Evolution and Human Behavior 38 (2017) 561–571

Contents lists available at ScienceDirect

Evolution and Human Behavior

journal homepage: www.ehbonline.org

Original Article

Effect of artificial surveillance cues on reported moral judgment: Experimental failures to replicate and two meta-analyses

Stefanie B. Northover [a,b,*], William C. Pedersen [c], Adam B. Cohen [b], Paul W. Andrews [a]

[a] Department of Psychology, Neuroscience & Behaviour, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada
[b] Department of Psychology, Arizona State University, 651 E. University Drive, P.O. Box 871104, Tempe, AZ 85287, USA
[c] Department of Psychology, California State University, Long Beach, 1250 Bellflower Boulevard, Long Beach, CA 90840, USA

ARTICLE INFO

Article history:
Initial receipt 3 August 2016
Final revision received 20 December 2016

Keywords:
Surveillance cues
Cues of being watched
Observation cues
Eyespots
Moral judgment
Meta-analysis

ABSTRACT

Several papers have reported that artificial surveillance cues, such as images of watching eyes, cause anonymous participants to behave as if they are actually under surveillance, thus increasing moral behavior. In a series of four experiments, we found no evidence that artificial surveillance cues impact reported moral judgment, self-rated possession of positive traits, or religiosity. Two small meta-analyses, both comprising six experiments investigating the effect of artificial surveillance cues on moral judgment, provided mixed conclusions. One meta-analysis produced a mean effect size not significantly different from zero and the other produced a mean effect size on the edge of significance. On the whole, artificial surveillance cues have inconsistent effects, or possibly no effect, on moral outcomes.

1. Introduction

When people are under surveillance, they tend to behave more prosocially than they otherwise would (Kurzban, 2001; Kurzban, DeScioli, & O'Brien, 2007; Piazza & Bering, 2008; Satow, 1975; van Rompay, Vonk, & Fransen, 2009). Even artificial cues of surveillance, such as stylized images of eyes, have apparently increased prosocial behavior in lab and field experiments (e.g., Haley & Fessler, 2005; Pfattheicher & Keller, 2015). Participants seemingly behave like they are being watched when they are exposed to artificial cues of being watched, even though participants are consciously aware that they are not actually being watched. Outcomes in such experiments have included donating to charity (Pfattheicher, 2015), hand washing (Carbon & Hesslinger, 2011), and picking up litter (Ernest-Jones, Nettle, & Bateson, 2011).

However, surveillance cue effects sometimes fail to replicate (Carbon & Hesslinger, 2011; Matsugasaki, Tsukamoto, & Ohtsubo, 2015). Two meta-analyses investigating the impact of artificial surveillance cues on generosity produced small mean effect sizes that were not significantly different from zero (Northover, Pedersen, Cohen, & Andrews, 2017). Although artificial surveillance cues may not impact generosity, more work should be done to investigate additional behavioral outcomes. The goal of the present paper is to investigate the effect of artificial surveillance cues on moral judgment, an outcome sufficiently different from generosity to warrant separate consideration.

Our primary measure of moral judgment is that used by Bourrat, Baumard, and McKay (2011), who asked participants to rate the moral acceptability of two misdeeds: returning a lost wallet but keeping the money, and falsifying information on a résumé (Schnall, Haidt, Clore, & Jordan, 2008). Participants who were exposed to an image of watching eyes rated the transgressions more harshly than participants exposed to an image of flowers, suggesting that the artificial surveillance cue caused the participants to respond like they were truly under surveillance.

In many cases, reported surveillance cue effects are seemingly conditional on features of the environment, qualities of the surveillance cue, participant traits, or methods of data analysis (Northover et al., 2017). Although many moderating variables have been proposed, findings are inconsistent. One potential moderator is the masculinity or femininity of the surveillance cue. In a field experiment conducted by Bateson et al. (2006), anonymous people contributed more money to an honesty box, used to collect funds for coffee, when masculine eyes were displayed than when feminine eyes were displayed (but see Carbon & Hesslinger, 2011). Matland and Murray (2015) also found a greater

* Corresponding author at: Department of Psychology, Neuroscience & Behaviour, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada.
E-mail addresses: stefanie.northover@asu.edu (S.B. Northover), bill.pedersen@csulb.edu (W.C. Pedersen), adamcohen@asu.edu (A.B. Cohen), pandrew@mcmaster.ca (P.W. Andrews).

http://dx.doi.org/10.1016/j.evolhumbehav.2016.12.003
1090-5138/

effect from masculine eyes. However, other studies found no significant difference between masculine and feminine eye cues (Nettle et al., 2013; Panagopoulos, 2014).

Another potential moderator is the familiarity of the surveillance cue. A familiar face may induce the feeling of being watched by a member of the community. People are more likely to behave prosocially in less densely-populated areas (Korte & Kerr, 1975; Levine, Martinez, Brase, & Sorenson, 1994; Rushton, 1978; Yousif & Korte, 1995). This may be because the less dense the population of a community, the easier it is to build and maintain a reputation. Therefore, people should behave more prosocially when they are observed by familiar individuals than they do when they are observed by unfamiliar individuals.

We conducted four experiments investigating artificial surveillance cue effects. Initially, we were interested in multiple dependent variables – religiosity, positive traits, and moral judgment (Experiment 1). However, we turned our attention to a single dependent variable – moral judgment – when we were unable to conceptually replicate Bourrat et al.'s (2011) moral judgment results. Experiments 2, 3, and 4 were increasingly precise replications of Bourrat and colleagues. We also investigated the femininity/masculinity and the familiarity of the surveillance cues as possible moderators. In addition, we evaluated several moderating variables in post hoc fashion after multiple experiments failed to replicate the findings of Bourrat et al. These variables included the length of surveillance cue exposure, the location of the surveillance cue, whether the experimenters drew attention to the surveillance cue, and the location of the experiment. None of our experiments resulted in significant surveillance cue effects.[1] In addition to our experiments, we conducted small meta-analyses of the six studies which investigated the effect of surveillance cues on Bourrat and colleagues' moral judgment task.

2. Experiment 1

2.1. Surveillance cue traits

In addition to exploring effects of surveillance cues generally, we investigated different attributes of surveillance cues: familiarity and masculinity/femininity.

2.2. Dependent measures

In Experiment 1, we investigated two dependent measures in addition to moral judgment: self-rated possession of positive traits and religiosity. If surveillance increases the likelihood of reputation-boosting behavior, then any traits that are desirable in social exchanges may be displayed or exaggerated. Thus, a watched individual may behave in a way that implies the possession of positive traits such as kindness, honesty, generosity, or reliability.

Additionally, people who are being watched may wish to appear religious. Religion tends to be associated with morality and trustworthiness (Edgell, Gerteis, & Hartmann, 2006; Farkas, Johnson, Foleno, Duffett, & Foley, 2001; Hall, Cohen, Meyer, Varley & Brewer, 2015; Tan & Vogel, 2008), whereas atheists tend to be viewed as untrustworthy (Gervais, Shariff, & Norenzayan, 2011) and incite negative feelings in others (Pew Research Center, 2014). Although the Canadian province of Ontario, from which our sample came, is not a particularly religious region (23.14% of people claimed no religious affiliation in a 2011 census; Statistics Canada, 2013), a meta-analysis conducted by Sedikides and Gebauer (2010) showed a significant positive correlation between intrinsic religiosity and socially desirable responding among Canadians. The authors proposed that this relationship exists because religiosity can be used by people to self-enhance. If the authors are correct, their findings suggest that religiosity is valued by Canadian culture. Therefore, Canadian participants may exaggerate their religiosity when they feel like they are being watched.

2.3. Method

2.3.1. Participants

We recruited 338 psychology students from McMaster University, located in southern Ontario. Participants were given course credit for their participation. The mean age of the participants was 19.1 years; there were 83 men, 253 women, and 2 of unreported gender; about 50% were White, 40% Asian, 6.5% Middle Eastern, and 5% indicated some other ethnicity.

