MORAL JUDGMENT: SURVEILLANCE CUES DEBUNKED

MORAL JUDGMENT: SURVEILLANCE CUES DEBUNKED

By STEFANIE BLAIR NORTHOVER, B. A.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Master of Science

McMaster University

MASTER OF SCIENCE (2014)
McMaster University
Psychology, Neuroscience, and Behaviour

TITLE: Moral Judgment: Surveillance Cues Debunked

AUTHOR: Stefanie Northover, B. A. (California State University, Long Beach)

SUPERVISOR: Dr. Paul Andrews

PAGES: viii, 69

Abstract

Several studies have seemingly demonstrated that artificial surveillance cues, such as images of watching eyes, increase prosocial behaviour. One of these studies investigated the effect of observation cues on moral judgment (Bourrat et al., 2011). Participants rated the moral acceptability of two misdeeds: falsifying information on a resume and keeping the cash found in a lost wallet. The moral acceptability ratings were lower for participants who were presented with an image of watching eyes than they were for participants exposed to a control image of flowers. The authors suggested that false cues of being watched triggered evolved cognitive mechanisms for recognizing when one is being observed. These mechanisms may have driven the cued participants to behave in a way that would have protected their reputations if they really had been watched; that is, by expressing disapproval of immoral behaviour. Inspired by Bourrat et al., I conducted an experiment investigating the effect of surveillance cues on self-rated positive traits, religiosity, and moral judgment. I found no evidence for an effect on any of these variables, including moral judgment. I conducted 3 more experiments, each increasingly similar in design to Bourrat et al., to determine the reason for the discrepancy in our results. None of my experiments replicated the surveillance cue effect. I suggest the most likely explanation is that Bourrat et al. obtained a false positive. My experimental results call into question the effect of surveillance cues on moral judgment; thus, it is appropriate to be skeptical of surveillance cues generally. I conducted a meta-analysis of studies investigating the effect of surveillance cues on generosity. The resulting funnel plot is consistent with publication bias in favour of significant results; it may also indicate that the surveillance cue effect on generosity, though perhaps a real phenomenon, is smaller than the literature implies.

Acknowledgements

This research would not have been possible without the help of several people. First, I would like to thank my mentor, William Pedersen, for helping me develop the idea for this study. I thank my advisor, Paul Andrews, for sharing his ideas and guiding the course of this project. I am grateful to the members of my committee, Bruce Milliken and David Feinberg, for patiently answering my various questions. I thank Gregory Atkinson and Chris McAllister for their clever website design. I am grateful to Ben Bolker and Steve Gangestad for helping with statistics. I appreciate Jacqueline Walsh, Gray Boyko, Samantha Celli, Andrea Tankel, and Katrina Pullia, my army of experimenters, for collecting data. I thank Daniel Nettle, Claus-Christian Carbon, Adam Sparks, Terry Burnham, Ryo Oda, Stefan Pfattheicher, Aurelien Baillon, Mary Rigdon, Keiko Ishii, Shinobu Kitayama, Toshio Yamagishi, and especially Pierrick Bourrat for providing me with data, statistics, and information about their studies.

I am also grateful to those who supported me during a difficult time. My mother has done far more for me than I deserve. My father has never stopped believing in me. My sister always has my back. I was lucky to meet Gray Boyko and Jacqueline Walsh, whose friendship made me feel less like a stranger in a strange land. Finally, I thank Alexis Alfaro, who means more to me than I can say. Alexis, here's to the next sixteen years.

Table of Contents

List of Figures

List of Tables

In 2011, Bourrat, Baumard, and McKay described an experiment in which participants presented with an image of watching eyes rated moral transgressions as less morally acceptable than participants presented with an image of flowers. Bourrat and colleagues suggested that the image of watching eyes, which served as a false cue of being watched, triggered evolved cognitive mechanisms for recognizing when one is being observed by others. These evolved mechanisms may have impelled the participants to express greater disapproval of the moral transgressions, as this behaviour would have made a good impression on real observers. Although several papers have reported significant surveillance cue effects, I believe Bourrat et al. obtained false positive results. In the present study, I review the surveillance cue literature. Then I describe four experiments I conducted in an attempt to replicate the findings of Bourrat and colleagues, none of which did so. Finally, I describe a meta-analysis I conducted of studies evaluating surveillance cues in the context of generosity. The results of the meta-analysis are consistent with publication bias in the surveillance cue literature, suggesting the surveillance cue effect may not be as strong or reliable as the literature implies.

People who know or believe their actions are being observed by others behave differently (e.g., Triplett, 1898; Zajonc, 1965; Putz, 1975; Latané, 1981; Bond & Titus, 1983; Aiello & Svec, 1993; Risko & Kingstone, 2011). For example, they are more generous (Satow, 1975; Kurzban, 2001), more helpful (van Rompay, Vonk, & Fransen, 2009), and more likely to participate in moralistic punishment (Kurzban, DeScioli, & O'Brien, 2007; Piazza & Bering, 2008). Additionally, several studies have seemingly demonstrated that under certain conditions, people who are presented with obviously artificial cues of being observed, such as images of watching eyes, will behave as if they really are being observed. If true, this suggests that when people are being watched, their conscious belief of

observation is not required for changes in behaviour. An unconscious, automatic component seems to come into play; one that may be triggered by surveillance cues.

Many of the artificial surveillance cue studies have utilized the dictator game, an economic game with two players. One of the players, the dictator, receives money and decides how to allocate it among the two players. The second player merely accepts what the dictator has given him/her, if the dictator chooses to give anything at all. Haley and Fessler (2005) published one of the first papers on artificial surveillance cues in the context of the dictator game. Half of their dictators were presented with a stylized image of Horus eyes on their computer desktop; the other half were presented with a control desktop image. Dictators allocated more money, on average, in the eyes condition. Evidence for increased generosity due to images of watching eyes has been found in other dictator game studies as well (e.g., Oda, Niwa, Honma, & Hiraishi, 2011; Baillon, Selim, & van Dolder, 2013), although the results have often been mixed or conditional (e.g., Rigdon, Ishii, Watabe, & Kitayama, 2009; Mifune, Hashimoto, & Yamagishi, 2010; Raihani & Bshary, 2012; Nettle et al., 2013; Sparks & Barclay, 2013). Some researchers did not find significantly increased generosity among dictators presented with images of eyes (e.g., Sparks, 2010). One such experiment was conducted by Tane and Takezawa (2011). The authors suggested that their use of a dark, sound-proof room in which participants sat alone canceled out surveillance cue effects. I find this explanation plausible, but Tane and Takezawa's results are exactly what one would expect them to be if surveillance cues don't affect behaviour.

Artificial surveillance cues have been used in other kinds of economic games with ambiguous results. Kismet, a robot created at the Massachusetts Institute of Technology, served as a surveillance cue in Burnham and Hare's (2007) public goods game. Participants with the image of Kismet on their

computer screen, watching them with his large eyes, contributed more money than participants in a control group. On the other hand, Fehr and Schneider (2010) found no significant differences between groups in a trust game experiment.

The relationship between surveillance cues and generosity has also been examined with charity donation studies. For instance, Ekström (2012) placed images on recycling machines in Swedish supermarkets. Customers used the machines to recycle cans and bottles and were given a choice of whether to keep the money earned or donate it to a charity. An image of eyes was displayed for half the time and a control image of flowers was displayed the other half. The results of this study are difficult to interpret; when considering only the days on which few recycling customers visited the stores, there was a 30% increase in the amount of money donated in the surveillance cue condition, but overall, there was no difference in the amount donated by customers when the machines displayed eyes compared to flowers.

Powell, Roberts, and Nettle (2012) conducted a supermarket experiment as well. Buckets located at checkouts were used to collect donations to a charity. Half of the buckets displayed an image of eyes and the other half displayed an image of three stars. Donations were 48% higher to eyes than control buckets. As in Ekström's (2012) study, surveillance cues apparently affected donations more strongly when there were fewer customers present. During slow weeks, the eyes buckets received 59% more in donations per thousand customers; during busier weeks, the eyes buckets received only 28% more.

Keller and Pfattheicher (2011) conducted a charity donation study in a lab. Participants who were high in prevention-focused self-regulation (i.e., highly focused on preventing negative events in

their lives) donated more money to charity in the eyes condition than they did in the no-eyes condition. Participants who were low in prevention-focused self-regulation showed the opposite pattern: They donated less money to charity in the eyes condition than they did in the no-eyes condition. The authors suggested that people who are habitual prevention-focused self-regulators are especially sensitive to cues of being monitored.

As far as I know, the first published field study on the effects of images of watching eyes was conducted by Bateson, Nettle, and Roberts (2006). An honesty box, located in a university coffee room, was used to collect money to pay for coffee and tea. Anonymous people could put as much money into the box as they wanted. When a photograph of eyes was displayed on a cupboard door, nearly three times as much money was contributed to the honesty box compared to when a photograph of flowers was displayed (but see Carbon & Hesslinger, 2011, for a critique).

Francey and Bergmüller (2012) conducted a field experiment in which trash was placed on bus stop benches in Geneva. People spent a significantly greater amount of time disposing of garbage when images of eyes were displayed compared to when images of flowers were displayed, although the actual proportion of people who threw away trash did not differ significantly between groups. In another field experiment, litter was left on fewer tables in a cafeteria when photographs of eyes were placed on the walls (Ernest-Jones, Nettle, & Bateson, 2011). This was especially true when the cafeteria contained relatively few people. The outcome of this experiment as well as that of Ekström's (2012) recycling machine study and Powell et al.'s (2012) charity donation bucket study suggests that surveillance cues may be redundant in the presence of large numbers of people. It seems likely that an individual in this situation is already being cued by the crowd of real people in the vicinity. The

possibility that the noise of a large crowd distracts individuals and decreases the likelihood of the surveillance cue being noticed has not been ruled out, however (Ekström, 2012). On the other hand, another field study of littering behaviour found the opposite effect: Bicyclists on a university campus who were exposed to images of watching eyes were less likely to litter than those who weren't exposed to surveillance cues, but only when there was a *greater* number of people in the vicinity (Bateson, Callow, Holmes, Redmond Roche, & Nettle, 2013).

A field study conducted by Panagopoulos (2014) investigated surveillance cues in the context of voting behaviour. Panagopoulos mailed postcards to registered voters in Key West, Florida. All of the postcards displayed a message urging recipients to vote in an upcoming mayoral election. Next to the message was an image which varied depending on condition: watching eyes (surveillance cue), an American flag (political control), or a palm tree (apolitical control). A larger proportion of voters in the surveillance cue condition participated in the election compared to voters in the control conditions.

Most of these studies suggest that surveillance cues increase generosity and cooperation. Assuming the reality of the surveillance cue effect, the process probably begins with the detection of watching eyes. When a person detects a face or eyes, certain areas of the brain activate. Evidence suggests a region in the lateral fusiform gyrus is in charge of identifying faces (George et al., 1999; Hoffman & Haxby, 2000) and a region in the superior temporal sulcus appears to be involved in the representation of the changeable aspects of faces, such as gaze direction and mouth movements (Puce, Allison, Bentin, Gore, & McCarthy, 1998; Hoffman & Haxby, 2000; Emery, 2000). It is possible that this activation of the brain results in unconscious alteration of behaviour in various ways that serve to

protect and improve the individual's reputation. This automatic response may have been selected for due to the fitness-enhancing effects a positive reputation provides.

Throughout time, natural selection has probably favoured individuals who behave prosocially when they are being watched and thus create positive reputations for themselves (e.g., Wedekind & Milinski, 2000; Smead, 2010). The importance of reputation can be understood in the context of indirect reciprocity. While direct reciprocity proposes that individuals are likely to help people who have helped them in the past, indirect reciprocity proposes that individuals are likely to help people who have helped others (Alexander, 1987; Nowak & Sigmund, 1998; Nowak & Sigmund, 2005). People learn about the generosity of an individual either by watching the individual interact with others or by word of mouth (Axelrod, 1984; Wedekind & Milinski, 2000; Nowak & Sigmund, 2005). When indirect reciprocity is evaluated mathematically, it is shown that natural selection can favour strategies that benefit people based on their reputation (Nowak & Sigmund, 2005). Someone with a reputation as an altruist tends to receive help from others, so it is in an individual's best interest to advertise altruism; i.e., to be cooperative, generous, and altruistic when being watched.

Empirical evidence indicates that a good reputation reaps rewards. This is often demonstrated in games. For example, Wedekind and Milinski (2000) found that players of an economic game were more generous toward individuals who had built up good reputations during the course of the game than they were toward individuals whose reputations were less glowing.

