)	3.42 (1.55)	$t(57) = 1.56$ $p = .125$
How attractive did you find the woman in this video?	5.21 (.96)	5.08 (.74)	$t(57) = .59$ $p = .557$
How serious was this invasion of privacy?	6.24 (.87)	6.15 (.97)	$t(57) = .37$ $p = .713$
How uneasy did you feel about this video while you were watching it?	4.94 (1.56)	4.23 (1.99)	$t(57) = 1.54$ $p = .130$
How accurately do you believe that trained researchers can gauge sexual arousal through observing pupil dilation and eye blink rate?	4.61 (1.17)	4.62 (1.53)	$t(57) = .03$ $p = .978$
How accurately do you believe that the trained researcher in this study was able to gauge your sexual arousal through the information about your pupil dilation and eye blink rate that we collected with our high definition camera?	4.27 (1.35)	4.62 (1.68)	$t(57) = .87$ $p = .388$
To what degree are you generally concerned about invasions of privacy?	4.70 (1.57)	4.34 (1.72)	$t(57) = .82$ $p = .417$
Negative affect score (PANAS)	5.76 (.95)	5.13 (1.35)	$t(31) = 1.58$ $p = .124$
Positive affect score (PANAS)	3.86 (1.05)	3.94 (.80)	$t(31) = .23$ $p = .820$

Note. Column 1 shows measures. Columns 2 and 3 show means and standard deviations. Column 4 shows values for t and p. Because some participants skipped some items, degrees of freedom may differ between measures. PANAS = Positive and Negative Affect Schedule.

and eyeblink rate provide plausibly deniable evidence of sexual arousal.

Discussion

Everyone has thoughts and feelings of which they are not proud. No one wants to admit to feeling frightened when a Black man walks toward them, or aroused when a student walks away. So what would ever compel people to acknowledge such unseemly private reactions? One might expect public evidence to do the trick. People may be tempted to deny having had a racist thought or an inappropriate sexual impulse, but when there is evidence of these mental transgressions, they may be forced to acknowledge them.

And yet, in our studies, precisely the opposite happened. Our participants were less willing to admit to being racially biased or to experiencing inappropriate sexual arousal when they were told that another person had objective evidence of their thoughts and feelings. Interestingly, that evidence made participants less willing to acknowledge their mental transgressions only when they learned about it *before* they had an opportunity to engage in self-reflection, and it had no impact when they learned about it only *after* they had that opportunity. Although we cannot say with certainty what participants were or were not doing during the reflection period, we do know that those who were confronted with evidence beforehand were or were not doing something that demonstrably decreased their subsequent willingness to acknowledge their mental transgressions. Given what psychologists know about the ways in which people typically respond to threats to their reputations and identities (Dweck & Elliott-Moskwa, 2010; Leary et al., 2009; Tesser, 2000), it seems likely that what these participants *were* doing was thinking defensively, and what they were *not* doing was reflecting openly and honestly on their personal shortcomings.

Our studies show that under some circumstances, confronting people with evidence of their biases can lead them to deny those biases. Does that mean that confrontation is bad policy? Not necessarily. Research shows that confronting people about their biases can have a variety of positive consequences, not the least of which is that it can lead people to change their overt behavior. For instance, when White participants are induced to use a stereotypical word to describe a Black person and are then confronted by a confederate ("You should really try to think about Blacks in other ways that are less prejudiced. It just seems that you sound like some kind of racist to me"), they are less likely to use that word again, and less likely to explicitly endorse statements such as "I would rather not have Blacks in the same apartment building that I live in" (Czop, Monteith, & Mark, 2006). Confrontation can be a powerful method for changing the way people behave. But changing the way people think and feel is another story. People who are confronted may change their behavior either because they privately acknowledged their mental transgressions ("I may be a bit racist, so I'll try to do better") or because they privately denied them ("I am definitely not a racist, and I'm going to prove it"). Our studies suggest that if we want people to change their minds rather than simply treading lightly and minding their manners, then it is important to recognize that confronting them with evidence of their mental transgressions can sometimes backfire. It seems that before people are confronted by others, they may need some time to confront themselves.

References

Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2007). A dynamic model of guilt: Implications for motivation and self-regulation in the context of prejudice. *Psychological Science, 18,* 524–530. http://dx.doi.org/10.1111/j.1467-9280.2007.01933.x

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24,* 349–354. http://dx.doi.org/10.1037/h0047358

Czopp, A. M., Monteith, M. J., & Mark, A. Y. (2006). Standing up for a change: Reducing bias through interpersonal confrontation. *Journal of Personality and Social Psychology, 90,* 784–803. http://dx.doi.org/10.1037/0022-3514.90.5.784

Dweck, C. S., & Elliott-Moskwa, E. S. (2010). Self-theories: The roots of defensiveness. *Social Psychological Foundations of Clinical Psychology, 136*–153.