2.3.2. Procedure

Each participant was seated alone in a small room with the door closed, isolated from other people to provide for anonymity and privacy. The participants' task was to complete a computer questionnaire made up of three parts designed to measure religiosity, self-rated possession of positive traits, and moral judgment. The computer screen was split into two frames. The left frame contained the questionnaire, which was administered through LimeSurvey (www.limesurvey.org). The contents of the right frame depended on which of four conditions the participant had been randomly assigned to – familiar face, unfamiliar face, chair (an image control condition), or no image (blank screen). For the familiar face condition, the image was of a celebrity's face. For the unfamiliar face condition, the image was of the face of a person who was not well known in North America. For the chair condition, the image was of a chair on a white background.

The cover story told to the participants was, "We're studying simultaneous processing of various types of visual stimuli. All conditions will have words. Some conditions will also have images. Some conditions will *not* have images. At the end of the experiment, you'll be asked questions about any images you see if you have them, so *please pay careful attention to them*."

To ensure the experimenters were blind to condition, the experimenters clicked a button on the computer screen as soon as they were finished giving directions to each participant. Immediately after clicking the button, the experimenters left the experiment room. Clicking the button started a ten second countdown, then the right frame loaded either an image (familiar face, unfamiliar face, or chair conditions) or a blank page (no image condition).

At the end of the experiment, participants were probed for suspicion. Data were removed for those who correctly guessed the purpose of the experiment.[2]

2.3.3. Stimuli

Participants in the familiar face, unfamiliar face, and chair conditions were presented with images. Six different images were used for each of these conditions. Each participant in these three conditions was shown just one of the images. We used a monitor with a viewable image size of 59.69 cm and a screen resolution of 1920 by 1080 pixels.

The individuals chosen for the familiar face condition were Kristen Stewart, Rihanna, Taylor Swift, Barack Obama, Danielle Radcliffe, and Tom Hanks. The individuals selected for the unfamiliar face condition were mostly models or celebrities from outside North America, chosen because their images were similar in style and attractiveness to those in the familiar face condition. We attempted to match the familiar and unfamiliar faces on gender, approximate age, and ethnicity. Half of the faces were male and half were female, so we were able to investigate the dependent measures according to the masculinity/femininity of the surveillance cues. Each face image had an interpupillary distance of 115 or 116 pixels. All faces were aligned so there was no head tilt.

[1] Data from all four experiments are available at the first author's website.

[2] Unfortunately, we do not know the exact number of participants whose data were removed, as these records were lost; our best estimate is 4 or 5. These data were removed before any data analysis.

They were all looking at the camera straight-on or nearly so, which made them appear to look at the participants. The mean area of the face images was 170,871 square pixels. See Fig. 1 for examples of familiar and unfamiliar faces.

The chair images were obtained from the internet and were chosen because we judged them unlikely to elicit any emotions. The mean area of the chair images was 192,134 square pixels. See Fig. 2 for an example of a chair image.

2.3.4. Instruments

The survey was made up of different sections in counterbalanced order: religiosity, possession of positive traits, and moral judgment.

For the religiosity section, participants were asked to indicate their agreement with the statements "I believe in God", "We'd be better off if religion played a bigger role in people's lives", and "Religious beliefs are important to me in my everyday decisions" (Li, Cohen, Weeden, & Kenrick, 2010). Ratings, which varied from 1 (*very strongly disagree*) to 9 (*very strongly agree*), were summed to create a religiosity score which could range from 3 (least religiosity) to 27 (most religiosity). Cronbach's alpha was 0.89, indicating very good reliability.

For the positive traits section, participants used the same rating scale to indicate how strongly they agreed that they possessed certain positive traits (kind, competent, attractive, brave, generous, and intelligent), as well as certain negative traits (dishonest, unreliable, weak, and insecure). After reverse-scoring the negative traits, scores for all ten statements were summed to create a positive traits score, which could range from 10 (lowest possible rating of positive traits) to 90 (greatest possible rating). Cronbach's alpha was 0.73, indicating acceptable reliability.

The moral acceptability task was modeled after Bourrat et al.'s (2011) study. Bourrat and colleagues' participants read two vignettes which were originally published by Schnall et al. (2008). In Experiment 1, our participants read the same vignettes. One of the vignettes ("wallet") read as follows: "You are walking down the street when you come

Fig. 2. Example of image used for the chair condition in Experiment 1 and Experiment 2.

Fig. 1. Examples of images used for the familiar face and unfamiliar face conditions in Experiment 1. On the left is a familiar face and on the right is an unfamiliar face, matched for approximate age, gender, and ethnicity.

across a wallet lying on the ground. You open the wallet and find that it contains several hundred dollars in cash as well the owner's driver's license. From the credit cards and other items in the wallet it's very clear that the wallet's owner is wealthy. You, on the other hand, have been hit by hard times recently and could really use some extra money. You consider sending the wallet back to the owner without the cash, keeping the cash for yourself. How wrong is it for you to keep the money you found in the wallet in order to have more money for yourself?"

The "résumé" vignette read: "You have a friend who has been trying to find a job lately without much success. He figured that he would be more likely to get hired if he had a more impressive resume. He decided to put some false information on his resume in order to make it more impressive. By doing this he ultimately managed to get hired, beating out several candidates who were actually more qualified than he. How wrong was it for your friend to put false information on his resume in order to help him find employment?"

The participants rated the moral acceptability of each of the vignettes on a 9-point scale, with 1 = *morally unacceptable* and 9 = *morally acceptable*.

2.4. Results

2.4.1. Manipulation check on familiarity

At the end of the experiment, participants in the familiar face and unfamiliar face conditions were asked to rate the familiarity of the face in their images on a 4-point scale, from 1 (*not familiar at all*) to 4 (*very familiar*). Participants in the familiar face condition rated their faces ($Mdn = 4$) significantly more familiar than participants in the unfamiliar face condition ($Mdn = 1$), $\chi^2(3, N = 169) = 131$, $p < 0.001$, Cramér's $V = 0.88$.

2.4.2. Effect of familiar and unfamiliar faces

Kolmogorov-Smirnov tests indicated significant deviation from normality for all four dependent measures (all $ps < 0.001$). Therefore, Kruskal-Wallis (non-parametric) tests were conducted with the image condition (familiar face, unfamiliar face, chair, and no image) as the independent variable. There were no statistically significant differences in religiosity between the conditions, $H(3, N = 336) = 0.85$, $p = 0.84$. No significant differences were found in moral acceptability[3] for either the wallet vignette ($H(3, N = 338) = 5.00$, $p = 0.17$) or the résumé vignette ($H(3, N = 338) = 3.93$, $p = 0.27$). The Kruskal-Wallis test was marginally significant for positive traits, $H(3, N = 329) = 7.35$, $p = 0.062$, but the results indicated no meaningful pattern: The chair and unfamiliar face conditions had the highest medians ($Mdn = 67$), followed by the familiar face condition ($Mdn = 65.5$) and finally the no image condition ($Mdn = 65$). Descriptive statistics for religiosity and positive traits, broken down by conditions, are provided in Table 1. Descriptive statistics for the moral acceptability ratings are provided in Table 2.

[3] Although not relevant to the purpose of the experiment, it is interesting to note that men reported greater levels of moral acceptability, on average, than women. A two-tailed Mann-Whitney *U* Test revealed the difference between men and women to be significant for the wallet vignette, $U = 8154$, $N_{men} = 83$, $N_{women} = 253$, $p = 0.001$. The difference between men and women was not significant for the résumé vignette, $U = 9857$, $N_{men} = 83$, $N_{women} = 253$, $p = 0.39$, but the data showed the same pattern as the wallet vignette ratings, with higher means and medians for men. These data suggest that men are more morally lenient than women for these particular scenarios.
To make sure this sex difference did not affect the results of our independent variables, we conducted a Chi square test of independence to see if men and women were assigned to the different conditions in similar proportions. The proportions of men and women who were assigned to the no image, chair, familiar face, and unfamiliar face conditions were not statistically different, $\chi^2(3, N = 336) = 3.67$, $p = 0.30$, Cramér's $V = 0.10$. Likewise, the proportions of men and women who were assigned to the no image, chair, female face, and male face conditions were not statistically different, $\chi^2(3, N = 336) = 1.83$, $p = 0.61$, Cramér's $V = 0.07$.

