PSYCHOLOGY

Prevalence-induced concept change in human judgment

David E. Levari[1], Daniel T. Gilbert[1]*, Timothy D. Wilson[2], Beau Sievers[3], David M. Amodio[4], Thalia Wheatley[3]

Why do some social problems seem so intractable? In a series of experiments, we show that people often respond to decreases in the prevalence of a stimulus by expanding their concept of it. When blue dots became rare, participants began to see purple dots as blue; when threatening faces became rare, participants began to see neutral faces as threatening; and when unethical requests became rare, participants began to see innocuous requests as unethical. This "prevalence-induced concept change" occurred even when participants were forewarned about it and even when they were instructed and paid to resist it. Social problems may seem intractable in part because reductions in their prevalence lead people to see more of them.

The deformation of a solid under load is known as "creep." But in the past few years, that term has crept beyond materials science and has come to describe almost any kind of unintended expansion of a boundary. Software developers worry about feature creep (the unintended expansion of a product's function over time), project managers worry about scope creep (the unintended expansion of a team's mandate over time), and military commanders worry about mission creep (the unintended expansion of a campaign's objectives over time). As it turns out, abstract concepts can creep, too. For example, in 1960, Webster's dictionary defined "aggression" as "an unprovoked attack or invasion," but today that concept can include behaviors such as making insufficient eye contact or asking people where they are from (*1*). Many other concepts, such as abuse, bullying, mental disorder, trauma, addiction, and prejudice, have expanded of late as well (*2*). Some take these expansions as signs of political correctness and others as signs of social awakening. We take no position on whether these expansions are good or bad. Rather, we seek to understand what makes them happen. Why do concepts creep?

Psychologists have long known that stimuli are judged in the context of the other relevant stimuli that surround them in space or precede them in time (*3*–*8*), and the perceived aggressiveness of a particular behavior will naturally depend on the aggressiveness of the other behaviors the observer is seeing or has seen. When instances of a concept become less prevalent—for example, when unprovoked attacks and invasions decline—the context in which new instances are judged changes as well. If most behaviors are less aggressive than they once were, then some behaviors will seem more aggressive than they once did, which may lead observers to mistakenly conclude that the prevalence of aggression has not declined. When instances of a concept become less prevalent, the concept may expand to include instances that it previously excluded, thereby masking the magnitude of its own decline.

This phenomenon—which we call "prevalence-induced concept change"—can be a problem. When yellow bananas become less prevalent, a shopper's concept of "ripe" should expand to include speckled ones, but when violent crimes become less prevalent, a police officer's concept of "assault" should not expand to include jaywalking. What counts as a ripe fruit should depend on the other fruits one can see, but what counts as a felony, a field goal, or a tumor should not, and when these things are absent, police officers, referees, and radiologists should not expand their concepts and find them anyway. Modern life often requires people to use concepts that are meant to be held constant and should not be allowed to expand (*9*–*16*). Alas, research suggests that the brain computes the value of most stimuli by comparing them to other relevant stimuli (*17*–*19*); thus, holding concepts constant may be an evolutionarily recent requirement that the brain's standard computational mechanisms are ill equipped to meet (*20*, *21*).

Are people susceptible to prevalence-induced concept change? To answer this question, we showed participants in seven studies a series of stimuli and asked them to determine whether each stimulus was or was not an instance of a concept. The concepts ranged from simple ("Is this dot blue?") to complex ("Is this research proposal ethical?"). After participants did this for a while, we changed the prevalence of the concept's instances and then measured whether the concept had expanded—that is, whether it had come to include instances that it had previously excluded.