Wedekind and Braithwaite (2002) conducted an experiment in which students played an indirect reciprocity game for money. Although committing acts of generosity hurt generous individuals in the short-term, they built up good reputations that paid off for them in the long run because they were

rewarded by third parties. Those with the best reputations benefited in a subsequent direct reciprocity game as well, perhaps because they were easier to trust.

Barclay and Willer (2007) conducted a cooperative monetary game in which participants with the chance to be chosen for beneficial partnerships attempted to outdo other participants in the matter of altruism. This 'competitive altruism' paid off for the most generous participants, who were chosen to be partners over less altruistic participants.

Another experiment in which participants demonstrated competitive altruism involved a donation game (Milinski, Semmann, & Krambeck, 2002). Participants who publicly donated money to a charity received more money from the group (via donations) and were elected to represent the participants in a students' council.

Due to the large body of research suggesting that a reputation for being altruistic reaps rewards for the reputation-holder, surveillance cue effects on altruism and generosity are at least plausible. But what about reputational concerns besides generosity? One possible reaction to being watched is to express or demonstrate agreement with the moral values held (or seemingly held) by members of the community. Studies have demonstrated that when people judge whether the intentions of a certain individual are hostile or friendly, they are more concerned with the morality of the individual than other traits (Leach, Ellemers, & Barreto, 2007; Brambilla, Rusconi, Sacchi, & Cherubini, 2011). If an individual is believed to share the same moral compass as others, then those others should be more likely to trust the individual and to choose the individual for mutually beneficial partnerships and exchanges. Consider the issue of moral judgment, for example. If individuals persuade others that they condemn stealing (or dishonesty, or adultery, etc.), then others should trust that they will not steal

(or lie, or cheat, etc.) and these apparently virtuous individuals will be able to enter into beneficial social interactions.

Moral judgment was the topic evaluated by Bourrat and colleagues (2011), whose study I am most interested in. Specifically, participants were asked to rate the moral acceptability of two misdeeds: returning a lost wallet but keeping the money, and falsifying information on a resume (Schnall, Haidt, Clore, & Jordan, 2008). Participants with a photograph of eyes on their paper rated the violations as less morally acceptable than did participants with a photograph of flowers.

Experiments Overview

I carried out four experiments investigating the phenomenon of artificial surveillance cue effects. Initially, I was interested in multiple variables – religiosity, positive traits, and moral judgment (Experiment 1). When I was unable to conceptually replicate Bourrat et al.'s (2011) moral judgment results, however, I focused on determining why that was (Experiments 2, 3, and 4).

Experiment 1

Surveillance cue traits. In addition to exploring effects of surveillance cues generally, I investigated different properties of surveillance cues. First, I explored the possibility that using a familiar face as a surveillance cue would affect behaviour to a higher degree than an unfamiliar face would. A familiar face may alert individuals that they are being watched by a member of their community. Research has shown that people are more likely to behave prosocially in less densely-populated areas (Korte & Kerr, 1975; Rushton, 1978; Levine, Martinez, Brase, & Sorenzon, 1994; Yousif & Korte, 1995). This may be due to the fact that the denser the population of a community, the more difficult it is to build and maintain a reputation – there are too many people to keep track of.

Therefore, people should behave better when they are observed by familiar individuals, among whom they have reputations, than they do when they are observed by unfamiliar individuals. If they do, it follows that reputation-enhancing behaviour, such as condemnation of moral transgressions, may be more strongly affected by cues of being watched by familiar people than unfamiliar people. Consistent with this, drug store customers who witnessed a confederate shoplifting were more likely to report the theft if they had grown up in a rural environment rather than an urban one (Gelfand, Hartmann, Walder, & Page, 1973).

I also investigated potential differences between masculine and feminine surveillance cues (i.e., watching male faces vs. watching female faces). In Bateson et al.'s (2006) field experiment, described previously, anonymous people contributed more money to an honesty box, used to collect funds for coffee, when masculine eyes were displayed than they did when feminine eyes were displayed (but see Carbon & Hesslinger, 2011). Other studies found no differences between masculine and feminine eye cues, however (e.g., Nettle et al., 2013).

Additional dependent measures. In Experiment 1, I 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 behaviour, 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 display alignment with the community's religious beliefs. Experimental evidence indicates that religiosity fosters trust (Tan & Vogel, 2008) and that atheists are viewed as untrustworthy (Gervais, Shariff, & Norenzayan, 2011). Most Americans

associate religion with morality and trustworthiness (Edgell et al., 2006). In a Public Agenda survey, 69% of Americans reported that "more religion is the best way to strengthen family values and moral behavior in America" (Farkas et al., 2001). Individuals who play up their religious beliefs (or conceal a lack of them) may convince others that they are trustworthy and moral.

Method

Participants. I recruited 338 psychology students from McMaster University. Subjects 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 sex; about 50% were White, 40% Asian, 6.5% Middle Eastern, and 5% indicated some other ethnicity.

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' only task was to complete a computer questionnaire made up of three main 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. If the participant was in the no image condition, then the right frame was blank. If the participant was in any of the other three conditions, the right frame displayed an image. 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.

In order to decrease suspicion of the images, a cover story was given at the beginning of the

experiment. The participants were told, "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 experimenter blindness, 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, at the end of which the right frame loaded either an image (for participants in the familiar face, unfamiliar face, or chair conditions) or a blank page (for participants in the no image condition).

At the end of the experiment, participants were probed for suspicion. Data were removed for a few participants who correctly guessed the purpose of the experiment.

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.

The individuals chosen for the familiar face condition were those I believed would be easily recognized by most of the participants: Kristen Stewart, Rihanna, Taylor Swift, Barack Obama, Danielle Radcliffe, and Tom Hanks. For our hunter-gatherer ancestors, familiarity of a face probably meant that the face belonged to a member of the same community. Today we recognize the faces of many individuals whom we do not know personally, such as celebrities and politicians. Until fairly recently, evolutionarily, situations in which a familiar face belonged to someone outside of one's own community were probably rare, so there was no need to discriminate between the familiar face of a

fellow community member and the familiar face of a member of an out-group. I reasoned that if this was true, and if it was also the case that surveillance cues affect behaviour, then it might also be true that many of the mechanisms that are triggered when one sees a person one knows (a friend, a co-worker, a family member, etc.) are also triggered when one sees a familiar celebrity (whom one has never met).

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. I 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 I was able to investigate the dependent measures according to the masculinity/femininity of the surveillance cues. See Figure 1 for an example of a familiar face and an unfamiliar face.

The chair images were taken off the internet and were chosen because they were simple and boring. My goal was to use chair images that didn't incite any emotions. See Figure 2 for an example of a chair image.

Each face image had an interpupillary distance of 115 or 116 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 look at the participants. I used a monitor with a viewable image size of 59.69 cm and a screen resolution of 1,920 by 1,080 pixels. The mean area of the face images was 170,871 square pixels. The mean area of the chair images was 192,134 square pixels.

Instruments. The survey was made up of different sections. Participants first completed either the religiosity section, the possession of positive traits section, or the moral judgment section. The

order of these three main sections was counterbalanced.

For the religiosity section, participants were asked to rate, on a 9-point Likert scale, how strongly they agreed with the following three statements (Li, Cohen, Weeden, & Kenrick, 2010):

1. I believe in God.
2. We'd be better off if religion played a bigger role in people's lives.
3. Religious beliefs are important to me in my everyday decisions.

Larger values indicate stronger agreement. Ratings for the three statements were added up to create a total religiosity score which could range from 3 (least religiosity) to 27 (most religiosity).

For the positive traits section, participants were asked to rate, on a 9-point Likert scale, how strongly they agreed with the following statements:

1. I am kind.
2. I am competent.
3. I am dishonest.
4. I am unreliable.
5. I am weak.
6. I am attractive.
7. I am brave.
8. I am insecure.
9. I am generous.
10. I am intelligent.

These traits were chosen because they are related either to prosocial behaviour (e.g., generous) and/or

they are traits one might find attractive in a mate (e.g., competent, intelligent). Items 3, 4, 5, and 8 were reverse scored. Larger values again indicate greater agreement. Scores for all ten traits were added together to create a total positive traits score, which could range from 10 (lowest possible rating of positive traits) to 90 (greatest possible rating).

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, my participants read the same vignettes. One of the vignettes, which I will refer to as the "wallet vignette", read as follows: "You are walking down the street when you come 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 other vignette, which I will refer to as the "resume vignette", read as follows: "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 the vignettes on a 9-point Likert scale, with 1 = *morally unacceptable* and 9 = *morally acceptable*.

Results

Manipulation check on familiarity. A chi-square test of independence was conducted to determine if the faces in the familiar face and unfamiliar face conditions successfully differed in familiarity. At the end of the experiment, participants in these two conditions were asked to rate the familiarity of the face in their images on a 4-point Likert scale, with 1 being *not familiar at all*, 2 being *slightly familiar*, 3 being *familiar*, and 4 being *very familiar*. Participants in the familiar face condition ranked their faces ($Mdn = 4$) significantly differently from participants in the unfamiliar face condition ($Mdn = 1$), $\chi^2(3, N = 169) = 131$, $p < .001$, Cramér's $V = .88$. That is, faces in the familiar face condition were more familiar than faces in the unfamiliar face condition.

Effect of familiar and unfamiliar faces. Since none of the dependent measures were normally distributed, a series of Kruskal-Wallis 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, $H(3, N = 336) = 0.85$, $p = .84$. No significant differences were found in moral acceptability for either the wallet vignette ($H(3, N = 338) = 5.00$, $p = .17$) or the resume vignette ($H(3, N = 338) = 3.93$, $p = .27$). The Kruskal-Wallis test was marginally significant for positive traits, $H(3, N = 329) = 7.35$, $p = .062$, 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.

Effect of masculine and feminine faces. To determine if the masculinity or femininity of the surveillance cues was important, a series of Kruskal-Wallis tests were conducted with male face, female face, chair, and no image groups. No differences were found in religiosity ($H(3, N = 336) = 1.27$, $p = .74$), positive traits ($H(3, N = 329) = 6.31$, $p = .098$), or moral acceptability ratings for the resume vignette ($H(3, N = 338) = 3.78$, $p = .29$). The differences approached significance for the wallet vignette, $H(3, N = 338) = 7.69$, $p = .053$. Investigating this further, I found that this marginally significant result was driven by the women, so I conducted a couple of 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 for the resume vignette were still not significant ($H(3, N = 253) = 5.76$, $p = .12$) but the differences for the wallet vignette became significant ($H(3, N = 253) = 12.32$, $p = .006$). On average, the women who were presented with an image of a male face reported lower moral acceptability scores for the wallet vignette than women presented with an image of a female face, an image of a chair, or no image at all. Table 1 displays descriptive statistics for religiosity and positive traits. Descriptive statistics for the moral acceptability ratings are provided in Table 2.

Surveillance cue duration. Sparks and Barclay (2013) hypothesized that images of eyes have the greatest effect on behaviour when they appear briefly; those that remain visible for too long lose their effectiveness. Indeed, a meta-analysis they conducted supports this hypothesis. I 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. There were six different versions of the questionnaire, one for each possible ordering of the three tasks (e.g., religiosity first, moral acceptability second, and positive traits third). Each participant was randomly assigned to one of these task orders. Since there were two different combinations of task orders out of the six that resulted in a given task being taken first, 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 ten seconds 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. In other words, participants were exposed to images only briefly before beginning their first task.

A couple of Kruskal-Wallis tests were conducted examining positive traits using only data from participants for whom the positive traits questionnaire was the first task; in other words, 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, and found no significant differences ($H(3, N = 108) = 6.98, p = .073$). The other test compared the female face, male face, chair, and no image conditions and also found no significant differences ($H(3, N = 108) = 3.64, p = .30$).

Next, a couple of 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 ($H(3, N = 112) = .99, p = .80$). The other test compared the female face, male face, chair,

and no image conditions and also found no significant differences ($H(3, N = 112) = .96$, $p = .81$).

Finally, a series of 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 ($H(3, N = 113) = .27$, $p = .97$ for the wallet vignette; $H(3, N = 113) = 3.53$, $p = .32$ for the resume vignette). There were also no significant differences when comparing the female face, male face, chair, and no image conditions ($H(3, N = 113) = 2.18$, $p = .54$ for the wallet vignette; $H(3, N = 113) = 3.13$, $p = .37$ the resume vignette).