Table 1
Experiment 1: positive traits and religiosity ratings for the familiar face, unfamiliar face, chair, no image, female face, and male face conditions.

	Positive traits					Religiosity				
Condition	*M*	*Mdn*	Mode	*SD*	*n*	*M*	*Mdn*	Mode	*SD*	*n*
Familiar face	65.35	65.50	64	7.71	80	15.19	16.00	9	6.83	84
Unfamiliar face	66.32	67.00	64	8.10	82	15.68	17.00	18	7.28	85
Chair	67.26	67.00	66	7.13	84	14.98	16.00	19	7.24	84
No image	63.64	65.00	68	8.56	83	14.86	15.00	13,15	6.42	83
Female face	65.80	66.00	66	7.83	79	15.05	16.00	17,18	7.04	85
Male face	65.88	66.00	64	8.01	83	15.83	17.00	19	7.07	84

2.4.3. Effect of masculine and feminine faces

To determine if masculine surveillance cues had a different effect than feminine surveillance cues, Kruskal-Wallis tests were conducted with male face, female face, chair, and no image groups. No significant differences were found in religiosity ($H(3, N = 336) = 1.27$, $p = 0.74$). The differences between conditions were again marginally significant for positive traits ($H(3, N = 329) = 6.31$, $p = 0.098$). No significant differences were found in moral acceptability ratings for the résumé vignette ($H(3, N = 338) = 3.78$, $p = 0.29$), but the differences were marginally significant for the wallet vignette, $H(3, N = 338) = 7.69$, $p = 0.053$. Investigating this further, we found that for both vignettes, women who were presented with an image of a male face reported lower moral acceptability scores than women presented with an image of a female face, an image of a chair, or no image at all. We conducted two Kruskal-Wallis tests using only data from the women; the moral acceptability rating was the dependent variable and the male face, female face, chair, and no image groups were compared. The differences were still nonsignificant for therésumé vignette ($H(3, N = 253) = 5.76$, $p = 0.12$), but the differences were significant for the wallet vignette ($H(3, N = 253) = 12.32$, $p = 0.006$). Table 1 displays descriptive statistics for religiosity and positive traits. Descriptive statistics for the moral acceptability ratings are provided in Table 2.

2.4.4. Surveillance cue duration

Sparks and Barclay (2013) hypothesized that images of eyes have the greatest effect on behavior when the eyes appear briefly. Therefore, we investigated the possibility that self-reported religiosity, positive traits, or moral judgment was affected by briefly-appearing surveillance cues. The order in which the participants completed the tasks (i.e., moral judgment, religiosity, and positive traits) affected the amount of time the participants were exposed to the images before completing any given task. About one-third of the participants completed the moral judgment task first, about one-third completed the religiosity task first, and about one-third completed the positive traits task first. As explained previously, images were loaded 10 s after experimenters finished giving instructions to participants. Ten seconds was about the time it took for participants to load the first questions and read the directions to their first task, so exposure to images began about the same time participants began their first task. Therefore, participants were exposed to images only briefly before beginning their first task.

Table 2
Experiment 1: moral acceptability ratings for the familiar face, unfamiliar face, chair, no image, female face, and male face conditions.

	Wallet vignette					Résumé vignette				
Condition	*M*	*Mdn*	Mode	*SD*	*n*	*M*	*Mdn*	Mode	*SD*	*n*
Familiar face	2.43	2.00	1	1.87	84	2.60	2.00	1,2,3	1.61	84
Unfamiliar face	2.25	1.00	1	1.91	85	2.69	2.00	1	2.06	85
Chair	2.26	2.00	1	1.64	84	2.95	3.00	2	1.69	84
No image	2.65	2.00	1	1.78	85	2.75	2.00	1	2.04	85
Female face	2.54	2.00	1	1.90	85	2.60	2.00	1,3	1.75	85
Male face	2.13	1.00	1	1.86	84	2.69	2.00	1	1.95	84

Kruskal-Wallis tests were conducted examining positive traits using only data from participants for whom the positive traits questionnaire was the first task (i.e., participants who were exposed to their image only briefly before starting the positive traits task). One of the tests compared the familiar face, unfamiliar face, chair, and no image groups; there was a marginally significant difference, as there was when the task order was not taken into account ($H(3, N = 108) = 6.98$, $p = 0.073$). The other test compared the female face, male face, chair, and no image conditions and found no significant differences ($H(3, N = 108) = 3.64$, $p = 0.30$).

Next, Kruskal-Wallis tests were conducted examining religiosity using only data from participants who completed the religiosity task before the other two tasks. One of the tests compared the familiar face, unfamiliar face, chair, and no image groups, and found no significant differences ($H(3, N = 112) = 0.99$, $p = 0.80$). The other test compared the female face, male face, chair, and no image conditions and also found no significant differences ($H(3, N = 112) = 0.96$, $p = 0.81$).

Finally, Kruskal-Wallis tests were conducted examining moral judgment using data only from participants who completed the moral judgment task before completing the religiosity and positive traits tasks. Once again, no differences were found when comparing the familiar face, unfamiliar face, chair, and no image groups ($H(3, N = 113) = 0.27$, $p = 0.97$ for the wallet vignette; $H(3, N = 113) = 3.53$, $p = 0.32$ for the résumé vignette). There were also no significant differences when comparing the female face, male face, chair, and no image conditions ($H(3, N = 113) = 2.18$, $p = 0.54$ for the wallet vignette; $H(3, N = 113) = 3.13$, $p = 0.37$ the résumé vignette).

2.5. Discussion

Cues of being watched did not seem to affect self-reported religiosity or positive traits, regardless of the familiarity or gender of the face used for the surveillance cue. Results were less clear for moral judgment. When analyzing data only from women, moral acceptability ratings of the wallet vignette were lower for those presented with a male face than they were for the other groups. Before drawing any conclusions about this finding, we wanted to replicate it. We had not predicted this outcome in advance, and the probability of obtaining a false positive result of some kind was high; in Experiment 1, we conducted several statistical analyses without correcting our alpha level, thus inflating the probability of a Type I error.

In Experiment 1, we used images of celebrities to test the effect of cues of being watched by a familiar person. Although these images were rated as significantly more familiar than the images in the unfamiliar face condition, images of the participants' acquaintances may have been more appropriate to use as familiar faces (Gobbini, Leibenluft, Santiago, & Haxby, 2004). However, our less than optimal choice of familiar faces should not have affected our ability to test the more general hypothesis that artificial cues of being watched increase prosocial behavior. When combining the unfamiliar face condition with the familiar face condition to create a single face condition, and comparing that with a single control condition comprised of the chair condition and no image condition, there were no significant differences for any of the dependent measures (all $ps > 0.14$).

In greatest need of explanation is why the moral acceptability ratings from the surveillance cue groups did not differ from those of the control groups. Bourrat et al. (2011) found significant differences using the same vignettes and rating scales we did. There were, however, some differences between our studies. Perhaps one of those differences was the cause of our disparate results.

First, Bourrat et al. (2011) placed their surveillance cue directly above the rating scale. In Experiment 1, our surveillance cue was located on the right-hand side of a computer monitor, several inches away from the rating scale. Our surveillance cue may have been too far away from the participants' line of sight to influence their responses.

Second, Bourrat et al. (2011) did not give their participants any information about the surveillance cue on their paper. In Experiment 1, participants were told that they might see images, and, if so, to pay careful attention to them. Perhaps any feelings of being watched induced by our surveillance cues were attributed to the surveillance cues and thus discounted. (For a discussion of the role of attribution in memory, see Jacoby, Kelley, & Dywan, 1989.)

We also considered the length of surveillance cue exposure. Sparks and Barclay (2013) found that effects were more likely to appear when the surveillance cue was revealed only briefly, possibly because people habituate to surveillance cues if they are visible for too long. We found no effects among our relatively briefly-exposed participants. Therefore, we do not believe that the duration of surveillance cue visibility explains why Bourrat et al. obtained significant results and we did not. However, perhaps our surveillance cues were not brief enough.