In Study 1, we showed participants a series of 1000 dots that varied on a continuum from very purple to very blue (see fig. S1) and asked them to decide whether each dot was or was not blue. After 200 trials, we decreased the prevalence of blue dots for participants in the decreasing prevalence condition but not for participants in the stable prevalence condition. Figure 1 shows the percentage of dots at each point along the continuum that participants identified as blue on the initial 200 trials and on the final 200 trials. The two curves in Fig. 1A are nearly perfectly superimposed, indicating that participants in the stable prevalence condition were just as likely to identify a dot as blue when it appeared on an initial trial as when it appeared on a final trial. But the two curves in Fig. 1B are offset, indicating that participants in the decreasing prevalence condition were more likely to identify dots as blue when those dots appeared on a final trial than when those dots appeared on an initial trial. In other words, when the prevalence of blue dots decreased, participants' concept of blue expanded to include dots that it had previously excluded. Complete methods and results for this and all subsequent studies may be found in the supplementary materials.

In Studies 2 through 5, we examined the robustness of this effect. In Study 2, we replicated the procedure for Study 1, except that instead of telling participants in the decreasing prevalence condition that the prevalence of blue dots "might

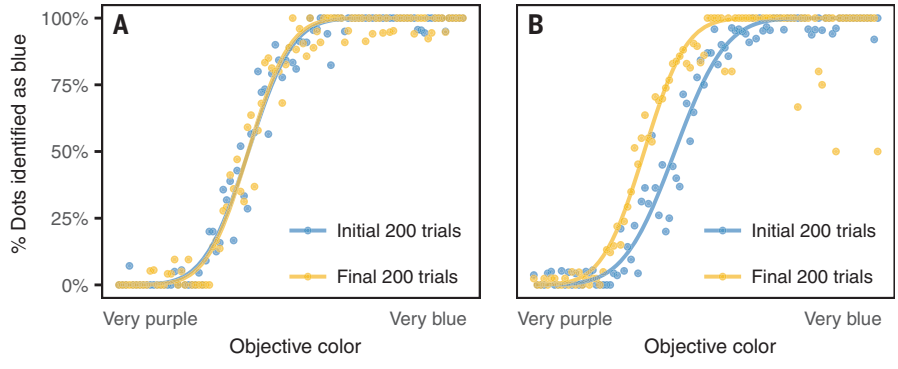


Fig. 1. Results for Study 1. (**A**) shows the stable prevalence condition, and (**B**) shows the decreasing prevalence condition. The *x* axes show the dot's objective color, and the *y* axes show the percentage of trials on which participants identified that dot as blue.

[1]Harvard University, Cambridge, MA, USA. [2]University of Virginia, Charlottesville, VA, USA. [3]Dartmouth University, Hanover, NH, USA. [4]New York University, New York, NY, USA.
*Corresponding author. Email: gilbert@wjh.harvard.edu

change" over trials, we told them that the prevalence of blue dots would "definitely decrease" over trials. The effect seen in Study 1 was replicated (see fig. S2). In Study 3, we replicated the procedure for Study 1, except that this time a third of the participants in the decreasing prevalence condition were explicitly instructed to "be consistent" and to not allow their concept of blue to change over the course of the study, and another third were given the same instruction and were also offered a monetary incentive for following it. In all conditions, the effect seen in Study 1 was replicated (see fig. S3). In Study 4, we replicated the procedure for Study 1, except that this time we decreased the prevalence of blue dots gradually for some participants (as we did in the previous studies) and abruptly for others. In all conditions, the effect seen in Study 1 was replicated (see fig. S4). Finally, in Study 5, we replicated the procedure for Study 1, except that this time instead of decreasing the prevalence of blue dots, we increased the prevalence of blue dots. As expected, this change reversed the effect seen in Study 1: When the prevalence of blue dots was increased, participants were less likely to identify a dot as blue when it appeared on a final trial than when it appeared on an initial trial (see fig. S5). In short, the prevalence-induced concept change seen in Study 1 proved remarkably robust and was not eliminated by forewarning (Study 2), by instructions and incentives (Study 3), by sudden decreases in prevalence (Study 4), or by a reversal in the direction of the change in prevalence (Study 5).