Comparing men and women. 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 the women (see Table 3). Mann-Whitney U-Tests revealed the difference between men and women to be significant for the wallet vignette, $U = 8155$, $N_{men} = 83$, $N_{women} = 253$, $p = .001$. The difference between men and women was not significant for the resume vignette, $U = 9857$, $N_{men} = 83$, $N_{women} = 253$, $p = .39$, but the data showed the same pattern as the wallet vignette ratings, with higher means and medians for men. These data suggest men are more lenient than women.

Discussion

Cues of being watched didn't 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 considering only women, moral acceptability ratings of the wallet vignette were lower for those presented with a male face than they were for the other groups. This was a result I was not expecting nor had any theoretical explanation for.

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

Length of surveillance cue exposure. Sparks and Barclay (2013) found that surveillance cue effects were more likely to appear when the surveillance cue was revealed only briefly, possibly because people habituate to surveillance cues if they're visible for too long. I investigated the possibility that surveillance cue effects would show up in my results if I restricted analyses to participants who were exposed to surveillance cues briefly. Participants in my study who completed the moral judgment task before the other two tasks were exposed to the surveillance cue for about the same amount of time as participants in Bourrat and colleagues' (2011) study. I found no surveillance cue effects among these briefly-exposed participants. Therefore, I don't believe surveillance cue exposure duration can explain why Bourrat et al. obtained significant results and I did not.

Surveillance cue location. Bourrat and colleagues placed their surveillance cue, an image of watching eyes, directly above the rating scale. In Experiment 1, my surveillance cue was located on the right-hand side of a computer monitor, several inches away from the rating scale. My surveillance cue may have been too far away from the participants' line of sight to influence their responses. It's possible some participants didn't see the surveillance cues well enough to be affected by them.

Participants' conscious awareness of surveillance cue. In Bourrat et al.'s study, participants were not given any information about the image on their paper. In Experiment 1, however,

experimenters told participants that they might see images, and, if so, to pay careful attention to them. Perhaps surveillance cues are less effective when participants are consciously thinking about and focusing on them. A speculative explanation for this is that participants presented with an image of a watching face in Experiment 1 experienced feelings of being watched, but attributed the feelings to the surveillance cue that they were told to pay attention to; thus, they discounted the feelings of observation. (For a discussion of the role of attribution in memory, see Jacoby, Kelley, & Dywan, 1989.) If my participants hadn't been told to pay attention to the face, perhaps they wouldn't have identified the source of their feelings of being watched and thus wouldn't have discounted them.

Experiment 2

I made a few changes in Experiment 2. First, I moved the surveillance cue closer to the rating scale. Second, I attempted to reduce the participants' level of conscious attention to the surveillance cue by providing no explanation for it. Third, I cut down on the length of surveillance cue exposure. These changes made Experiment 2 more similar to Bourrat and colleagues' (2011) study than Experiment 1. I tested the possibility that the changes would bring out surveillance cue effects.

I also used Experiment 2 to further investigate the apparent interaction between participant sex and surveillance cue gender that I found in Experiment 1. I hadn't expected to find this interaction and I had no theoretical explanation for it, so it was important to replicate it before assuming there was validity to the finding. In Experiment 2, I also evaluated surveillance cue traits that may have helped account for the interaction.

Method

Participants. I 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.

Procedure. Experiment 2 consisted of just the moral judgment task, although a third vignette was added. Once again, I administered the moral judgment task with LimeSurvey (www.limesurvey.org). This time, however, the participants completed the task on their own computers, in their homes or wherever else they were working.

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 Likert scale used to rate the moral acceptability of the behaviour 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.

The results of the first experiment suggested that the moral judgment of women had been affected by images of male faces. I attempted to replicate that finding with Experiment 2, and

additionally I considered facial traits that might have helped explain the effect of male faces on women's moral judgment: attractiveness, authoritativeness, judgment, and disgustingness. After completing the moral judgment task, participants who were exposed to images of faces were asked to rate the faces on those traits.

I included the following question at the end of the experiment in order 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 either indicated that they hadn't 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).

Stimuli. The images fell into two categories, a face or a chair. Fifteen different images of chairs were used and 30 different images of faces. The chairs, which were located on a white background, were again chosen because they were simple and boring. The faces were chosen with variety in mind. Fifteen were female and 15 were male. Some of the people in the images were famous; most were not. I chose images which I believed varied in attractiveness, from extremely attractive to extremely unattractive. 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 look at the participants. The mean area of the face images was 134,426 square pixels. The mean area of the chair images was 144,002 square pixels. Each participant was randomly assigned and shown only one image.

Instruments. In addition to the original two vignettes, I added a third vignette about infidelity.

The 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 described previously, participants who were exposed to images of faces were asked to rate the faces on attractiveness, authoritativeness, judgment, and disgustingness. Attractiveness was rated on a 9-point Likert scale (from 1 = *extremely unattractive* to 9 = *extremely attractive*). The other three traits were rated on 5-point Likert scales (from 0 = *not at all* to 4 = *extremely*). The wording was as follows: "How attractive is this individual?"; "Authority is the power to control or command people. How authoritative does this individual seem to be?"; "How judgmental does this individual seem to be?"; and "How disgusting is this individual?".

Results

I conducted Kruskal-Wallis tests comparing the male face, female face, and chair conditions. There were no significant differences for the wallet vignette ($H(2, N = 573) = .25$, $p = .88$), the resume vignette ($H(2, N = 569) = .38$, $p = .83$), or the infidelity vignette ($H(2, N = 572) = 1.41$, $p = .49$). Descriptive statistics are displayed in Table 4.

As explained previously, I asked participants who were exposed to an image of a face to rate the face on various traits in order to help me understand the apparent effect of male faces on the moral judgment of women. A series of Kruskal-Wallis tests, again comparing the male face, female face, and chair conditions but conducted using data only from women, found no significant differences for the wallet vignette ($H(2, N = 234) = .091$, $p = .96$), the resume vignette ($H(2, N = 228) = .14$, $p = .93$), or

the infidelity vignette ($H(2, N = 232) = .97$, $p = .62$). In other words, when examining only women, there were no differences between those who were shown a male face and the rest of the women; therefore, the result from Experiment 1 did not replicate. I decided to explore the relationship between the facial trait variables and moral acceptability anyway. I used Poisson loglinear generalized linear models to predict the moral acceptability ratings using the facial traits (attractive, authoritative, judgmental, and disgusting), as well as gender, as predictor variables (see Table 5). Gender was a significant predictor of the wallet and resume vignette moral acceptability ratings, with men reporting higher ratings on average than women, but not of the infidelity vignette moral acceptability ratings. Disgust was a highly significant predictor of all three moral acceptability ratings (all ps < .001), with participants who rated their face images as more disgusting reporting higher moral acceptability ratings. This relationship between disgust and moral leniency is inconsistent with research suggesting that induced feelings of physical disgust can increase reported moral condemnation (Wheatley & Haidt, 2005; Schnall et al., 2008; Horberg, Oveis, Keltner, & Cohen, 2009). Table 6 displays moral acceptability rating statistics for men and women.

Discussion

In Experiment 1, I found a difference in moral acceptability ratings between women presented with an image of a male face and the rest of the women. This result did not replicate in Experiment 2. I believe the original finding was a false positive; it's hardly surprising the occasional Type I error (an incorrect rejection of the null hypothesis) would slip in among so many statistical analyses.

In Experiment 1, the moral acceptability ratings were higher for men than they were for women. This pattern replicated in Experiment 2 (see Table 6), again suggesting that men are more lenient than

women.

In Experiment 1, I found no differences between groups based on the apparent gender of the surveillance cues; I did not find these differences in Experiment 2, either. More importantly, I found no differences in moral acceptability ratings between the surveillance cue conditions and the control condition, despite the strategic changes in design. Moving my surveillance cue to a location that was closer to the rating scale, and thus closer to the participants' line of vision while they were choosing their rating, didn't bring out the effect Bourrat and colleagues (2011) found; nor did my lack of acknowledgment and explanation for the surveillance cues. Therefore, I don't believe the surveillance cue's location nor the attention drawn to it can explain the differences between Bourrat et al.'s results and mine. In light of the findings of Sparks and Barclay (2013) that surveillance cues are more likely to affect behaviour when they are presented briefly, I reduced exposure time in Experiment 2. Since I still found no surveillance cue effects, cue exposure duration can probably be ruled out as an explanation for my null results as well.

Experiment 2 was more like Bourrat and colleagues' study than Experiment 1, yet I 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. Experiments 1 and 2 were attempts at conceptual replications; for Experiment 3, I tried a closer replication.

Experiment 3

Two conceptual replications of Bourrat et al. (2011) did not reveal an effect of surveillance cues on moral judgment ratings. Experiment 3 was a truer replication of Bourrat and colleagues.

Method

Participants. I recruited 93 psychology students from McMaster University. Subjects received course credit for their participation. In order to increase feelings of anonymity, Bourrat et al. (2011) did not collect demographic data; therefore, I did not collect demographic data either.

Procedure. The procedure for Experiment 3 was similar to that of Bourrat and colleagues (2011). In both studies, the participants completed the moral judgment task on a piece of paper with the wallet vignette on one side and the resume vignette on the other. I did not include 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 (see Figure 3). I used the same images as Bourrat et al. (2011) and scaled them to the same size they did (47 x 17 mm). Participants circled the number of their choice on the rating scale with a pen.

Experiment 3 was meant to copy 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, wherever they were recruited. Some of the participants were by themselves, while others were not (P. Bourrat, personal communication, May 3, 2013). My participants completed their task in a laboratory, however, isolated in their own rooms. Furthermore, Bourrat and colleagues' participants did not receive compensation, while my participants received course credit.

Results

Since I compared only two groups for Experiment 3, I used Mann-Whitney U tests. For the resume vignette, the median moral acceptability rating was lower for the eyes condition (*Mdn* = 2) than it was for the flowers condition (*Mdn* = 3), but moral acceptability ratings were not significantly

different ($U = 922$, $N_{eyes} = 46$, $N_{flowers} = 47$, $p = .21$).

For the wallet vignette, the medians were the same for both conditions ($Mdn = 2$), but moral acceptability scores were marginally significantly greater for the eyes condition ($U = 877$, $N_{eyes} = 46$, $N_{flowers} = 47$, $p = .10$). This difference is going in the opposite direction of Bourrat et al. (2011), however. Descriptive statistics for the moral acceptability ratings, broken down by condition, are presented in Table 7.

Discussion

There were no significant differences in moral acceptability ratings of the resume vignette between the eyes group and the flowers group. I did find a marginally significant difference in ratings of the wallet vignette, but it was going the "wrong" way; ratings were *higher* for the eyes condition. This is the opposite of what Bourrat and colleagues found.

Experiment 4

My first three experiments did not replicate the findings of Bourrat et al. (2011). Each of my experiments was increasingly similar to Bourrat and colleagues' experiment, with Experiment 3 being almost the same. Location was still a major difference between my experiments and Bourrat et al.'s, however. My first three experiments were conducted in laboratories, while Bourrat et al.'s study was conducted in university libraries. Perhaps there was some factor unique to one of the environments, such as participant privacy, that was responsible for our disparate results. Bourrat and colleagues conducted their experiment in a public location where the number of people near the participants was not controlled for. Some of the participants may have been isolated, while others may have been near other people. I considered the possibility that for Bourrat et al.'s participants, surveillance cues

amplified feelings of being monitored by real people in the vicinity, perhaps by reminding the participants that they were in a public place surrounded by other people who may have been watching them. All of my participants had been isolated in private rooms. Perhaps my participants felt too secure in their privacy to feel like they were being watched, regardless of the kind of image that was presented to them. Even if privacy had no effect on the outcome of the experiments, there may have been something about public libraries versus laboratories that was key to bringing out a surveillance cue effect. In Experiment 4, I tested the possibility that the laboratory environment from the first three experiments was the reason for my lack of significant results. I attempted to recreate the environment from Bourrat et al. (2011) in this final experiment.

Method

Participants. I recruited 96 participants from McMaster University libraries. In order to avoid selection bias, I tried to ask every person in a room to participate. Since Bourrat and colleagues did not compensate their participants, I did not compensate mine. And as Bourrat et al. did not collect demographic data, I did not either.

Procedure. The procedure for Experiment 4 was designed to be as similar to that of Bourrat et al. (2011) as possible. 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.