3. Experiment 2

To address the issues discussed above, we made three changes in Experiment 2 relative to Experiment 1. First, we reduced the length of surveillance cue exposure. Second, we placed the surveillance cue directly above the rating scale. Third, we attempted to reduce the participants' attention to the surveillance cue by providing no explanation for it. These changes made Experiment 2 more like Bourrat et al.'s (2011) study.

Experiment 2 further investigated the seeming interaction between participant sex and surveillance gender found in Experiment 1. In Experiment 1, women indicated harsher judgment of moral transgressions when they were exposed to cues of being watched by men. If this is a robust effect, it might be related to cues of being watched by potential mates. We further reasoned that if this account was correct, cues of being watched by men would have their greatest effect on women's judgment of behavior that men find particularly unattractive in a mate, such as infidelity (Buss & Schmitt, 1993). Therefore, Experiment 2 included a third vignette about infidelity. Experiment 2 also considered the possibility that the attractiveness of the (male) surveillance cue would moderate this effect, as women may try to appeal to attractive men but less so to unattractive men.

3.1. Method

3.1.1. Participants

We recruited 612 participants online via Amazon Mechanical Turk. Each participant was paid 20 cents. The data from 12 participants were removed (see explanation below), leaving 600 participants total for analyses. The mean age of the participants was 31.8 years; 236 were women, 328 were men, and 36 did not report their sex; about 56% were Asian, 29% were White, and 15% reported some other ethnicity.

Table 3
Experiment 2: moral acceptability ratings for the female face, male face, and chair conditions.

	Wallet vignette					Résumé vignette					Infidelity vignette				
Condition	*M*	*Mdn*	Mode	*SD*	*n*	*M*	*Mdn*	Mode	*SD*	*n*	*M*	*Mdn*	Mode	*SD*	*n*
Female face	2.83	2.00	1	2.25	187	2.99	3.00	1	2.07	180	1.84	1.00	1	1.63	181
Male face	2.77	2.00	1	2.32	203	2.87	2.00	1	1.98	202	2.03	1.00	1	1.83	204
Chair	2.83	2.00	1	2.28	183	3.03	2.00	1	2.14	187	1.72	1.00	1	1.47	187

Table 4
Experiment 2: women's moral acceptability ratings for the chair condition, the female face condition, and the male face condition broken down by attractiveness ratings.

	Résumé vignette			Wallet vignette			Infidelity vignette		
Condition	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*
Chair	2.79	2.17	76	2.43	2.12	77	1.60	1.44	78
Female face	2.64	1.96	77	2.50	2.16	80	1.73	1.54	78
Low attractiveness male face	2.47	1.83	47	2.10	1.79	48	1.96	1.82	47
High attractiveness male face	2.88	1.90	25	2.81	2.38	26	1.85	1.64	26

3.1.2. Procedure

Experiment 2 consisted of just the moral judgment task, although a third vignette (infidelity) was added. Once again, we administered the moral judgment task with LimeSurvey (www.limesurvey.org). This time, however, the participants completed the task on their own computers.

Each participant read and rated the vignettes in randomized order. After each of the three vignettes was displayed, the participants were asked to confirm that they had read the passage. Once the participants did so, an image appeared on the screen directly above the scale used to rate the moral acceptability of the behavior described in the passage.

Once the participants rated a vignette, they went on to a new page for the next vignette and the image disappeared until the participants indicated that they had read the next passage. In this way, exposure to the image was probably quite brief for most participants, and certainly briefer than it was for participants in Experiment 1, including those who received the briefest possible exposure by completing the moral judgment task first.

3.1.3. Stimuli

The images fell into two categories: faces and chairs. Fifteen different images of chairs and 30 different images of faces were used. Each participant was randomly assigned and shown only one image. The chairs met the same criteria as in Experiment 1. The mean area of the chair images was 144,002 square pixels.

Fifteen of the faces were female and 15 were male. Some of the people in the images were famous; most were not. Ethnicity was uncertain for most of the individuals, because their identity was unknown. We estimate that one-third were Asian, one-third were White, and one-sixth were Black; we are unsure about the rest. Because ethnicity was unrelated to our hypotheses, we simply strove for an ethnically diverse group of individuals. The faces varied widely in attractiveness.[4] Each face had an interpupillary distance of 109 or 110 pixels. All faces were aligned so there was no head tilt. They were all looking at the camera straight-on or nearly so, which made them appear to be looking at the participants. The mean area of the face images was 134,426 square pixels.

3.1.4. Instruments

The added vignette read as follows: "A young woman has been in a romantic relationship with her boyfriend for about two years now. She likes her boyfriend but occasionally she has sex with other men. Her boyfriend is not aware of this and believes she has been faithful to him. How wrong is it for the woman to secretly have sex outside of her relationship?" As before, participants rated the moral acceptability of the vignettes from 1 = *morally unacceptable* to 9 = *morally acceptable*. Furthermore, participants who were exposed to images of faces were asked to rate the attractiveness of the faces from 1 = *extremely unattractive* to 9 = *extremely attractive*.

3.2. Results

3.2.1. Manipulation check

We included the following question at the end of the experiment to weed out data from anyone who had trouble loading images or was not paying attention: "You should have seen an image appear multiple times during the survey. What kind of image did you see?" The answers, which were displayed in random order, were "A chair", "A face", or "I didn't see any images". Twelve participants were removed from analyses because they indicated that they had not seen any images or they chose the wrong kind of image (e.g., they chose "A face" when they were shown an image of a chair). Out of 612 participants, 600 correctly identified the kind of image displayed.

3.2.2. Surveillance cue gender effect on moral judgment

Kolmogorov-Smirnov tests indicated that the distribution of moral acceptability scores deviated significantly from normality for all three vignettes (all $ps < 0.001$). Therefore, we conducted nonparametric Kruskal-Wallis tests comparing moral judgment scores from the male face, female face, and chair conditions. There were no significant differences for the wallet vignette ($H(2, N = 573) = 0.25, p = 0.88$), the résumé vignette ($H(2, N = 569) = 0.38, p = 0.83$), or the infidelity vignette ($H(2, N = 572) = 1.41, p = 0.49$).[5] Descriptive statistics are displayed in Table 3.

3.2.3. Interaction between surveillance cue gender and participant sex

Kruskal-Wallis tests comparing the male face, female face, and chair conditions were again conducted, this time only including data from women. These tests found no significant differences for the wallet vignette ($H(2, N = 234) = 0.091, p = 0.96$), the résumé vignette ($H(2, N = 228) = 0.14, p = 0.93$), or the infidelity vignette ($H(2, N = 232) = 0.97, p = 0.62$). In other words, there were no significant differences in moral acceptability scores between women who were shown a male face and the rest of the women; therefore, the interaction from Experiment 1 did not replicate.

An effect of watching male faces could be moderated by the attractiveness of the watching face, so we investigated the possibility that women's moral acceptability ratings were differently affected by attractive and unattractive male faces. The skew of the data[6] precluded parametric tests, so we simply inspected the moral acceptability ratings given by women in the chair condition, the female face condition, and the male face condition. We split the male face condition into two conditions – low attractiveness male face and high attractiveness male face – consisting of women who rated the male face in their surveillance cue image as low (1–5) or high (6–9) in attractiveness. If women judge moral transgressions more harshly when exposed to cues of highly attractive watching men, moral acceptability ratings from women exposed to high attractiveness male faces should be lower than ratings from women exposed to images of low attractiveness male faces, female faces, and chairs. Inspection of ratings revealed the opposite pattern: Mean moral acceptability ratings for the résumé and wallet

[4] Participants who were exposed to images of faces were asked to rate the attractiveness of the faces on a 9-point scale (from 1 = *extremely unattractive* to 9 = *extremely attractive*). The attractiveness ratings created a bimodal distribution, with 14.3 and 13.7% of participants rating their image as a 5 and a 7, respectively, and all other ratings being chosen by 3 to 8.2% of participants.