Does this finding generalize from simple concepts to complex ones? To find out, in Study 6, we showed participants a series of 800 computer-generated human faces that (according to raters) varied on a continuum from very threatening to not very threatening (see fig. S6). We asked participants to determine whether the person whose face they saw (the "target") was or was not a threat. After 200 trials, we decreased the prevalence of threatening targets for participants in the decreasing prevalence condition but not for participants in the stable prevalence condition. Figure 2 shows the percentage of targets at each point on the continuum whom participants identified as a threat on the initial 200 trials and on the final 200 trials. Participants in the stable prevalence condition (Fig. 2A) were just as likely to identify a target as a threat when that target appeared on a final trial as when that target appeared on an initial trial, but participants in the decreasing prevalence condition (Fig. 2B) were more likely to identify a target as a threat when the target appeared on a final trial than when the target appeared on an initial trial. In other words, when the prevalence of threatening targets decreased, participants' concept of threat expanded to include targets that it had previously excluded.

The foregoing studies suggest that concepts expand when the prevalence of their instances decreases. Does this effect also occur when people are asked to make decisions about purely conceptual rather than visual stimuli? To find out, in Study 7 we asked participants to play the role of a reviewer on an Institutional Review Board. We showed participants a series of 240 proposals for scientific studies that (according to raters) varied on a continuum from very ethical to very unethical, and we asked participants to decide whether researchers should or should not be allowed to conduct the study. After 96 trials, we decreased the prevalence of unethical proposals for participants in the decreasing prevalence condition but not for participants in the stable prevalence condition. Figure 3 shows the percentage of proposals that participants rejected on the initial 48 trials and on the final 48 trials. Participants in the stable prevalence condition (Fig. 3A) were just as likely to reject ethically ambiguous proposals that appeared on a final trial and on an initial trial, but participants in the decreasing prevalence condition (Fig. 3B) were more likely to reject ethically ambiguous proposals that appeared on a final trial than on an initial trial. In other words, when the prevalence of unethical research proposals decreased, participants' concept of unethical expanded to include proposals that it had previously excluded.

Across seven studies, prevalence-induced concept change occurred when it should not have. When blue dots became rare, purple dots began to look blue; when threatening faces became rare, neutral faces began to appear threatening; and when unethical research proposals became rare, ambiguous research proposals began to seem unethical. This happened even when the change in the prevalence of instances was abrupt, even when participants were explicitly told that the prevalence of instances would change, and even when participants were instructed and paid to ignore these changes.

These results may have sobering implications. Many organizations and institutions are dedicated to identifying and reducing the prevalence of social problems, from unethical research to unwarranted aggressions. But our studies suggest that even well-meaning agents may sometimes fail to recognize the success of their own efforts, simply because they view each new instance in the decreasingly problematic context that they themselves have brought about. Although modern societies have made extraordinary progress in solving a wide range of social problems, from poverty and illiteracy to violence and infant mortality (*22*, *23*), the majority of people believe that the world is getting worse (*24*). The fact that concepts grow larger when their instances grow smaller may be one source of that pessimism.

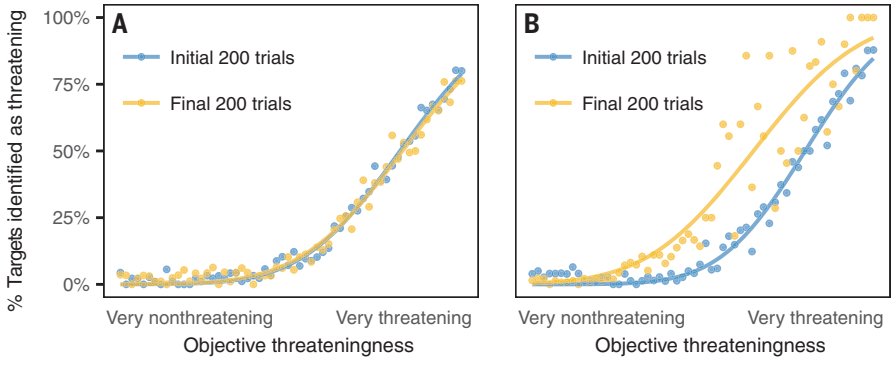


Fig. 2. Results for Study 6. (**A**) shows the stable prevalence condition, and (**B**) shows the decreasing prevalence condition. The *x* axes show the target's objective threateningness (as determined by human raters), and the *y* axes show the percentage of trials on which participants identified that target as a threat.