Results

Once again, I conducted Mann-Whitney U tests. No significant differences between the eyes and the flowers conditions were found for either vignette ($U = 945$, $N_{eyes} = 48$, $N_{flowers} = 46$, $p = .22$ for

the resume vignette; $U = 1016$, $N_{eyes} = 48$, $N_{flowers} = 48$, $p = .29$ for the wallet vignette). This time, the results of the resume vignette ratings went in the opposite direction of Bourrat et al., with a larger mean, median, and mode for participants in the eyes condition compared to the flowers condition. Descriptive statistics for the moral acceptability ratings, broken down by condition, are presented in Table 8.

Discussion

Our final experiment did not find significant differences between participants exposed to a surveillance cue and those exposed to a control image. This happened despite the fact that Experiment 4 was designed to be as similar to Bourrat et al.'s (2011) study as possible. Bourrat et al.'s participants completed their task in university libraries with little privacy, so participants in Experiment 4 did the same. Still, my results were null. Therefore, I have probably ruled out privacy and location as explanations for the differences in my results compared to those of Bourrat and colleagues.

All Experiments Combined

I combined the data from all four of my experiments and analyzed them. The faces conditions from Experiments 1 and 2 and the eyes condition from Experiments 3 and 4 were combined into a single surveillance cue condition. The other conditions were combined into a single control condition. A pair of Mann-Whitney U tests were conducted. The p-values were not significant for either the wallet vignette ($U = 142914$, $N_{surveillance} = 653$, $N_{control} = 447$, $p = .54$) or the resume vignette ($U = 140758$, $N_{surveillance} = 645$, $N_{control} = 449$, $p = .42$). The means and standard errors are displayed in Figure 4.

Experiments Discussion

Descriptive statistics for the moral acceptability scores for the surveillance cue and control conditions are presented in Tables 9 (Experiment 1), 10 (Experiment 2), 7 (Experiment 3), 8 (Experiment 4) and 11 (All experimental data combined). There is no consistent pattern in the data; sometimes moral acceptability scores are lower for the surveillance cue conditions, and sometimes they are lower for the control conditions. When looking at statistics for the entire combined data set, medians and modes are identical for the surveillance cue and control conditions, and the means are similar (Table 11).

I attempted to replicate the results from Bourrat et al. (2011) four times. The first two experiments were conceptual replications; the third and fourth were more "literal" replications. All four experiments achieved null results. I considered and ruled out multiple explanations for my lack of significant findings: 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, and experiment location.

If Bourrat and colleagues had truly demonstrated that an artificial surveillance cue could affect reported moral judgment, then I should have been able to replicate the effect. With a combined sample size of 1,100 participants (compared to 91 in Bourrat et al.'s experiment), my study had much more power. I believe the best explanation for Bourrat et al.'s significant results is a statistical false positive.

If artificial surveillance cues didn't affect the participants in my experiments, where does moral judgment fit into the surveillance cue literature at large? It's possible surveillance cues can, in fact, affect indicators of moral judgment; it may be the *rating of moral judgment in a survey* that isn't

affected. If artificial surveillance cues trigger automatic cognitive mechanisms that lead cued individuals to behave as if they are being watched, surveillance cues should have a stronger effect on behaviour that is easier to observe.

Support for this hypothesis comes from dictator games. Nettle and colleagues (2013) found no differences in mean donations between the eyes condition and the control condition in their dictator game experiment. However, a greater proportion of participants in the eyes condition donated something, rather than nothing, compared to participants in the control condition (80% versus 63%). Nettle and colleagues pointed out that this "increased probability of donation effect" was found in most published surveillance cue experiments involving the dictator game or a similar design (i.e., Haley & Fessler, 2005; Rigdon et al., 2009; Keller & Pfattheicher, 2011; Raihani & Bshary, 2012), whereas a mean donation effect was a less common outcome (but see Baillon et al., 2013). In other words, surveillance cues seem to increase the probability of dictators giving something, rather than nothing, more than they seem to increase the mean amount dictators give. In many real-life situations, the mere act of giving can be easily observed, whereas giving more doesn't necessarily look different to observers than giving less. Unfortunately, I do not know if the same pattern holds true for the charity field studies (e.g., Bateson et al., 2006; Ekström, 2012; Powell et al., 2012) because these studies do not have information on the actions of individuals.

Another study consistent with the notion that surveillance cues affect overt, highly visible actions was conducted by Nettle, Nott, and Bateson (2012). Rather than investigating positive, prosocial behaviour such as donating to charity or throwing away garbage, this study examined deterrence of antisocial behaviour. Large signs displaying watching eyes were placed at three bicycle

rack locations on a university campus; bike theft, a highly visible behaviour, decreased by 62% at these locations.

Finally, a field experiment conducted by Francey and Bergmüller (2012) provides mixed support for the idea that surveillance cues have their greatest impact on conspicuous behaviour. Trash was placed on bus stop benches in Geneva. People spent about twice as long disposing of garbage when images of eyes were displayed compared to when images of flowers were displayed. This effect seemed to be driven by participants in the eyes condition throwing away a greater number of items. Throwing away garbage in a public location is a highly noticeable behaviour, and it seems to have been affected by the surveillance cues in this experiment. On the other hand, the actual proportion of people who threw away trash did not differ significantly between groups, despite the fact that throwing away trash should be noticeably different from not throwing away trash.

Few published studies have investigated the effect of surveillance cues on survey responses. One of these studies involved the assessment of several variables important in social contexts (Carbon & Hesslinger, 2011). Participants completed questionnaires measuring self-reported tendencies to behave prosocially, to help others, and to seek revenge. Participants were also asked to indicate how often they washed their hands and how much free will they believed they had. Surveillance cues had no significant effects on these variables. Perhaps surveillance cues would have affected how likely participants were to actually behave prosocially, help others, seek revenge, and wash their hands – rather than their self-reported tendencies of these actions.

In the case of moral judgment, maybe surveillance cue effects would be found with a more conspicuous indicator of moral judgment. This may be an interesting topic for future research.

Surveillance Cues Meta-Analysis

I believe the most probable explanation for Bourrat et al.'s significant results is a false positive (Simmons, Nelson, & Simonsohn, 2011; Francis, 2012). This calls into question the reality of surveillance cue effects, not just in the context of moral judgment, but generally.

Looking at the surveillance cues literature, one may come to the conclusion that the watching eyes effect is well established. The literature of a given topic will be misleading if there are biases in reporting, analysis, and publication, however (Ioannidis & Trikalinos, 2007). Furthermore, many research practices lead to an increased probability of committing a Type I error (Simmons et al., 2011; John, Loewenstein, & Prelec, 2012).

Publication bias refers to the increased probability of a paper being published if it reports statistically significant results. There are a few reasons to suspect the surveillance cues literature suffers from publication bias. First, surveillance cue experiments are potentially quite easy and inexpensive to carry out (for example, I collected all the data for Experiment 4 in less than a day). This means there is less incentive for researchers who have obtained null results to go to the trouble of writing papers and attempting to publish them. Since it can be difficult to get null results published, it may not seem worth the trouble of attempting it when little was invested in the study to begin with. Second, the papers that were published with null surveillance cue results tend to report some kind of interesting finding in addition to the null results. For example, Fehr and Schneider (2010) found no evidence for a surveillance cue effect in their trust game, but they reported a "large effect" of making participants' previous game decisions known to the other players. Another experiment lacking evidence for surveillance cue effects involved a one-shot dictator game in which participants sat alone

in a dark, sound-proof room (Tane & Takezawa, 2011). The authors published null results but also suggested they had made an interesting finding: The darkness of the room negated the observation effect. Raihani and Bshary's (2012) dictator game experiment did not find a surveillance cue effect either; participants exposed to an image of eyes were *less* generous than participants exposed to an image of flowers. These results were contrary to the typical published claims of eye images increasing the generosity of dictators, but Raihani and Bshary were still able to report a significant finding since they described the outcome as the result of a positive effect of *flowers* on generosity. If studies like these exist, studies with nothing but null results should also exist. I suspect they do, but few are published.

Third, many of the published papers report mixed or qualified results, such as the surveillance cue effect appearing in men but not women (Rigdon et al., 2009) or among an ingroup but not an outgroup (Mifune et al., 2010), or surveillance cues having the opposite effect on participants high in prevention-focused self-regulation than they do on those low in prevention-focused self-regulation (Keller & Pfattheicher, 2011), or surveillance cues only working under certain conditions, such as when they are shown briefly (Sparks & Barclay, 2013) or when the number of people in an area is low (Ekström, 2012) or, conversely, high (Bateson et al., 2013), or surveillance cues causing people to spend more time throwing away trash but not increasing the proportion of people who throw away trash (Francey & Bergmüller, 2012). The conditional surveillance cue effects reported in these papers contradict findings from other papers and/or haven't been consistently replicated. When groups of participants are broken down into multiple sub-groups, covariates and moderating variables are considered (some of which are quite flexibly defined—what is a “larger number” of people?), and

several statistical analyses are made on a single data set, false positives are likely to appear.

For these reasons, I investigated the possibility of publication bias in the surveillance cues literature. I decided to do this with a meta-analysis and funnel plot. After I selected the studies for the meta-analysis, I noticed that only one of those studies (Keller & Pfattheicher, 2011) reported a significant replication. Lack of published replications may be a symptom of publication bias, so this was a fourth reason to move forward with a publication bias test.

Literature Search and Inclusion Criteria

As discussed previously in this paper, the surveillance cues literature describes the use of surveillance cues in various situations and their purported effects on different kinds of behaviour. A meta-analysis is only useful if the studies included examine the same phenomenon, however (Lipsey & Wilson, 2001). I chose to investigate the effect of surveillance cues on generosity since a large number of the published studies fit reasonably well into this category of behaviour (see Table 12).

More specifically, I analyzed studies that investigated the effect of visual, artificial, cues of being watched on the generosity of people. I operationalized generosity as *giving material resources to others without receiving material resources in return*. Material resources refers to money or objects that could be exchanged for money, such as points or tokens used in economic games, or lottery tickets in the case of one study (Tane & Takezawa, 2011). All of the studies were either economic games (mostly the dictator game) or involved donations to charity. The kind of observation cues I selected were those that resembled a watching face or eyes; they were generally photographs or stylized images of eyes. I was not interested in surveillance cues that may have led participants to believe a real human could see them, such as one-way mirrors or surveillance cameras. I also omitted studies in which

participants were shown images of their game partners or people who were purportedly their game partners (e.g., Burnham, 2003; Smith et al., 2009). Although these images may have served as good surveillance cues, they also introduced variables besides the cue of being watched by providing participants with information about their game partners that participants in control groups did not have. In a dictator game, for example, dictators provided with photos of their recipients may have cared more about their recipients than they would have if they had not seen the photos (the "identifiable victim effect": Schelling, 1968; Jenni & Loewenstein, 1997; Burnham, 2003). In studies like these, if dictators in the surveillance cue conditions gave their recipients more money than dictators in the control conditions, who's to say they didn't do so because they cared more about their recipients than the control group dictators? Since I was interested in surveillance cues specifically, I felt it was best to omit studies such as these for which the independent variable served as more than just a surveillance cue. Finally, I required studies for which the means, standard deviations, and sample sizes for the surveillance cue and control conditions were known. As explained below, these statistics were necessary for my analysis. Bateson et al. (2006) knew the size of their population but not the size of their sample, so I had to exclude their study, which otherwise would have been a good fit for the meta-analysis.

I excluded a proportion of the data from one study that met the inclusion criteria: Fehr and Schneider (2010). Fehr and Schneider (2010) described a series of one-shot trust games which consisted of two players, a trustor and a trustee. The trustor and trustee both began with ten points each. The trustor decided how many of his/her points to give to the trustee. The trustor could choose to give one point, four points, seven points, or all ten points. The trustee received the trustor's points

multiplied by four. If the trustor chose to send ten points, for example, the trustee received forty. The trustee then chose how many points to return to the trustor. The trustee could choose to return nothing, *compensate* the trustor by sending back the same amount of points the trustor had sent, or *equalize* by sending back the amount of points that would result in both players possessing an equal number. The researchers were interested in whether surveillance cues would affect the transfer decisions made by both trustors and trustees, so half of the trustors and half of the trustees were exposed to an image of stylized Horus eyes on their computer screen. For the meta-analysis, I included data on the points back-transferred by trustees but excluded data on the points transferred by trustors. I was only interested in the effect of surveillance cues on generosity; when evaluating the decisions of the trustors, there was no way to tease apart selfish from generous motives. A trustor may invest points in the hope of receiving a greater number of points when the trustee back-transfers. In other words, a trustor may give up points in the hope of eventually gaining a greater number of points than he/she gave up and thus benefiting with a net gain. Therefore, the number of points transferred by trustors is a poor operationalization of generosity.