[5] In Experiment 1, men rated the moral violations more leniently than women. This was also the case in Experiment 2; the men's mean ratings were higher for all three vignettes and their median ratings were higher for the wallet and résumé vignettes. Two-tailed Mann-Whitney *U* Tests revealed the difference between men and women to be significant for the wallet vignette, $U = 31.542, N_{men} = 319, N_{women} = 234, p = 0.001$ and for the résumé vignette, $U = 32.023, N_{men} = 322, N_{women} = 228, p = 0.009$, and marginally significant for the infidelity vignette, $U = 34.814, N_{men} = 323, N_{women} = 232, p = 0.09$. However, the proportions of men and women who were assigned to the male face, female face, and chair conditions were not significantly different, $\chi^2(2, N = 564) = 1.35, p = 0.51$, Cramér's $V = 0.05$.

[6] The infidelity vignette distribution had a skewness of 2.23 ($SE = 0.10$), the wallet vignette distribution had a skewness of 1.13 ($SE = 0.10$), and the résumé vignette distribution had a skewness of 0.87 ($SE = 0.10$).

Fig. 3. Images used for the surveillance cue and control conditions in Experiments 3 and 4. These images were initially used by Bourrat et al. (2011).

transgressions were *highest* from women exposed to high attractiveness male faces, and for the infidelity transgression, they were second highest. In other words, exposure to images of highly attractive male faces did not reduce women's reported moral acceptability in Experiment 2. Descriptive statistics are available in Table 4.

3.3. Discussion

In Experiment 2, we found no differences in moral acceptability ratings between conditions, despite our changes in design to more closely replicate Bourrat et al. (2011). Therefore, we do not believe the surveillance cue's location nor the attention drawn to it can explain the differences between Bourrat et al.'s results and ours. In light of the findings of Sparks and Barclay (2013) that surveillance cues are more likely to affect behavior when they are presented briefly, we reduced presentation time in Experiment 2. We still found no surveillance cue effects; furthermore, in an experiment conducted after we conducted ours, Sparks and Barclay (2015) found no effect of either long or short duration surveillance cues on the moral judgment task. Therefore, cue presentation duration is an unlikely explanation for our null results as well.

In Experiment 1, we found a difference in moral acceptability ratings between women presented with an image of a male face and women presented with an image of a female face, an image of a chair, or no image. This result did not replicate in Experiment 2. We therefore believe the original finding was a false positive. We also found no main effect for the apparent gender of the person in the surveillance cue images.

Experiment 2 was more like Bourrat et al.'s (2011) study than Experiment 1, yet we still obtained null results. However, perhaps there was some key difference not yet explored between their study and ours that brought out a surveillance cue effect in the former. For Experiment 3, we tried an even closer replication.

4. Experiment 3

The methods for Experiment 3 resembled more closely those used by Bourrat et al. (2011), making Experiment 3 a truer replication than our first two experiments. We used Bourrat et al.'s stimuli images, administered the task on paper as they did, and forwent the collection of demographic data as they did.

4.1. Method

4.1.1. Participants

We recruited psychology students from McMaster University, resulting in a sample size of 93, similar to Bourrat et al.'s (2011) sample size of 91. Subjects received course credit for their participation. In order to increase feelings of anonymity, Bourrat et al. did not collect demographic data; therefore, we did not either.

Table 5
Experiment 3: moral acceptability ratings for the control and surveillance cue conditions.

	Wallet vignette					Résumé vignette				
Condition	*M*	*Mdn*	Mode	*SD*	*n*	*M*	*Mdn*	Mode	*SD*	*n*
Control	1.96	2.00	1	1.20	47	3.01	3.00	2	1.76	47
Surveillance	2.46	2.00	1	1.63	46	2.63	2.00	1	1.77	46

4.1.2. Procedure

The procedure for Experiment 3 was similar to that of Bourrat et al. (2011). In both studies, participants completed the moral judgment task on a piece of paper with the wallet vignette printed on one side and the résumé vignette on the other. Because we were attempting a high-fidelity replication of Bourrat et al., we included only the wallet and résumé vignettes they used (and thus excluded the infidelity vignette from Experiment 2).

Half of the participants had an image of watching eyes above the Likert scale, and half had an image of flowers (the surveillance cue and control conditions, respectively; see Figure 3). We used the same images Bourrat et al. used and scaled them to the same size they did (47×17 mm). Participants circled the number of their choice on the rating scale with a pen.

Experiment 3 was meant to replicate the procedure described by Bourrat et al. (2011), but there were some differences. Bourrat et al.'s participants completed the moral judgment task in university libraries. Some of the participants were by themselves, whereas others were not (P. Bourrat, personal communication, May 3, 2013). Our participants completed their task in a laboratory, however, and were always isolated in their own rooms. Furthermore, Bourrat and colleagues' participants did not receive compensation, whereas our participants received course credit.

4.2. Results

Because we compared only two groups for Experiment 3, we analyzed our data the same way Bourrat et al. (2011) analyzed theirs – with Mann-Whitney *U* tests. Descriptive statistics for the moral acceptability ratings, broken down by condition, are presented in Table 5.

For the résumé vignette, the median moral acceptability rating was lower for the surveillance cue condition ($Mdn = 2$) than it was for the control condition ($Mdn = 3$), but moral acceptability ratings were not significantly different ($U = 922$, $N_{\text{surveillance}} = 46$, $N_{\text{control}} = 47$, $p = 0.21$).[7]

For the wallet vignette, the medians were the same for both conditions ($Mdn = 2$), but moral acceptability scores were marginally greater for the surveillance cue condition ($U = 877$, $N_{\text{surveillance}} = 46$, $N_{\text{control}} = 47$, $p = 0.10$). This difference is in the opposite direction of that obtained by Bourrat et al. (2011).

4.3. Discussion

There were no significant differences in moral acceptability ratings of the résumé vignette between the surveillance cue group and the control group. We did find a marginally significant difference in ratings of the wallet vignette, but moral acceptability ratings were *higher* for the surveillance cue condition. This is the opposite of what Bourrat et al. (2011) found.

5. Experiment 4

Our first three experiments did not replicate the findings of Bourrat et al. (2011). Each of our experiments was increasingly similar to

[7] All Mann-Whitney *U* Tests reported in the present study were two-tailed.

Table 6
Experiment 4: moral acceptability ratings for the control and surveillance cue conditions.

	Wallet vignette					Résumé vignette				
Condition	*M*	*Mdn*	Mode	*SD*	*n*	*M*	*Mdn*	Mode	*SD*	*n*
Control	2.83	2.00	1	2.40	48	3.11	2.00	1	2.38	46
Surveillance	2.17	1.50	1	1.67	48	3.31	3.00	2,3	1.96	48

Table 7
Experiment 1: moral acceptability ratings for the control and surveillance cue conditions.

	Wallet vignette					Résumé vignette				
Condition	*M*	*Mdn*	Mode	*SD*	*n*	*M*	*Mdn*	Mode	*SD*	*n*
Control	2.46	2.00	1	1.72	169	2.85	2.00	1	1.87	169
Surveillance	2.34	1.00	1	1.88	169	2.64	2.00	1	1.85	169

Bourrat and colleagues' experiment, though experiment location still differed. Our first three experiments were conducted in laboratories, whereas Bourrat et al.'s study was conducted in university libraries. In Experiment 4, we also conducted our experiment in libraries.

5.1. Method

5.1.1. Participants

We recruited 96 participants from McMaster University libraries. In order to avoid selection bias, we tried to ask every person in a room to participate. Because Bourrat et al. (2011) did not compensate their participants, we did not compensate ours. And as Bourrat et al. did not collect demographic data, we did not either.

5.1.2. Procedure

Library patrons who agreed to participate completed the moral acceptability task at the desk or table where they were found. Other than the lack of privacy and the location, the moral acceptability task was the same as in Experiment 3.

5.2. Results

Once again, we conducted Mann-Whitney *U* tests. No significant differences between the surveillance cue and the control conditions were found for either vignette (résumé: $U = 945$, $N_{surveillance} = 48$, $N_{control} = 46$, $p = 0.22$; wallet: $U = 1016$, $N_{surveillance} = 48$, $N_{control} = 48$, $p = 0.29$). This time, the difference in *résumé* vignette ratings went in the opposite direction of that obtained by Bourrat et al. (2011), with a larger mean, median, and mode for participants in the surveillance cue condition compared to the control condition. Descriptive statistics for the moral acceptability ratings, broken down by condition, are presented in Table 6.