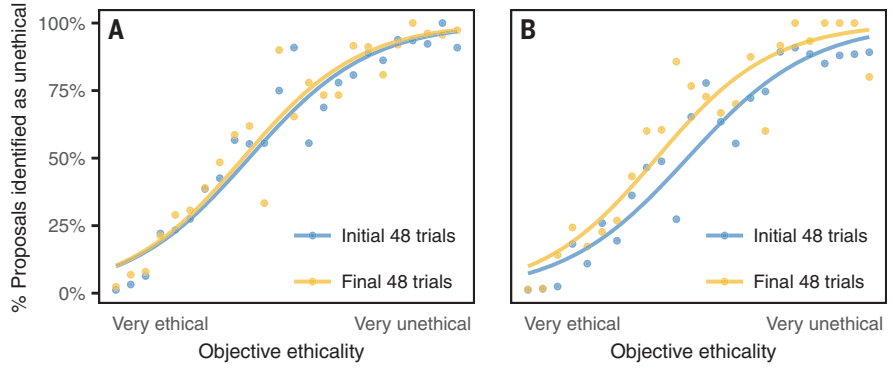


Fig. 3. Results for Study 7. (**A**) shows the stable prevalence condition, and (**B**) shows the decreasing prevalence condition. The *x* axes show the proposal's objective ethicality (as determined by raters), and the *y* axes show the percentage of trials on which participants rejected the proposal.

REFERENCES AND NOTES

1. S. O. Lilienfeld, *Perspect. Psychol. Sci.* **12**, 138–169 (2017).
2. N. Haslam, *Psychol. Inq.* **27**, 1–17 (2016).
3. A. Parducci, *Psychol. Rev.* **72**, 407–418 (1965).
4. H. Helson, *Adaptation-Level Theory* (Harper & Row, New York, 1964).
5. C. W. G. Clifford *et al*., *Vision Res.* **47**, 3125–3131 (2007).
6. C. Summerfield, F. P. de Lange, *Nat. Rev. Neurosci.* **15**, 745–756 (2014).
7. X. X. Wei, A. A. Stocker, *Nat. Neurosci.* **18**, 1509–1517 (2015).
8. N. Stewart, N. Chater, G. D. A. Brown, *Cognit. Psychol.* **53**, 1–26 (2006).
9. A. Pepitone, M. DiNubile, *J. Pers. Soc. Psychol.* **33**, 448–459 (1976).
10. G. Rodríguez, S. Blanco, *Anu. Psicol. Juridica* **26**, 107–113 (2016).
11. D. L. Chen, T. J. Moskowitz, K. Shue, *Q. J. Econ.* **131**, 1181–1242 (2016).
12. U. Simonsohn, F. Gino, *Psychol. Sci.* **24**, 219–224 (2013).
13. S. Bhargava, R. Fisman, *Rev. Econ. Stat.* **96**, 444–457 (2014).
14. U. Simonsohn, *Rev. Econ. Stat.* **88**, 1–9 (2006).
15. U. Simonsohn, G. Loewenstein, *Econ. J. (Lond.)* **116**, 175–199 (2006).
16. L. Damisch, T. Mussweiler, H. Plessner, *J. Exp. Psychol. Appl.* **12**, 166–178 (2006).
17. M. Carandini, D. J. Heeger, *Nat. Rev. Neurosci.* **13**, 51–62 (2011).
18. K. Louie, M. W. Khaw, P. W. Glimcher, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 6139–6144 (2013).
19. A. Parducci, *Sci. Am.* **219**, 84–90 (1968).
20. J. M. Wolfe, T. S. Horowitz, N. M. Kenner, *Nature* **435**, 439–440 (2005).
21. M. C. Hout, S. C. Walenchok, S. D. Goldinger, J. M. Wolfe, *J. Exp. Psychol. Hum. Percept. Perform.* **41**, 977–994 (2015).
22. M. Roser, The short history of global living conditions and why it matters that we know it (2017); available at https://ourworldindata.org/a-history-of-global-living-conditions-in-5-charts.
23. S. Pinker, *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress* (Viking Press, New York, 2018).
24. YouGov Survey (2015); available at https://d25d2506sfb94s.cloudfront.net/cumulus_uploads/document/z2knhgzguv/GB_Website.pdf.
25. D. Levari, Prevalence-induced concept change in human judgment. Zenodo, Version v1.0.0 (2018); 10.5281/zenodo.1219833.