I conducted a search on Web of Science, most recently on May 10, 2014, for studies involving artificial surveillance cues. I limited results to the following categories: behavioral sciences, psychology experimental, psychology, evolutionary biology, multidisciplinary sciences, psychology biological, psychology developmental, psychology multidisciplinary, psychology social, psychology applied, education educational research, psychology educational, social sciences interdisciplinary, sociology, economics, and anthropology. I used the search terms "cues of being watched" OR "eyespots" OR "social cues" OR "eye cues" OR "surveillance cues" OR "eye images" OR "eye-like"

OR "perception of human face" OR "watching eyes" OR "images of eyes" OR "observation cues" OR "cues of observation". I also did a search on PsycINFO, most recently on May 11, 2014, using the same search terms. The references cited by the papers I obtained with these searches provided two additional studies. My search strategy is outlined in Figure 5.

Calculation of Effect Sizes and Statistical Approach

I included 14 papers in the meta-analysis (see Table 12). For some of the papers, usable data were broken down by the researchers into more than one set (e.g., men and women; study 1 and study 2), so the total number of data points for the analysis was 18.

I merged conditions for two papers. Sparks and Barclay (2013) employed two distinct surveillance cue conditions; short exposure and prolonged exposure (explained in greater detail below). I combined these two surveillance cue conditions into one. Raihani and Bshary (2012) utilized four conditions: an image of eyes, an image of flowers, an image of a black square, and no image. I combined the latter three conditions into one control condition.

In the case of Sparks and Barclay (2013), there were two surveillance cue conditions but only one control condition. Rather than choosing one surveillance cue condition to compare with the control condition, I combined the surveillance cue conditions into one. I was interested in surveillance cues, generally, as opposed to different kinds of surveillance cues, so choosing one kind over the other would have been an arbitrary decision resulting in fewer data available for my meta-analysis. Sparks and Barclay made a reasonable a priori case for treating their two surveillance cue conditions differently, however, so I evaluate their argument later in this paper.

In the case of Raihani and Bshary (2012), there was only one surveillance cue condition but

three non-surveillance cue conditions. Again, I had the option of selecting a subset of conditions to compare. Although the participants in Raihani and Bshary's flowers condition were significantly more generous than participants shown a black square or no image, the large number of studies with opposite results (i.e., participants shown images of flowers were less generous than participants shown surveillance cues) leads me to believe Raihani and Bshary most likely obtained a false positive result, and in fact, the authors had expected the opposite result when they began. It is likely that flower conditions are not really different from other non-surveillance cue conditions. Therefore, I treated the flowers condition like a control condition, rather than an experimental condition, and merged it with the black square and no image conditions. The goal of my meta-analysis was to compare any surveillance cue conditions that met my criteria to any non-surveillance cue conditions; it was not to make distinctions between different kinds of surveillance cues or different kinds of control conditions.

The statistics I utilized were the means, standard deviations, and sample sizes for the surveillance cue conditions and the control conditions. These statistics were often reported in the papers; when they weren't, I asked the researchers for them. Some of the researchers sent me their raw data; in those cases, I calculated the statistics myself. For one of the studies (Fehr & Schneider, 2010), I extracted standard errors (from which the standard deviations were mathematically derived) from Figure 2a. I used two methods to do this: Data Thief and GNU Image Manipulation Program (GIMP). I decided to use the GIMP-derived standard errors due to the fact I was able to use a higher resolution image of the chart with GIMP than I could with the more commonly-used Data Thief. Correlations between the values derived from the two methods were quite high, however; $r(18) = .999$, $p < .001$.

The goal of my meta-analysis was to evaluate the possibility of publication bias. I followed the

procedure outlined by Lipsey and Wilson (2001). Each data point consists of a comparison between the mean amount of resources (usually money) given by participants in a study's surveillance cue condition and those given by participants in the control condition. For each comparison, I calculated the pooled standard deviation according to

$$s_p = \sqrt{\frac{(n_{G1} - 1)s_{G1}^2 + (n_{G2} - 1)s_{G2}^2}{(n_{G1} - 1) + (n_{G2} - 1)}} \quad (1)$$

where n_{G1} is the sample size for the eyes condition, n_{G2} is the sample size for the control condition, s^2_{G1} is the variance for the eyes condition, and s^2_{G2} is the variance for the control condition. Next, I calculated the standardized mean difference effect size for each data point according to

$$ES_{sm} = \frac{\overline{X}_{G1} - \overline{X}_{G2}}{s_p} \quad (2)$$

where $\overline{X}_{G1}$ is the mean for the eyes condition, $\overline{X}_{G2}$ is the mean for the control condition, and s_p is the pooled standard deviation as defined by Equation 1. Because this value of the effect size is biased, I used a correction (Hedges, 1981) to calculate an unbiased effect size according to

$$ES'_{sm} = \left[1 - \frac{3}{4N - 9}\right] ES_{sm} \quad (3)$$

where N is the total sample size and ES_{sm} is the biased standardized mean difference effect size as defined by Equation 2. The next step was to calculate the standard error of the standardized mean difference effect size according to

$$SE_{sm} = \sqrt{\frac{n_{G1} + n_{G2}}{n_{G1} n_{G2}} + \frac{(ES'_{sm})^2}{2(n_{G1} + n_{G2})}} \quad (4)$$

where ES'_{sm} is the unbiased standardized mean difference effect size, as defined by Equation 3.

Once these statistics were calculated for each comparison, I created a forest plot (see Figure 6) and a funnel plot (see Figure 7). The forest plot displays the effect size and 95% confidence intervals for the individual studies, which are organized by year of publication. A funnel plot, also known as a funnel graph diagram, is a kind of scatter plot that may be used to check a literature for publication bias (Light & Pillemer, 1984; Elvik, 1998; Sterne & Egger, 2001). A measure of effect size is plotted along the horizontal axis while the vertical axis measures some value of study size. If there is no bias in the literature, the data points should create the shape of a symmetrical funnel. Studies with smaller sample sizes form the wide part of the funnel because their imprecision results in a large amount of variance in effect size; studies with larger sample sizes make up the narrow part of the funnel due to greater accuracy. If a proportion of the studies with null results or negative effects aren't published, then the funnel will be lopsided. Asymmetry may also be an indication of small studies overestimating the true size of an effect (Sterne & Egger, 2001).

I plotted the unbiased standardized mean difference effect size along the x-axis. Sterne and Egger (2001) suggested that the standard error is the best statistic to plot along the vertical axis, so I placed the standard error of the standardized mean difference effect size on the y-axis. Simply looking at the funnel plot (Figure 7), one can see that the funnel is asymmetrical. As the standard error increases (and the studies become less reliable), variation in effect size increases, as expected, but the funnel mouth is wider on the positive end of the graph than it is on the negative end. This may indicate publication bias in favour of publishing positive results. Consider the three data points located outside of the 95% confidence intervals. The two with positive effect sizes have relatively large standard errors

(0.26 and 0.29) while the one with a negative effect size has a small standard error (0.12). Since studies with small standard errors are more reliable than studies with large standard errors, the fact that the outlier with a small standard error has a negative effect size implies that the funnel is missing a chunk of its negative side. It's likely the tip of the funnel would converge on an effect size close to zero if the negative studies which are probably sitting in various file drawers were included. The lopsided shape of the funnel plot is consistent with publication bias.

Another possible interpretation is that the true effect size has been exaggerated by smaller studies. Figure 7 shows the funnel plot with 95% confidence intervals; as the standard error decreases and the studies become more reliable, the confidence intervals narrow noticeably. Based only on published studies, the mean effect size is 0.065, with an unbiased standard error of 0.046 and a 95% confidence interval of -0.026 to 0.156. The fact that the confidence interval includes zero casts doubt on the reality of surveillance cue effects; moreover, if there is a publication bias in favour of significant positive results, then the mean effect size must be smaller than the one I have calculated. In summary, the effect of surveillance cues on generosity may be smaller than the literature indicates.

Sparks and Barclay (2013) posited that surveillance cues increase generosity so long as they are shown to participants only briefly. If the duration of surveillance cue exposure is key to bringing out an effect on behaviour, perhaps evidence for publication bias disappears when controlling for length of exposure. I modified my meta-analysis in order to investigate this possibility.

Sparks and Barclay reviewed surveillance cue studies and categorized the exposure of surveillance cues as short, ambiguous, or prolonged. They explained their categorization system as such:

> We classify a method as providing short exposure when eye stimuli are suddenly visible or attention is drawn to them shortly before the DV is measured. When eye stimuli are visible and in the participant's line of vision for several minutes, we classify the methods as a prolonged exposure. When stimuli are constantly visible but participant attention may be focused elsewhere, exposure length is ambiguous. (Sparks & Barclay, 2013, p. 318)

Most of the studies included in my meta-analysis utilized methods categorized by Sparks and Barclay as short exposure. I removed data from two papers with methods categorized by Sparks and Barclay as prolonged exposure (Fehr & Schneider, 2010; Tane & Takezawa, 2011). I removed one paper with methods classified by Sparks and Barclay as ambiguous exposure (Nettle et al., 2013). I removed one study that was published too late for inclusion in Sparks and Barclay's analysis but which had methods I determined to be prolonged exposure, due to the fact that the surveillance cues were visible to participants for the whole experiment (Baillon et al., 2013). I also modified one of my data points: Sparks and Barclay (2013) conducted an experiment of their own which included a short exposure surveillance cue condition, a prolonged exposure surveillance cue condition, and a control condition. For the general meta-analysis, I compared the control condition to a combined surveillance cue condition consisting of both short and prolonged exposure surveillance cue groups. For the "short exposure" version of the meta-analysis, I compared the control condition to just the short exposure surveillance cue condition. All of these modifications are noted in Table 12.

Once these modifications were made, the meta-analysis consisted only of comparisons (13 data points in all) in which participants were exposed to surveillance cues briefly. Consistent with Sparks and Barclay's hypothesis, the grand mean effect size based on short exposure studies is larger than the grand mean effect size based on all the meta-analysis studies (0.108 [unbiased standard error of 0.052] versus 0.065 [unbiased standard error of 0.046]). Furthermore, the 95% confidence interval is 0.006 to

0.211, which does not include zero. On the other hand, a funnel plot based on the modified meta-analysis (presented in Figure 8) looks even more biased in favour of significant surveillance cue effects than the general funnel plot. The lopsided short exposure funnel plot is consistent with publication bias, which means the grand mean effect size and the confidence intervals may be unreliable.

Sparks and Barclay (2013) presented a list of 14 short exposure studies and indicated that 13 of them found a surveillance cue effect. This may seem inconsistent with my modified short-exposure funnel plot which displays 5 data points with negative effect sizes. This apparent contradiction is a result of differing methods, however. Mifune et al. (2010) reported a surveillance cue effect for participants in an ingroup condition, but not for an outgroup condition. This resulted in two data points on my funnel plot, one with a positive effect size and one with a negative effect size; Sparks and Barclay indicated that an effect of eyes was found for the study. Rigdon et al. (2009) reported a surveillance cue effect for men, providing a positive data point for my funnel plot, but not for women, providing a negative data point for my funnel plot; Sparks and Barclay indicated that an effect of eyes was found for the study. Keller and Pfattheicher (2011) reported significant interaction effects between surveillance cues and prevention focus scores for two studies. Comparing the surveillance cue conditions to the control conditions without considering interactions, which my meta-analysis was not concerned with, resulted in a positive effect size for the first study and a negative effect size for the second study; Sparks and Barclay indicated that an effect of eyes was found. In Ekström's (2012) supermarket study, there was no overall effect of surveillance cues, but the authors reported a significant effect on slow days (i.e., when few customers visited the stores). This yielded a negative effect size for my funnel plot, while Sparks and Barclay indicated that an effect of eyes was found for

the study. Finally, Raihani and Bshary (2012) produced a negative effect size on my funnel plot because participants in the surveillance cue condition gave less money than participants in the control conditions. This makes sense because the dependent measure evaluated by my meta-analysis was mean amount given. On the other hand, there was an increased probability of giving for participants in Raihani and Bshary's surveillance cue condition, so Sparks and Barclay indicated that an effect of eyes was found for the study.