Evans, R. I., Hansen, W. B., & Mittelmark, M. B. (1977). Increasing the validity of self-reports of behavior in a smoking in children investigation. *Journal of Applied Psychology, 62,* 521–523. http://dx.doi.org/10.1037/0021-9010.62.4.521

Feagin, J. R. (1991). The continuing significance of race: Anti-black discrimination in public places. *American Sociological Review, 56,* 101–116. http://dx.doi.org/10.2307/2095676

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102,* 4–27. http://dx.doi.org/10.1037/0033-295X.102.1.4

Henderson, N.-M. (2014, October 28). "Vincent Sheheen accidentally said the word 'whore.' And Ann Romney is not happy about it". *Washington Post*. Retrieved from https://www.washingtonpost.com/news/the-fix/wp/2014/10/28/vincent-sheheen-accidentally-called-nikki-haley-a-whore-ann-romney-is-not-happy-about-it/

Jones, E. E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin, 76,* 349–364. http://dx.doi.org/10.1037/h0031617

Leary, M. R., Terry, M. L., Batts Allen, A., & Tate, E. B. (2009). The concept of ego threat in social and personality psychology: Is ego threat a viable scientific construct? *Personality and Social Psychology Review, 13,* 151–164. http://dx.doi.org/10.1177/1088868309342595

Monteith, M. J. (1993). Self-regulation of prejudiced responses: Implications for progress in prejudice-reduction efforts. *Journal of Personality and Social Psychology, 65,* 469–485. http://dx.doi.org/10.1037/0022-3514.65.3.469

Monteith, M. J., Ashburn-Nardo, L., Voils, C. I., & Czopp, A. M. (2002). Putting the brakes on prejudice: On the development and operation of cues for control. *Journal of Personality and Social Psychology, 83,* 1029–1050. http://dx.doi.org/10.1037/0022-3514.83.5.1029

Moston, S., Stephenson, G. M., & Williamson, T. M. (1992). The effects of case characteristics on suspect behavior during police questioning. *British Journal of Criminology, 32,* 23–40.

Murray, D. M., & Perry, C. L. (1987). The measurement of substance use among adolescents: When is the 'bogus pipeline' method needed? *Addictive Behaviors, 12,* 225–233. http://dx.doi.org/10.1016/0306-4603(87)90032-3

Orwell, G. (1949). *Nineteen eighty-four*. London, UK: Secker & Warburg.

Perillo, J. T., & Kassin, S. M. (2011). Inside interrogation: The lie, the bluff, and false confessions. *Law and Human Behavior, 35,* 327–337. http://dx.doi.org/10.1007/s10979-010-9244-2

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin, 28,* 369–381. http://dx.doi.org/10.1177/0146167202286008

Roese, N. J., & Jamieson, D. W. (1993). Twenty years of bogus pipeline research: A critical review and meta-analysis. *Psychological Bulletin, 114,* 363–375. http://dx.doi.org/10.1037/0033-2909.114.2.363

Ronson, J. (2015). *So you've been publically shamed*. New York, NY: Penguin.

Rudman, L. A., Dohn, M. C., & Fairchild, K. (2007). Implicit self-esteem compensation: Automatic threat defense. *Journal of Personality and Social Psychology, 93,* 798–813. http://dx.doi.org/10.1037/0022-3514.93.5.798

Swim, J. K., & Hyers, L. L. (1999). Excuse me—What did you just say?!: Women's public and private responses to sexist remarks. *Journal of Experimental Social Psychology, 35,* 68–88. http://dx.doi.org/10.1006/jesp.1998.1370

Swim, J. K., Hyers, L. L., Cohen, L. L., Fitzgerald, D. C., & Bylsma, W. H. (2003). African American college students' experiences with everyday racism: Characteristics of and responses to these incidents. *Journal of Black Psychology, 29,* 38–67. http://dx.doi.org/10.1177/0095798402239228

Tamaki, J. (February 14, 2001). "Bustamante voices regret for racial slur". *Los Angeles Times*. http://articles.latimes.com/2001/feb/14/news/mn-25173.

Tesser, A. (2000). On the confluence of self-esteem maintenance mechanisms. *Personality and Social Psychology Review, 4,* 290–299. http://dx.doi.org/10.1207/S15327957PSPR0404_1

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54,* 1063–1070. http://dx.doi.org/10.1037/0022-3514.54.6.1063

Woodlee, Y. (February 4, 1999). "D. C. mayor acted 'hastily,' will rehire aide". *Washington Post*. http://www.washingtonpost.com/wp-srv/local/longterm/williams/williams020499.htm

Received October 27, 2015
Revision received February 10, 2016
Accepted March 20, 2016 ■