ACKNOWLEDGMENTS

We thank T. Brady and M. Thornton for technical assistance and M. Banker, S. Carroll, P. Chan, R. Chmielinksi, A. Collinsworth, K. da Silva, I. Droney, C. Fitzgerald, S. Ganley, M. Graether, J. Green, L. Harris, U. Heller, T. Hicks, S. Hoffman, E. Kemp, B. Kollek, R. Lisner, Z. Lu, B. Martinez, T. Murphy, D. Park, D. Peng, M. Powell, M. Sanders, C. Shaw, G. Stern, R. Stramp, A. Strauss, L. Symes, H. Tieh, A. Wang, I. Wang, C. Wu, X. Zeng, and Y. Zheng for research assistance. Statistical support was provided by data science specialists S. Worthington and I. Zahn at the Institute for Quantitative Social Science, Harvard University. **Funding:** We acknowledge the support of National Science Foundation grant BCS-1423747 to T.D.W. and D.T.G. **Author contributions:** All authors contributed to the conceptual development of the work, D.E.L. collected and analyzed the data, and D.E.L. and D.T.G. designed the experiments and wrote the manuscript. **Competing interests:** None. **Data and materials availability:** The complete materials and data for all studies are available at Zenodo, https://zenodo.org/record/1219833 (*25*).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/360/6396/1465/suppl/DC1
Materials and Methods
Figs. S1 to S6
References (*26*–*31*)

5 September 2017; accepted 30 April 2018
10.1126/science.aap8731

Prevalence-induced concept change in human judgment

David E. Levari, Daniel T. Gilbert, Timothy D. Wilson, Beau Sievers, David M. Amodio and Thalia Wheatley

Science **360** (6396), 1465-1467.
DOI: 10.1126/science.aap8731

Perceptual and judgment creep

Do we think that a problem persists even when it has become less frequent? Levari *et al.* show experimentally that when the "signal" a person is searching for becomes rare, the person naturally responds by broadening his or her definition of the signal—and therefore continues to find it even when it is not there. From low-level perception of color to higher-level judgments of ethics, there is a robust tendency for perceptual and judgmental standards to "creep" when they ought not to. For example, when blue dots become rare, participants start calling purple dots blue, and when threatening faces become rare, participants start calling neutral faces threatening. This phenomenon has broad implications that may help explain why people whose job is to find and eliminate problems in the world often cannot tell when their work is done.

Science, this issue p. 1465

www.sciencemag.org/content/360/6396/1465/suppl/DC1

Supplementary Materials for

Prevalence-induced concept change in human judgment

David E. Levari, Daniel T. Gilbert*, Timothy D. Wilson, Beau Sievers, David M. Amodio, Thalia Wheatley

*Corresponding author. Email: gilbert@wjh.harvard.edu

Published 29 June 2018, *Science* **360**, 1465 (2018)
DOI: 10.1126/science.aap8731

This PDF file includes:

Methods

Methods for Study 1

Overview. In Study 1, we showed participants a series of dots on a computer screen and asked them to determine whether each dot was blue or not blue. After many trials, we decreased the prevalence of the blue dots for some participants. This and all subsequent studies were approved by the Harvard University Committee on the Use of Human Subjects.