Again, I'd like to point out that these reported surveillance cue effects are conditional and the conditions are inconsistent (i.e., different from one study to the next). These studies have many chances to produce significant results because they employ subgroups and moderating variables. Examining only the short exposure studies that were included in both my meta-analysis and Sparks and Barclay's list, a surveillance cue effect was indicated by Sparks and Barclay whenever the results were mixed or conditional. This rule for evaluating research surely portions a greater degree of confidence in the reliability of surveillance cue effects than is warranted. I concede that this method, while rather lenient, should not prevent a fair investigation of surveillance cue exposure as a moderating variable so long as the same criteria are used for prolonged and ambiguous exposure studies. Unfortunately, there are few surveillance cue studies that are something other than short exposure; Sparks and Barclay (2013) listed five ambiguous and four prolonged exposure studies, one of which did not employ artificial surveillance cues (Lamba & Mace, 2010), leaving three. The small number of prolonged exposure studies makes it difficult to interpret the role surveillance cue duration plays. Because short exposure studies represent such a large portion of the literature, short exposure studies may give the appearance of demonstrating true effects for the same reasons surveillance cue studies in general give

the appearance of demonstrating true effects: publication bias, the file-drawer effect, multiple analyses on single data sets, and the labeling of studies with mixed and conditional results as hits. The length of surveillance cue exposure may be a good variable to investigate further, but the evidence that surveillance cues can affect behaviour when they are visible briefly is currently tenuous. Based on the evidence I have presented, I believe that publication bias is a more likely explanation for the inconsistency in surveillance cue effects than duration of surveillance cue exposure.

General Discussion

Inspired by Bourrat and colleagues' (2011) study of the effect of surveillance cues on moral judgment, I conducted an experiment investigating the effect of surveillance cues on self-judged positive traits, religiosity, and moral judgment. I found no evidence for an effect on any of these variables. The lack of effect for moral judgment was puzzling, since Bourrat and colleagues' paper suggested the reality of this phenomenon. I conducted three more experiments, each increasingly similar in design to Bourrat et al. (2011), and thus systematically tested variables in search of an explanation for why Bourrat and colleagues obtained significant results and I didn't. None of my experiments replicated the surveillance cue effect. I suggest the most likely explanation for the disparate results is that Bourrat and colleagues obtained a false positive.

Once the surveillance cue effect on reported moral judgment was called into question, the effect of surveillance cues on other variables became suspect. I conducted a meta-analysis of studies investigating surveillance cues in the context of generosity. The results of my meta-analysis may indicate publication bias in favour of publishing significant results; they may also be a sign that the surveillance cue effect on generosity, though perhaps a real phenomenon, is smaller than the literature

implies.

My experiments were based on surveys, which involve subtle behaviour such as circling numbers on a piece of paper. Perhaps surveillance cues are more likely to work on highly noticeable behaviour. In dictator games, differences between surveillance cue and control conditions are usually found in the likelihood of participants giving something rather than nothing; they are less often found in the mean amounts of money given (Nettle et al., 2013). In a real-life situation, the mere act of giving is often easily observed, whereas giving more doesn't necessarily look different to observers than giving less. If surveillance cues trigger reputation-enhancing behaviour, then maybe surveillance cues have the potential to affect indications of moral judgment that are highly visible to onlookers and therefore capable of enhancing the reputation of the actor; good deeds cannot enhance reputations if they go unnoticed.

The conclusions that can be drawn from my meta-analysis are limited. First, I only considered the effect of surveillance cues on generosity; the results of my meta-analysis cannot be extrapolated to other dependent measures, such as theft or littering. Furthermore, generosity was measured in one specific way (i.e., the mean amount given); other ways of measuring generosity, such as the probability of giving, were not evaluated. Second, a funnel plot only allows for subjective judgments. It is my judgment that the shape of the funnel plot is consistent with publication bias, but this isn't proof of publication bias and readers may disagree with my interpretation. Third, my funnel plot only consists of 18 data points; a funnel plot comprised of a larger data set would be easier to interpret. Due to the limitations of my meta-analysis, the application of additional meta-analyses, involving methods different from those described here, could shed more light on the question of publication bias in the

surveillance cues literature (Ioannidis & Trikalinos, 2007).

The surveillance cue studies discussed in this paper generally compared two groups of people: participants who were shown artificial surveillance cues and participants who weren't. I would like to see experiments in which participants who are exposed to artificial surveillance cues are compared with participants who are exposed to real surveillance cues (e.g., actual people). This would allow researchers to compare the effect of artificial surveillance cues to the effect of genuine surveillance cues.

I have presented evidence that the surveillance cue effect is smaller than the literature indicates, and that it may not exist in the case of survey reports of moral judgment. In the future, I hope to see research conducted in a more rigorous fashion. If studies are conducted with larger sample sizes, the results will be more reliable. Null results should be published, regardless of whether or not researchers report interesting findings along with those null results. I also suggest that researchers try to replicate new findings before publishing. The goal of science is to determine truth to the best of our abilities, not to publish as many papers as possible.

References

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

Aiello, J. R., & Svec, C. M. (1993). Computer monitoring and work performance: Extending the social facilitation framework to electronic presence. *Journal of Applied Social Psychology, 23,* 537-548. doi:10.1111/j.1559-1816.1993.tb01102.x

Alexander, R. D. (1987). *The biology of moral systems*. Hawthorne, New York: A. de Gruyter.

Axelrod, R. M. (1984). *The evolution of cooperation*. New York: Basic Books.

*Baillon, A., Selim, A., & van Dolder, D. (2013). On the social nature of eyes: The effect of social cues in interaction and individual choice tasks. *Evolution and Human Behavior, 34,* 146-154. doi:10.1016/j.evolhumbehav.2012.12.001

Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B, 274,* 749-753. doi:10.1098/rspb.2006.0209

Bateson, M., Callow, L., Holmes, J. R., Redmond Roche, M. L., & Nettle, D. (2013). Do images of 'watching eyes' induce behaviour that is more pro-social or more normative? A field experiment on littering. *PLoS ONE, 8*(12), e82055. doi:10.1371/journal.pone.0082055

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

Bond, C. F., Jr., & Titus, L. J. (1983). Social facilitation: A meta-analysis of 241 studies. *Psychological Bulletin, 94*(2), 265-292. doi:10.1037/0033-2909.94.2.265

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

Brambilla, M., Rusconi, P., Sacchi, S., Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology, 41,* 135-143. doi:10.1002/ejsp.744

Burnham, T. C. (2003). Engineering altruism: A theoretical and experimental investigation of anonymity and gift giving. *Journal of Economic Behavior & Organization, 50,* 133-144. doi:10.1016/S0167-2681(02)00044-6

*Burnham, T. C., & Hare, B. (2007). Engineering human cooperation. *Human Nature, 18*, 88-108. doi:10.1007/s12110-007-9012-2

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. doi:10.1024/1421-0185/a000058

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

*Ekström, M. (2012). Do watching eyes affect charitable giving? Evidence from a field experiment. *Experimental Economics, 15,* 530-546. doi:10.1007/s10683-011-9312-6

Elvik, R. (1998). Evaluating the statistical conclusion validity of weighted mean results in meta-analysis by analyzing funnel graph diagrams. *Accident Analysis and Prevention, 30,* 255-266. doi:10.1016/S0001-4575(97)00076-6

Emery, N. J. (2000). The eyes have it: The neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews, 24,* 581-604. doi:10.1016/S0149-7634(00)00025-7

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. doi:10.1016/j.evolhumbehav.2010.10.006

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.

*Fehr, E., & Schneider, F. (2010). Eyes are on us, but nobody cares: are eye cues relevant for strong reciprocity? *Proceedings of the Royal Society B: Biological Sciences, 277*, 1315-1323. doi:10.1098/rspb.2009.1900

Francey, D., & Bergmüller, R. (2012). Images of eyes enhance investments in a real-life public good. *PLoS ONE, 7*(5): e37397. doi:10.1371/

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review, 19*(6), 975-991. doi:10.3758/s13423-012-0322-y

Gelfand, D. M., Hartmann, D. P., Walder, P., & Page, B. (1973). Who reports shoplifters? A field-experimental study. *Journal of Personality and Social Psychology, 25*(2), 276-285. doi:10.1037/h0033953

George, N., Dolan, R. J., Fink, G. R., Baylis, G. C., Russell, C., & Driver, J. (1999). Contrast polarity and face recognition in the human fusiform gyrus. *Nature Neuroscience, 2,* 574-580. doi:10.1038/9230

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. doi:10.1037/a0025882

*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. doi:10.1016/j.evolhumbehav.2005.01.002

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107-128. doi:10.3102/10769986006002107

Hoffman, E. A., & Haxby, J. V. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature Neuroscience, 3,* 80-84. doi:10.1038/71152

Horberg, E. J., Oveis, C., Keltner, D., & Cohen, A. B. (2009). Disgust and the moralization of purity. *Journal of Personality and Social Psychology, 97*(6), 963-976. doi:10.1037/a0017423

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials, 4,* 245-253. doi:10.1177/1740774507079441

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.

Jenni, K. E., & Loewenstein, G. (1997). Explaining the "identifiable victim effect". *Journal of Risk and Uncertainty, 14,* 235-257. doi:10.1023/A:1007740225484

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532. doi:10.1177/0956797611430953

*Keller, J., & Pfattheicher, S. (2011). Vigilant self-regulation, cues of being watched and

cooperativeness. *European Journal of Personality, 25,* 363-372. doi:10.1002/per.797

Kimball, S., Mattis, P., & the GIMP Development Team. (2013). GNU Image Manipulation Program (Version 2.8.4) [Computer software]. Retrieved from http://www.gimp.org

Korte, C., & Kerr, N. (1975). Response to altruistic opportunities in urban and nonurban settings. *The Journal of Social Psychology, 95*(2), 183-184. doi: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. doi:10.1023/A:1012563421824

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior, 28*, 75-84. doi:10.1016/j.evolhumbehav.2006.06.001

Lamba, S. & Mace, R. (2010). People recognize when they are really anonymous in an economic game. *Evolution and Human Behavior, 31,* 271-278. doi:10.1016/j.evolhumbehav.2010.02.002

Latané, B. (1981). The psychology of social impact. *The American Psychologist, 36,* 343-356. doi:10.1037//0003-066X.36.4.343

Leach, C. W., Ellemers, N., Barreto, M. (2007). Group virtue: the importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology, 93,* 234-249. doi:10.1037/0022-3514.93.2.234

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. doi: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. doi:10.1016/j.jesp.2009.10.017

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research.* Cambridge,

Massachusetts: Harvard University Press.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, California: Sage Publications.

*Mifune, N., Hashimoto, H., & Yamagishi, T. (2010). Altruism toward in-group members as a reputation mechanism. *Evolution and Human Behavior, 31*, 109-117. doi:10.1016/j.evolhumbehav.2009.09.004

Milinski, M., Semmann, D., & Krambeck, H. J. (2002). Donors to charity gain in both indirect reciprocity and political reputation. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 269*(1494), 881–883. doi:10.1098/rspb.2002.1964

*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. doi:10.1016/j.evolhumbehav.2012.08.004

Nettle, D., Nott, K., & Bateson, M. (2012). 'Cycle thieves, we are watching you': Impact of a simple signage intervention against bicycle theft. *PLoS ONE, 7*(12), e51738. doi:10.1371/journal.pone.0051738

Nowak, M. A., & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature, 393,* 573-577. doi:10.1038/31225

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437,* 1291-1298. doi:10.1038/nature04131

*Oda, R., Niwa, Y., Honma, A., & Hiraishi, K. (2011). An eye-like painting enhances the expectation of a good reputation. *Evolution and Human Behavior, 32,* 166-171.

doi:10.1016/j.evolhumbehav.2010.11.002

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

Panagopoulos, C. (2014). I've got my eyes on you: Implicit social-pressure cues and prosocial behavior. *Political Psychology, 35*(1), 23-33. doi:10.1111/pops.12074

*Powell, K. L., Roberts, G., & Nettle, D. (2012). Eye images increase charitable donations: Evidence from an opportunistic field experiment in a supermarket. *Ethology, 118,* 1096-1101. doi:10.1111/eth.12011

Puce, A., Allison, T., Bentin, S., Gore, J. C., & McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *The Journal of Neuroscience, 18,* 2188-2199. Retrieved from http://www.jneurosci.org/content/18/6/2188.full.pdf

Putz, V. R. (1975). The effects of different modes of supervision on vigilance behaviour. *British Journal of Psychology, 66*(2), 157-160. doi:10.1111/j.2044-8295.1975.tb01449.x

*Raihani, N. J., & Bshary, R. (2012). A positive effect of flowers rather than eye images in a large-scale, cross-cultural dictator game. *Proceedings of the Royal Society B, 279,* 3556-3564. doi:10.1098/rspb.2012.0758

*Rigdon, M., Ishii, K., Watabe, M., & Kitayama, S. (2009). Minimal social cues in the dictator game. *Journal of Economic Psychology, 30*, 358-367. doi:10.1016/j.joep.2009.02.002

Risko, E. F., & Kingstone, A. (2011). Eyes wide shut: Implied social presence, eye tracking and attention. *Attention, Perception, & Psychophysics, 73,* 291-296. doi:10.3758/s13414-010-0042-1

Rushton, J. P. (1978). Urban density and altruism: Helping strangers in a Canadian city, suburb, and small town. *Psychological Reports, 43*(3), 987-990. doi:10.2466/pr0.1978.43.3.987

Satow, K. (1975). Social approval and helping. *Journal of Experimental Social Psychology, 11,* 501-509. doi:10.1016/0022-1031(75)90001-3

Schelling, T. C. (1968). The life you save may be your own. In S. Chase (Ed.), *Problems in Public Expenditure Analysis*. Washington DC: The Brookings Institute.