5.3. Discussion

Our final experiment did not find significant differences between participants exposed to a surveillance cue and those exposed to a control image, conducted in as similar a location as Bourrat et al.'s (2011) experiment location as possible.

Table 9
Wallet vignette meta-analysis statistics.

Study	*N*	*SE*	ES	95% CI
Bourrat, Baumard, and McKay (2011)	91	0.46	−0.90	[−1.81, 0.00]
Sparks and Barclay (2015)	159	0.31	0.42	[−0.18, 1.02]
Present study - Experiment 1	338	0.20	−0.12	[−0.50, 0.27]
Present study - Experiment 2	573	0.20	−0.03	[−0.43, 0.37]
Present study - Experiment 3	93	0.30	0.50	[−0.08, 1.08]
Present study - Experiment 4	96	0.42	−0.67	[−1.49, 0.16]

Note. ES = unstandardized mean difference effect size, with negative values indicating lower moral acceptability ratings from participants in the surveillance cue conditions; CI = confidence interval.

6. Meta-analysis

In addition to Bourrat et al.'s (2011) original experiment and our four experiments, we know of only one other that investigated the impact of artificial cues of being watched on moral judgment (Sparks & Barclay, 2015). Participants completed the same moral judgment task we used in the present study. Like our experiments, this one obtained null results. To get a better picture of surveillance cue effects on moral judgment, we meta-analyzed these six moral judgment experiments. One meta-analysis included ratings of the wallet vignette and the other included ratings of the résumé vignette.

We utilized multiple conditions in our first two experiments, but the meta-analyses focused on the simple comparison of a surveillance cue condition to a control condition. Therefore, we combined some conditions. For Experiment 1, we combined the familiar face and unfamiliar face conditions into one surveillance cue condition, and we combined the chair and no image conditions into one control condition. The surveillance cue condition for Experiment 2 was formed by combining the female face and male face conditions. Descriptive statistics for the moral acceptability scores for the surveillance cue and control conditions for Experiments 1–4 are presented in Tables 5–8.

6.1. Methods

For the meta-analyses, we followed procedures outlined by Lipsey and Wilson (2001). Both meta-analyses consisted of six experiments: Experiments 1–4 from the present study, Bourrat et al. (2011), and Sparks and Barclay (2015). For each experiment, we compared the mean moral acceptability score for participants in the surveillance cue condition to the mean moral acceptability score for participants in the control condition. Because all six experiments used the same measures, we calculated unstandardized effect sizes. Each data point consisted of a comparison between the mean moral acceptability rating given by participants in an experiment's surveillance cue condition and the mean moral acceptability rating given by participants in the control condition. For both meta-analyses, we weighted the individual effect sizes using a random effects model and we calculated the overall mean effect size, the standard error of the overall mean effect size, and the 95% confidence interval of the overall mean effect size.

6.2. Results and discussion

The wallet vignette meta-analysis included 1350 participants. The findings for the wallet vignette meta-analysis are summarized in Table 9 and plotted in Fig. 4 (forest plot) and Fig. 5 (funnel plot). The mean

Table 8
Experiment 2: moral acceptability ratings for the control and surveillance cue conditions.

	Wallet vignette					Résumé vignette					Infidelity vignette				
Condition	*M*	*Mdn*	Mode	*SD*	*n*	*M*	*Mdn*	Mode	*SD*	*n*	*M*	*Mdn*	Mode	*SD*	*n*
Control	2.83	2.00	1	2.28	183	3.03	2.00	1	2.14	187	1.72	1.00	1	1.47	187
Surveillance	2.80	2.00	1	2.29	390	2.93	2.00	1	2.02	382	1.94	1.00	1	1.74	385

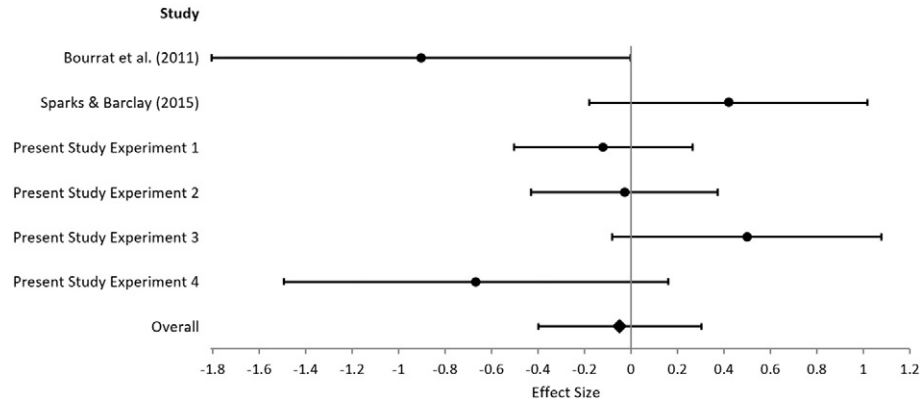

Fig. 4. Forest plot for wallet vignette meta-analysis: unstandardized mean difference effect size and 95% confidence interval for each study, and overall. A negative effect size indicates lower moral acceptability ratings from participants in the surveillance cue condition.

effect size, calculated as the unstandardized mean difference, was −0.05 ($SE = 0.18$), indicating lower moral acceptability ratings from participants in the surveillance cue conditions, as originally found by Bourrat et al. (2011). However, this effect size is tiny (an absolute value of 0.05 point on a 9-point scale) and not significantly different from zero, given that the 95% confidence interval for the effect size was −0.40 to 0.30. Thus, the wallet vignette meta-analysis does not provide evidence that artificial cues of being watched affect reported moral judgment.

The Q test for homogeneity of effect sizes was significant, $Q(5) = 11.48$, $p = 0.04$. This suggests that the effect size distribution was heterogeneous. Inspection of Fig. 5 reveals a distribution which does indeed appear heterogeneous; there are two negative effect sizes with relatively large standard errors, two positive effect sizes with medium-sized standard errors, and two effect sizes close to zero with relatively small standard errors. We have no ready explanation for this, since all the experiments utilized the same task. However, we used a random effects model, which does not assume homogeneity of effect sizes (Cumming, 2014; Lipsey & Wilson, 2001).

The résumé vignette meta-analysis included 1346 participants. The findings for the résumé vignette meta-analysis are summarized in Table 10 and plotted in Fig. 6 (forest plot) and Fig. 7 (funnel plot). The mean effect size was −0.21 ($SE = 0.11$), once again indicating lower moral acceptability ratings from participants in the surveillance cue conditions. The mean effect size was significantly different from zero with a 95% confidence interval of −0.43 to −0.00003. We note that the upper limit of the confidence interval was very close to 0, so the effect was just barely significant. Unlike the wallet vignette meta-analysis, the résumé vignette meta-analysis does provide some support for the claim that artificial surveillance cues affect reported moral judgment. The Q test for homogeneity of effect sizes was not significant for the résumé vignette, $Q(5) = 4.90$, $p = 0.43$, suggesting that the effect size distribution was homogeneous.

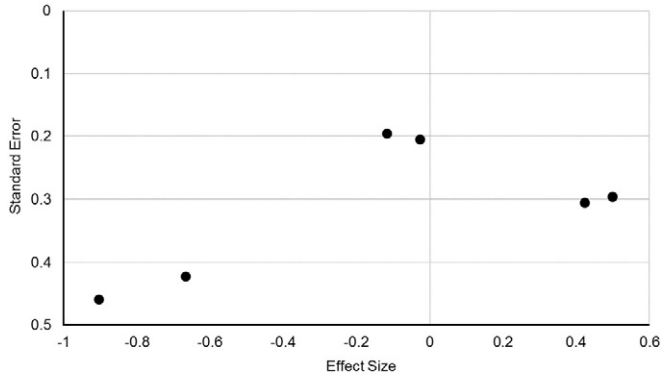

Fig. 5. Funnel plot for wallet vignette meta-analysis: unstandardized mean difference effect size and standard error of the unstandardized mean difference effect size for each study. A negative effect size indicates lower moral acceptability ratings from participants in the surveillance cue condition.

7. General discussion

Previously, two meta-analyses revealed no evidence for an effect of artificial surveillance cues on generosity (Northover et al., 2017). We still thought it important to examine another set of outcomes, as certain dependent measures may be more susceptible to surveillance cues than others. Inspired by Bourrat et al.'s (2011) study of the effect of surveillance cues on moral judgment, we conducted an experiment investigating the effect of surveillance cues on self-rated positive traits, religiosity, and moral judgment. We found no evidence for an effect on any of these variables. We conducted three additional moral judgment experiments, each increasingly similar in design to that of Bourrat and colleagues. None of our experiments replicated the surveillance cue effect on reported moral judgment.

We tested several possible moderators: the surveillance cue's location, the amount of attention that was drawn to the surveillance cue, the length of time the surveillance cue was displayed, participant privacy, experiment location, the apparent gender of the surveillance cue (a woman's face versus a man's face), and the familiarity of the surveillance cue (the face of a familiar person versus an unfamiliar person). Overall, our results do not provide compelling evidence that these variables moderate the effect of surveillance cues on reported moral judgment.

We then conducted two small meta-analyses of the six studies that have investigated the effect of artificial surveillance cues on a moral judgment task. In our view, the wallet vignette meta-analysis provides no evidence that artificial surveillance cues increase reported moral judgment, whereas the résumé vignette meta-analysis provides limited evidence. One possible explanation for this is that artificial surveillance cues cause people to report harsher judgment of certain moral

Table 10
Résumé vignette meta-analysis statistics.

Study	*N*	*SE*	ES	95% CI
Bourrat, Baumard, and McKay (2011)	93	0.45	−1.04	[−1.91, −0.17]
Sparks and Barclay (2015)	159	0.28	−0.23	[−0.77, 0.31]
Present study - Experiment 1	338	0.20	−0.21	[−0.60, 0.19]
Present study - Experiment 2	569	0.18	−0.10	[−0.46, 0.26]
Present study - Experiment 3	93	0.37	−0.38	[−1.10, 0.34]
Present study - Experiment 4	94	0.45	0.20	[−0.68, 1.08]

Note. ES = unstandardized mean difference effect size, with negative values indicating lower moral acceptability ratings from participants in the surveillance cue conditions; CI = confidence interval.

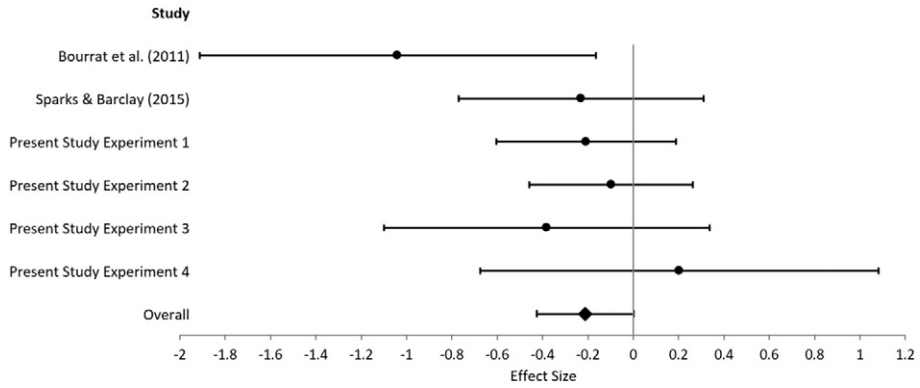

Fig. 6. Forest plot for résumé vignette meta-analysis: unstandardized mean difference effect size and 95% confidence interval for each study, and overall. A negative effect size indicates lower moral acceptability ratings from participants in the surveillance cue condition.

transgressions but not others. There may be something about the transgression described in the résumé vignette, as opposed to the wallet vignette, that is susceptible to the effects of surveillance cues. We do not know what that might be, however.

The meta-analysis results are mixed and based on a small number of studies; therefore, one cannot draw firm conclusions from them. Overall, however, we do not feel there is compelling evidence that surveillance cues affect moral judgment. In light of this, one possibility is that Bourrat et al. (2011) obtained a false positive (Francis, 2012; Simmons, Nelson, & Simonsohn, 2011).

It is, of course, possible that Bourrat et al. (2011) produced a true effect that we were unable to replicate. There may be some critical design feature that was present in Bourrat and colleagues' study but not present in ours (Higgins & Eitam, 2014). For example, Bourrat et al.'s study took place in France, whereas ours took place in Canada. Bourrat and colleagues presented the vignettes in French, whereas we used the original English language versions. It remains possible there is some cultural or language-related factor that is responsible for our lack of effect.

It is also possible that the number of people in the environment was perfect for producing a surveillance cue effect in Bourrat et al.'s (2011) study, whereas our experiments either had too few people in the environment (Experiments 1, 2, and 3) or too many (Experiment 4). Perhaps surveillance cues remind individuals that there are people in the area who can monitor their actions, and thus surveillance cues require the presence of at least some people in the area to affect behavior. Therefore, participants with total privacy, like our participants in Experiments 1, 2, and 3, but unlike the participants in Bourrat et al.'s study, may be immune to artificial surveillance cue effects. Artificial surveillance cues are also conceivably ineffective in crowded locations; they may be redundant in the presence of a large number of genuine surveillance cues. We conducted Experiment 4 in crowded libraries, whereas Bourrat and colleagues at least occasionally found participants sitting alone (P. Bourrat, personal communication, May 3, 2013). Population density may be an interesting moderating variable to consider in the future.

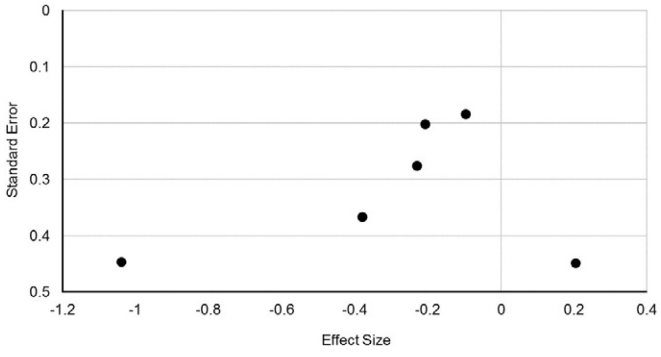

Fig. 7. Funnel plot for résumé vignette meta-analysis: unstandardized mean difference effect size and standard error of the unstandardized mean difference effect size for each study. A negative effect size indicates lower moral acceptability ratings from participants in the surveillance cue condition.

In conclusion, research conducted to date provides a mixed picture of artificial surveillance cues. Previously, two meta-analyses produced little evidence for artificial surveillance effects on generosity (Northover et al., 2017). In the present paper, we examined the effect of surveillance cues on moral judgment. The cumulative research on this topic is inconclusive, with five experiments finding no effect and one finding an effect, and yet an overall significant effect for one of the meta-analyses. If surveillance cues have true effects on reported moral judgment, perhaps they require specific circumstances, such as the right number of people in the environment, or a specific kind of moral transgression. Researchers may wish to consider such moderating variables if they choose to tackle this topic in the future. However, another possibility is that these effects are not robust and should be viewed with skepticism.

Acknowledgements

The authors would like to thank Bruce Milliken, David Feinberg, Gregory Atkinson, Chris McAllister, Jacqueline Walsh, Gray Boyko, Samantha Celli, Andrea Tankel, Katrina Pullia, Pierrick Bourrat, Adam Sparks, Stefan Pfattheicher, Yohsuke Ohtsubo, Ben Bolker, Oscar Gonzalez, and Steven Gangestad.

References**

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, *2*, 412–414. http://dx.doi.org/10.1098/rsbl.2006.0509.

*Bourrat, P., Baumard, N., & McKay, R. (2011). Surveillance cues enhance moral condemnation. *Evolutionary Psychology*, *9*(2), 193–199 (Retrieved from http://www.epjournal.net/wp-content/uploads/EP09193199.pdf).

Buss, D. M., & Schmitt, D. P. (1993). Sexual strategies theory: An evolutionary perspective on human mating. *Psychological Review*, *100*, 204–232. http://dx.doi.org/10.1037/0033-295x.100.2.204.