Schmitz, C. (2012). LimeSurvey (Version 2.00) [Computer software]. Hamburg, Germany: LimeSurvey Project Team. Retrieved from http://www.limesurvey.org

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. doi:10.1177/0146167208317771

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. doi:10.1177/0956797611417632

Smead, R. (2010). Indirect reciprocity and the evolution of "moral signals". *Biology & Philosophy, 25,* 33-51. doi:10.1007/s10539-009-9175-9

Smith, F. G., Debruine, L. M., Jones, B. C., Krupp, D. B., Welling, L. L. M., & Conway, C. A. (2009). Attractiveness qualifies the effect of observation on trusting behavior in an economic game. *Evolution and Human Behavior, 30,* 393-397. doi:10.1016/j.evolhumbehav.2009.06.003

Sparks, A. (2010). *Subtle cues and economic games* (Master's thesis, McMaster University, Hamilton, Ontario). Retrieved from http://digitalcommons.mcmaster.ca/opendissertations/4321

*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. doi:10.1016/j.evolhumbehav.2013.05.001

Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology, 54,* 1046-1055. doi:10.1016/S0895-4356(01)00377-8

Tan, J. H. W., & Vogel, C. (2008). Religion and trust: An experimental study. *Journal of Economic Psychology, 29*, 832-848. doi:10.1016/j.joep.2008.03.002

*Tane, K., & Takezawa, M. (2011). Perception of human face does not induce cooperation in darkness. *Letters on Evolutionary Behavioral Science, 2,* 24-27. doi:10.5178/lebs.2011.15

Triplett, N. (1898). The Dynamogenic Factors in Pacemaking and Competition. *The American Journal of Psychology, 9*(4), 507-533. doi:10.2307/1412188

Tummers, B. (2010). DataThief III (Version 1.6) [Computer software]. Retrieved from http://download.cnet.com/DataThief/3000-2053_4-80439.html

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. doi:10.1177/0013916507309996

Wedekind, C., & Braithwaite, V. A. (2002). The long-term benefits of human generosity in indirect reciprocity. *Current Biology, 12,* 1012-1015. doi:10.1016/S0960-9822(02)00890-4

Wedekind, C., & Milinski, M. (2000). Cooperation through image scoring in humans. *Science, 288*, 850-852. doi:10.1126/science.288.5467.850

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological*

Science, 16(10), 780-784. doi:10.1111/j.1467-9280.2005.01614.x

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. doi:10.1177/0022022195265002

Zajonc, R. B. (1965). Social facilitation. *Science, 149*(3681), 269-274. doi:10.2307/1715944

Figure 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.

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

Figure 3. Images used for the eyes and flowers conditions in Experiments 3 and 4. These images were initially used by Bourrat et al. (2011).

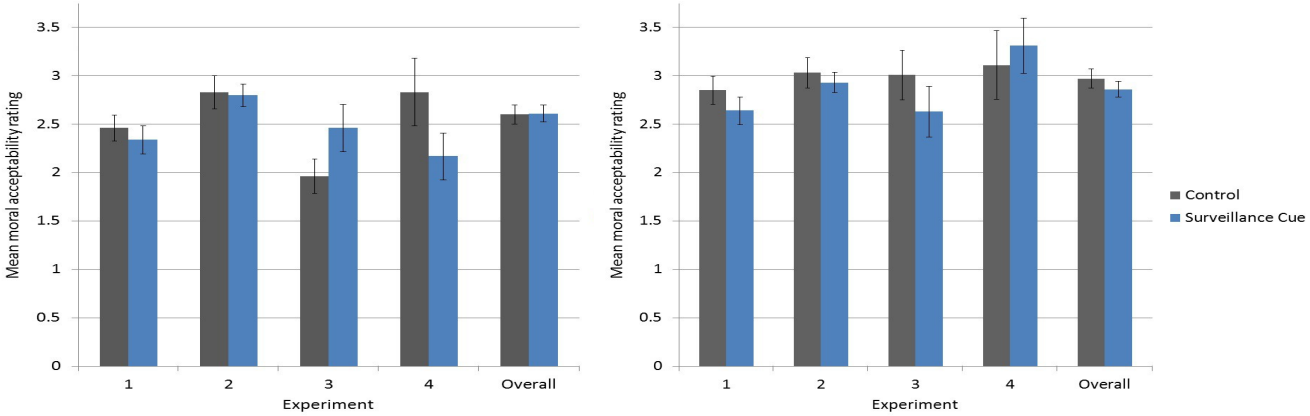

Figure 4. Bar plot of mean moral acceptability ratings for participants in the surveillance cue and control conditions for Experiments 1 – 4 and for all four experiments combined. Ratings for the wallet vignette are on the left; ratings for the resume vignette are on the right. Error bars represent one standard error above and below the mean.

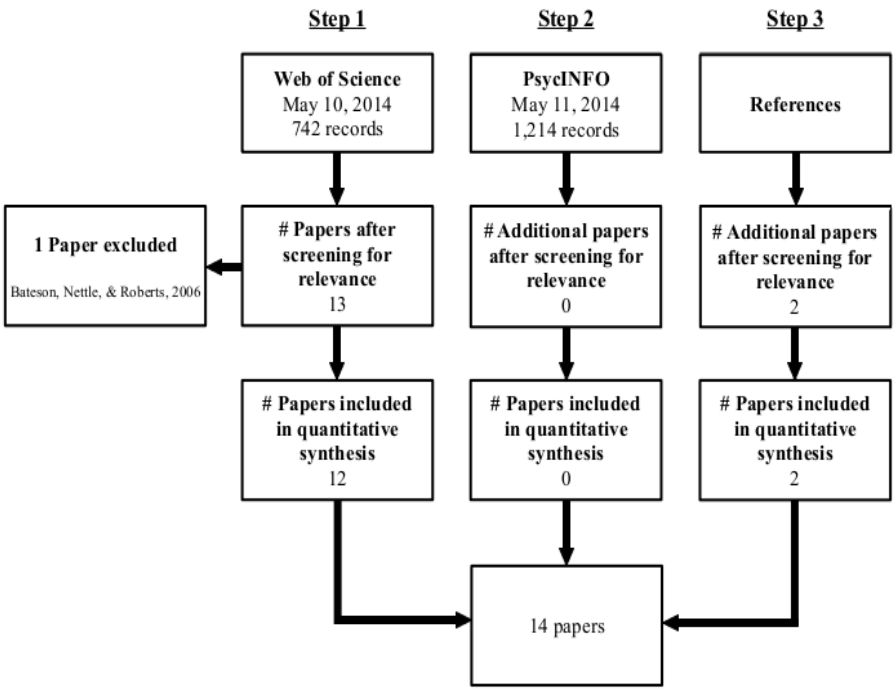

Figure 5. Meta-analysis search strategy flow diagram.

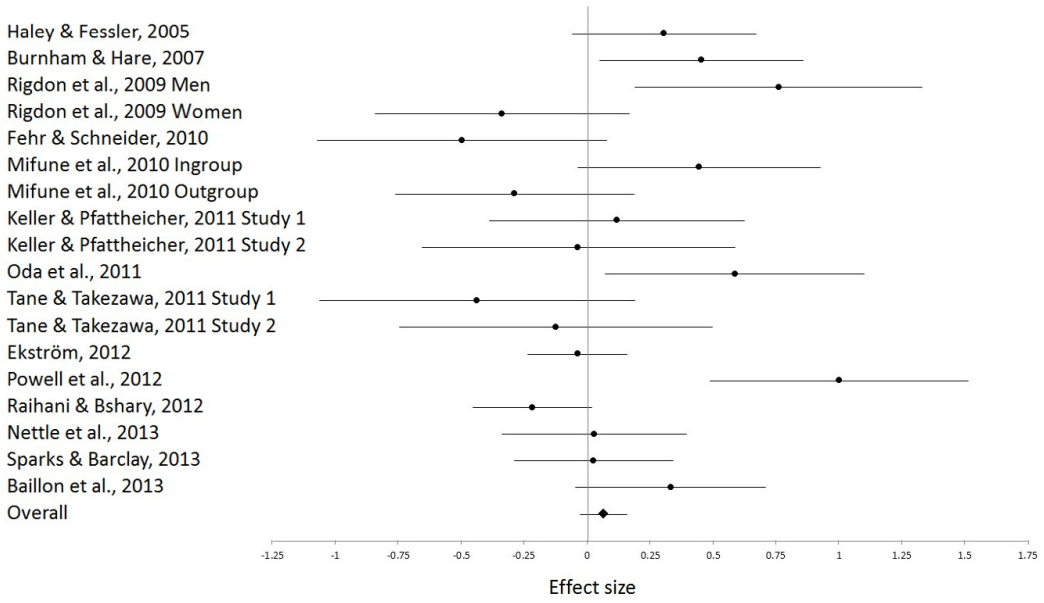

Figure 6. Forest plot of the unbiased standardized mean difference effect size and the 95% confidence intervals for the studies comprising the meta-analysis.

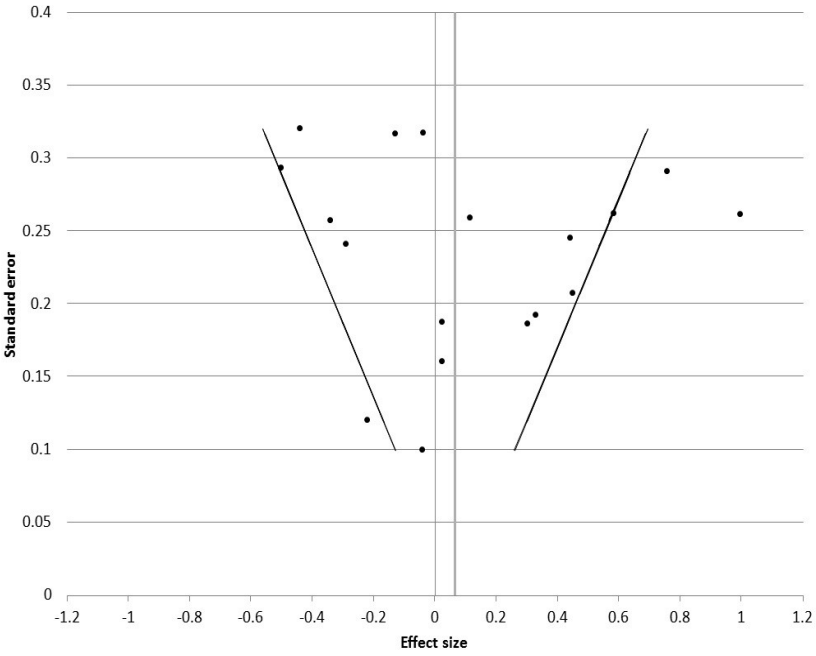

Figure 7. Funnel plot of meta-analysis studies with confidence intervals. The unbiased standardized mean difference effect size is plotted along the horizontal axis; the unbiased standard error of the standardized mean difference effect size is plotted along the vertical axis. The weighted mean effect size (0.065) is represented by the vertical line. The diagonal lines indicate the expected 95% confidence intervals around the mean effect size.