Carbon, C., & Hesslinger, V. M. (2011). Bateson et al."s (2006) cues-of-being-watched paradigm revisited. *Swiss Journal of Psychology*, *70*(4), 203–210. http://dx.doi.org/10.1024/1421-0185/a000058.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. http://dx.doi.org/10.1177/0956797613504966.

Edgell, P., Gerteis, J., & Hartmann, D. (2006). Atheists as "other": Moral boundaries and cultural membership in American society. *American Sociological Review*, *71*, 211–234. http://dx.doi.org/10.1177/000312240607100203.

Ernest-Jones, M., Nettle, D., & Bateson, M. (2011). Effects of eye images on everyday cooperative behavior: A field experiment. *Evolution and Human Behavior*, *32*, 172–178. http://dx.doi.org/10.1016/j.evolhumbehav.2010.10.006.

* *References marked with an asterisk indicate studies included in the meta-analysis.

Farkas, S., Johnson, J., Foleno, T., Duffett, A., & Foley, P. (2001). *For goodness' sake: Why so many want religion to play a greater role in American society.* New York, New York: Public Agenda.
Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*(6), 975–991. http://dx.doi.org/10.3758/s13423-012-0322-y.
Gervais, W. M., Shariff, A. F., & Norenzayan, A. (2011). Do you believe in atheists? Distrust is central to *anti*-atheist prejudice. *Journal of Personality and Social Psychology*, *101*(6), 1189–1206. http://dx.doi.org/10.1037/a0025882.
Gobbini, M. I., Leibenluft, E., Santiago, N., & Haxby, J. V. (2004). Social and emotional attachment in the neural representation of faces. *NeuroImage*, *22*, 1628–1635. http://dx.doi.org/10.1016/j.neuroimage.2004.03.049.
Haley, K. J., & Fessler, D. M. T. (2005). Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, *26*, 245–256. http://dx.doi.org/10.1016/j.evolhumbehav.2005.01.002.
Hall, D. L., Cohen, A. B., Meyer, K. K., Varley, A. H., & Brewer, G. A. (2015). Costly signaling increases trust, even across religious affiliations. *Psychological Science*, *26*, 1368–1376. http://dx.doi.org/10.1177/0956797615576473.
Higgins, E. T., & Eitam, B. (2014). Priming…shimming: It's about knowing when and why stimulated memory representations become active. *Social Cognition*, *32*, 225–242. http://dx.doi.org/10.1521/soco.2014.32.supp.225.
Jacoby, L. L., Kelley, C. M., & Dywan, J. (1989). Memory attributions. In H. L. Roediger, & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honor of Endel Tulving* (pp. 391–422). Hillsdale, NJ: Erlbaum.
Korte, C., & Kerr, N. (1975). Response to altruistic opportunities in urban and nonurban settings. *The Journal of Social Psychology*, *95*(2), 183–184. http://dx.doi.org/10.1080/00224545.1975.9918701.
Kurzban, R. (2001). The social psychophysics of cooperation: Nonverbal communication in a public goods game. *Journal of Nonverbal Behavior*, *25*(4), 241–259. http://dx.doi.org/10.1023/A:1012563421824.
Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, *28*, 75–84. http://dx.doi.org/10.1016/j.evolhumbehav.2006.06.001.
Levine, R. V., Martinez, T. S., Brase, G., & Sorenson, K. (1994). Helping in 36 U.S. cities. *Journal of Personality and Social Psychology*, *67*(1), 69–82. http://dx.doi.org/10.1037/0022-3514.67.1.69.
Li, Y. J., Cohen, A. B., Weeden, J., & Kenrick, D. T. (2010). Mating competitors increase religious beliefs. *Journal of Experimental Social Psychology*, *46*, 428–431. http://dx.doi.org/10.1016/j.jesp.2009.10.017.
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, California: Sage Publications.
Matland, R., & Murray, G. (2015). I only have eyes for you: Does implicit social pressure increase voter turnout? *Political Psychology*.. http://dx.doi.org/10.1111/pops.12275.
Matsugasaki, K., Tsukamoto, W., & Ohtsubo, Y. (2015). Two failed replications of the watching eyes effect. *Letters on Evolutionary Behavioral Science*, *6*, 17–20. http://dx.doi.org/10.5178/lebs.2015.36.
Nettle, D., Harper, Z., Kidson, A., Stone, R., Penton-Voak, I. S., & Bateson, M. (2013). The watching eyes effect in the dictator game: It's not how much you give, it's being seen to give something. *Evolution and Human Behavior*, *34*, 35–40. http://dx.doi.org/10.1016/j.evolhumbehav.2012.08.004.
Northover, S. B., Pedersen, W. C., Cohen, A. B., & Andrews, P. W. (2017). Artificial surveillance cues do not increase generosity: Two meta-analyses. *Evolution and Human Behavior*, *38*, 144–153. http://dx.doi.org/10.1016/j.evolhumbehav.2016.07.001.
Panagopoulos, C. (2014). Watchful eyes: Implicit observability cues and voting. *Evolution and Human Behavior*, *35*, 279–284. http://dx.doi.org/10.1016/j.evolhumbehav.2014.02.008.
Pew Research Center (2014). *How Americans feel about religious groups.* Washington, DC: Author (Retrieved from http://www.pewforum.org/2014/07/16/how-americans-feel-about-religious-groups).
Pfattheicher, S. (2015). A regulatory focus perspective on reputational concerns: The impact of prevention-focused self-regulation. *Motivation and Emotion*, *39*, 932–942. http://dx.doi.org/10.1007/s11031-015-9501-2.
Pfattheicher, S., & Keller, J. (2015). The watching eyes phenomenon: The role of a sense of being seen and public self-awareness. *European Journal of Social Psychology*, *45*, 560–566. http://dx.doi.org/10.1002/ejsp.2122.
Piazza, J., & Bering, J. (2008). The effects of perceived anonymity on altruistic punishment. *Evolutionary Psychology*, *6*, 487–501 (Retrieved from http://www.epjournal.net/wp-content/uploads/EP06487501.pdf).
Rushton, J. P. (1978). Urban density and altruism: Helping strangers in a Canadian city, suburb, and small town. *Psychological Reports*, *43*(3), 987–990. http://dx.doi.org/10.2466/pr0.1978.43.3.987.
Satow, K. (1975). Social approval and helping. *Journal of Experimental Social Psychology*, *11*, 501–509. http://dx.doi.org/10.1016/0022-1031(75)90001-3.
Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin*, *34*(8), 1096–1109. http://dx.doi.org/10.1177/0146167208317771.
Sedikides, C., & Gebauer, J. E. (2010). Religiosity as self-enhancement: A meta-analysis of the relation between socially desirable responding and religiosity. *Personality and Social Psychology Review*, *14*, 17–36. http://dx.doi.org/10.1177/1088868309351002.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. http://dx.doi.org/10.1177/0956797611417632.
Sparks, A., & Barclay, P. (2013). Eye images increase generosity, but not for long: The limited effect of a false cue. *Evolution and Human Behavior*, *34*, 317–322. http://dx.doi.org/10.1016/j.evolhumbehav.2013.05.001.
*Sparks, A., & Barclay, P. (2015). No effect on condemnation of short or long exposure to eye images. *Letters on Evolutionary Behavioral Science*, *6*, 13–16. http://dx.doi.org/10.5178/lebs.2015.35.
Statistics Canada. (2013). 2011 National Household Survey Health Profile (catalogue no. 82-228-XWE). (Retrieved from http://www12.statcan.gc.ca/health-sante/82-228/index.cfm?Lang=E).
Tan, J. H. W., & Vogel, C. (2008). Religion and trust: An experimental study. *Journal of Economic Psychology*, *29*, 832–848. http://dx.doi.org/10.1016/j.joep.2008.03.002.
van Rompay, T. J. L., Vonk, D. J., & Fransen, M. L. (2009). The eye of the camera: Effects of security cameras on prosocial behavior. *Environment and Behavior*, *41*(1), 60–74. http://dx.doi.org/10.1177/0013916507309996.
Yousif, Y., & Korte, C. (1995). Urbanization, culture, and helpfulness: Cross-cultural studies in England and the Sudan. *Journal of Cross-Cultural Psychology*, *26*(5), 474–489. http://dx.doi.org/10.1177/0022022195265002.