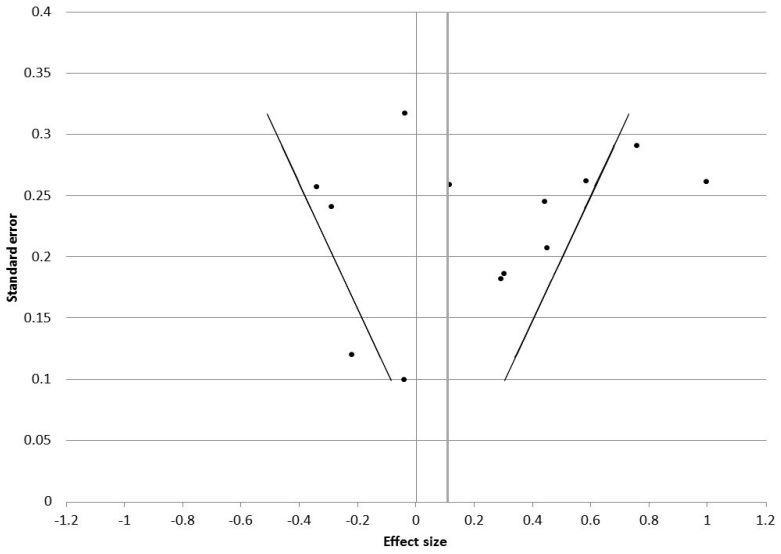

Figure 8. Funnel plot of short exposure meta-analysis studies. The unbiased standardized mean difference effect size is plotted along the horizontal axis; the unbiased standard error of the standardized mean difference effect size is plotted along the vertical axis. The weighted mean effect size (0.108) is represented by the vertical line. The diagonal lines indicate the expected 95% confidence intervals around the mean effect size.

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*	*SE*	*n*	*M*	*Mdn*	Mode	*SD*	*SE*	*n*
Familiar Face	65.35	65.50	64	7.71	0.86	80	15.19	16.00	9	6.83	0.75	84
Unfamiliar Face	66.32	67.00	64	8.10	0.89	82	15.68	17.00	18	7.28	0.79	85
Chair	67.26	67.00	66	7.13	0.78	84	14.98	16.00	19	7.24	0.79	84
No Image	63.64	65.00	68	8.56	0.94	83	14.86	15.00	13,15	6.42	0.71	83
Female Face	65.80	66.00	66	7.83	0.88	79	15.05	16.00	17,18	7.04	0.76	85
Male Face	65.88	66.00	64	8.01	0.88	83	15.83	17.00	19	7.07	0.77	84

Table 2

Experiment 1: Moral Acceptability Ratings for the Familiar Face, Unfamiliar Face, Chair, No Image, Female Face, and Male Face Conditions

	Wallet Vignette						Resume Vignette					
Condition	*M*	*Mdn*	Mode	*SD*	*SE*	*n*	*M*	*Mdn*	Mode	*SD*	*SE*	*n*
Familiar Face	2.43	2.00	1	1.87	0.20	84	2.60	2.00	1,2,3	1.61	0.18	84
Unfamiliar Face	2.25	1.00	1	1.91	0.21	85	2.69	2.00	1	2.06	0.22	85
Chair	2.26	2.00	1	1.64	0.18	84	2.95	3.00	2	1.69	0.18	84
No Image	2.65	2.00	1	1.78	0.19	85	2.75	2.00	1	2.04	0.22	85
Female Face	2.54	2.00	1	1.90	0.21	85	2.60	2.00	1,3	1.75	0.19	85
Male Face	2.13	1.00	1	1.86	0.20	84	2.69	2.00	1	1.95	0.21	84

Table 3

Experiment 1: Moral Acceptability Ratings for Men and Women

	Wallet Vignette						Resume Vignette					
Gender	*M*	*Mdn*	Mode	*SD*	*SE*	*n*	*M*	*Mdn*	Mode	*SD*	*SE*	*n*
Men	3.06	3.00	1	2.25	0.25	83	3.13	3.00	1	2.38	0.26	83
Women	2.15	2.00	1	1.52	0.10	253	2.61	2.00	1	1.62	0.10	253

Table 4

Experiment 2: Moral Acceptability Ratings for the Female Face, Male Face, and Chair Conditions

	Wallet Vignette						Resume Vignette						Infidelity Vignette					
Condition	*M*	*Mdn*	Mode	*SD*	*SE*	*n*	*M*	*Mdn*	Mode	*SD*	*SE*	*n*	*M*	*Mdn*	Mode	*SD*	*SE*	*n*
Female Face	2.83	2.00	1	2.25	0.17	187	2.99	3.00	1	2.07	0.15	180	1.84	1.00	1	1.63	0.12	181
Male Face	2.77	2.00	1	2.32	0.16	203	2.87	2.00	1	1.98	0.14	202	2.03	1.00	1	1.83	0.13	204
Chair	2.83	2.00	1	2.28	0.17	183	3.03	2.00	1	2.14	0.16	187	1.72	1.00	1	1.47	0.11	187

Table 5

Experiment 2: Predictors of Moral Acceptability Using Poisson Loglinear Generalized Linear Models

	Wallet Vignette				Resume Vignette				Infidelity Vignette			
Predictor	b	*SE*	Wald χ^2	p-value	b	*SE*	Wald χ^2	p-value	b	*SE*	Wald χ^2	p-value
Intercept	0.70	0.13	30.94	<.001	0.88	0.13	49.47	<.001	0.29	0.15	3.81	.051
Gender[a]	-0.20	0.07	8.66	.003	-0.15	0.07	5.01	.025	-0.09	0.08	1.23	.268
Attractive	0.01	0.02	0.08	.781	0.05	0.02	7.82	.005	0.02	0.02	0.80	.372
Authoritative	0.10	0.04	7.99	.005	-0.05	0.04	2.13	.145	0.01	0.04	0.05	.831
Judgmental	0.01	0.04	0.14	.707	-0.03	0.04	0.51	.476	0.04	0.04	0.67	.413
Disgusting	0.13	0.03	20.48	<.001	0.11	0.03	14.61	<.001	0.19	0.03	31.08	<.001

[a]Women, compared to men.

Table 6

Experiment 2: Moral Acceptability Ratings for Men and Women

	Wallet Vignette						Resume Vignette						Infidelity Vignette					
Gender	*M*	*Mdn*	Mode	*SD*	*SE*	*n*	*M*	*Mdn*	Mode	*SD*	*SE*	*n*	*M*	*Mdn*	Mode	*SD*	*SE*	*n*
Men	3.06	2.00	1	2.36	0.13	319	3.11	2.50	1	2.08	0.12	322	1.94	1.00	1	1.71	0.10	323
Women	2.45	1.00	1	2.11	0.14	234	2.69	2.00	1	1.99	0.13	228	1.75	1.00	1	1.58	0.10	232

Table 7

Experiment 3: Moral Acceptability Ratings for the Control and Surveillance Cue Conditions

	Wallet Vignette						Resume Vignette					
Condition	*M*	*Mdn*	Mode	*SD*	*SE*	*n*	*M*	*Mdn*	Mode	*SD*	*SE*	*n*
Control	1.96	2.00	1	1.20	0.17	47	3.01	3.00	2	1.76	0.26	47
Surveillance	2.46	2.00	1	1.63	0.24	46	2.63	2.00	1	1.77	0.26	46

Table 8

Experiment 4: Moral Acceptability Ratings for the Control and Surveillance Cue Conditions

	Wallet Vignette						Resume Vignette					
Condition	*M*	*Mdn*	Mode	*SD*	*SE*	*n*	*M*	*Mdn*	Mode	*SD*	*SE*	*n*
Control	2.83	2.00	1	2.40	0.35	48	3.11	2.00	1	2.38	0.35	46
Surveillance	2.17	1.50	1	1.67	0.24	48	3.31	3.00	2,3	1.96	0.28	48

Table 9

Experiment 1: Moral Acceptability Ratings for the Control and Surveillance Cue Conditions

	Wallet Vignette						Resume Vignette					
Condition	*M*	*Mdn*	Mode	*SD*	*SE*	*n*	*M*	*Mdn*	Mode	*SD*	*SE*	*n*
Control	2.46	2.00	1	1.72	0.13	169	2.85	2.00	1	1.87	0.14	169
Surveillance	2.34	1.00	1	1.88	0.14	169	2.64	2.00	1	1.85	0.14	169

Table 10

Experiment 2: Moral Acceptability Ratings for the Control and Surveillance Cue Conditions

	Wallet Vignette						Resume Vignette						Infidelity Vignette					
Condition	*M*	*Mdn*	Mode	*SD*	*SE*	*n*	*M*	*Mdn*	Mode	*SD*	*SE*	*n*	*M*	*Mdn*	Mode	*SD*	*SE*	*n*
Control	2.83	2.00	1	2.28	0.17	183	3.03	2.00	1	2.14	0.16	187	1.72	1.00	1	1.47	0.11	187
Surveillance	2.80	2.00	1	2.29	0.12	390	2.93	2.00	1	2.02	0.10	382	1.94	1.00	1	1.74	0.09	385

Table 11

All Experiment Data Combined: Moral Acceptability Ratings for the Control and Surveillance Cue Conditions

	Wallet Vignette						Resume Vignette					
Condition	*M*	*Mdn*	Mode	*SD*	*SE*	*n*	*M*	*Mdn*	Mode	*SD*	*SE*	*n*
Control	2.60	2.00	1	2.02	0.10	447	2.97	2.00	1	2.03	0.10	449
Surveillance	2.61	2.00	1	2.12	0.08	653	2.86	2.00	1	1.96	0.08	645

Table 12

Statistics for Surveillance Cue and Control Condition Comparisons From Studies Included in Meta-Analysis

		Surveillance			Control				
Study	Design	*M*	*SD*	*n*	*M*	*SD*	*n*	UES	USE
Haley & Fessler (2005)	Dictator game	3.14	2.33	77	2.38	2.74	47	0.30	0.19
Burnham & Hare (2007)	Public goods game	6.75	3.08	48	5.25	3.50	48	0.45	0.21
Rigdon et al. (2009) Men	Dictator game	3	2.04	24	1.41	2.08	27	0.76	0.29
Rigdon et al. (2009) Women	Dictator game	2.12	1.95	34	2.79	1.95	28	-0.34	0.26
Fehr & Schneider (2010) Trustees[a]	Trust game	5.36	1.91	24	6.28	1.71	24	-0.50	0.29
Mifune et al. (2010) Ingroup	Dictator game	421.84	135.54	38	357.10	153.91	31	0.44	0.24
Mifune et al. (2010) Outgroup	Dictator game	325.25	185.32	40	383.87	220.00	31	-0.29	0.24
Keller & Pfattheicher (2011) Study 1	Charity donation	1.05	1.02	30	0.93	0.94	30	0.12	0.26
Keller & Pfattheicher (2011) Study 2	Charity donation	0.91	1.15	21	0.95	1.10	19	-0.04	0.32
Oda et al. (2011)	Dictator game	330	178.4	30	238.7	125.6	31	0.59	0.26
Tane & Takezawa (2011) Study 1[a]	Dictator game	2.6	1.98	20	3.45	1.82	20	-0.44	0.32
Tane & Takezawa (2011) Study 2[a]	Dictator game	3.05	2.24	20	3.35	2.41	20	-0.13	0.32
Ekström (2012)	Charity donation	0.06	0.05	186	0.06	0.06	222	-0.04	0.10
Powell et al. (2012)	Charity donation	7.9	2.12	33	5.48	2.64	33	1.00	0.26
Raihani & Bshary (2012)	Dictator game	0.15	0.10	92	0.17	0.11	295	-0.22	0.12
Nettle et al. (2013)[b]	Dictator game	1.58	1.25	69	1.54	1.74	49	0.03	0.19
Baillon et al. (2013)[a]	Dictator game	13.93	12.66	55	9.75	12.48	55	0.33	0.19
Sparks & Barclay (2013)[c]	Dictator game	4.48	1.93	133	4.44	1.61	55	0.02	0.16
Sparks & Barclay (2013) Short Exposure[d]	Dictator game	4.96	1.87	69	4.44	1.61	55	0.29	0.18

Note. UES = unbiased effect size; USE = unbiased standard error.

[a] These study's methods provided prolonged exposure using the criteria of Sparks and Barclay (2013), and were therefore excluded from the short exposure version of the meta-analysis.

[b] This study's methods provided ambiguous exposure using the criteria of Sparks and Barclay (2013), and was therefore excluded from the short exposure version of the meta-analysis.

[c] Sparks and Barclay (2013) included three conditions: a short exposure surveillance cue condition, a prolonged exposure surveillance cue condition, and a control condition. For the general meta-analysis, I compared the control condition to a combined surveillance cue condition consisting of both short and prolonged exposure surveillance cue groups; that comparison is presented in this row.

[d] Sparks and Barclay (2013) included three conditions: a short exposure surveillance cue condition, a prolonged exposure surveillance cue condition, and a control condition. For the short exposure version of the meta-analysis, I compared the control condition to the short exposure surveillance cue condition; that comparison is presented in this row.