Sample. Participants were 22 students at Harvard University (6 males, 16 females, M_{age} = 22.5 years, SD = 1.9 years) who received either money or course credit in exchange for their participation. One female participant experienced a minor medical problem during the study and her data were excluded, leaving 21 participants in the data set. In this and all subsequent studies: (a) We set a minimum sample size based on previous research that had used similar methods and stimuli, (b) once we reached the minimum sample size, we continued to recruit participants through the end of the academic term, (c) we did not analyze our data until all participants had been recruited, (d) all manipulations, measures, and data exclusions are reported, and (e) data exclusions had no impact on the significance of the results.

Procedures. Upon arrival at the laboratory, participants were escorted to a room equipped with a computer display and keyboard, and they remained there for the duration of the study. Participants were told that a series of colored dots would appear on the screen, one at a time, and that their task was to decide whether each dot was blue or not blue, and to indicate their decision by pressing one of two keys on the keyboard that were respectively labeled “blue” and “not blue.”

On each trial, a colored dot appeared on a solid gray background. The color of the dot varied across trials from very purple (60% blue, RGB 100-0-155) to very blue (99.6% blue, RGB 1-0-254). Each dot appeared on the screen for 500 milliseconds and was then replaced by a question mark, which remained on the screen until participants pressed one of the response keys. Participants were told that there would be 1000 trials divided into 20 blocks, and that the prevalence of blue dots might vary across blocks. Specifically, they were told that some blocks "may have a lot of blue dots, and others may have only a few." Participants completed 10 practice trials to ensure they understood the procedure, and then completed 1000 test trials. To help participants remain attentive, we allowed them to take a break every 50 trials.

We created two conditions by dividing the color spectrum into two halves that we will refer to as the "purple spectrum" (RGB 100-0-155 through RGB 51-0-204) and the "blue spectrum" (RGB 50-0-205 through RGB 1-0-254), as shown in Figure S1. Half the participants were randomly assigned to the *stable* condition. In this condition, we determined the color of the dot shown on each trial by randomly sampling the two spectra with equal probability. We will refer to the probability that a dot was sampled from the blue spectrum as the *signal prevalence*. In the stable condition, the signal prevalence on trials 1-1000 was 50%. The remaining participants were assigned to the *decreasing* condition. In this condition, we sampled the two spectra with unequal probability on some trials. Specifically, in the decreasing condition the signal prevalence was 50% on trials 1-200; 40% on trials 201-250; 28% on trials 251-300; 16% on trials 301-350; and 6% on trials 351-1000.

Analyses and Results. The tasks that participants performed in this and all subsequent studies performed may be thought of as signal detection tasks. However, traditional signal detection tasks present participants with stimuli that can be objectively classified as either signal or noise, and the data are typically analyzed by using the number of correct and incorrect responses to calculate *d′* (sensitivity) and *c* (response threshold) for each participant. Because there are no "objectively correct" answers to questions such as "Is this dot blue?" or "Is this face threatening?" or "Is this proposal acceptable?" it is not possible to calculate these traditional parameters for our data. Our alternative analytic approach is described below. In addition, in this and subsequent studies, we used Generalized Linear Mixed Models to account for the nested and unbalanced structure of the data. It is worth noting that analyses using alternative strategies—such as Bayesian linear mixed-effect models implemented in R using the *blme* package (*25*) as well as repeated-measures Analysis of Variance with responses binned across trials—yielded the same basic pattern of results described below.

Did the decrease in the prevalence of blue dots cause participants' concepts of *blue* to expand? To find out, we fit a binomial generalized linear mixed model to our data in R (*26*) using the *lme4* package (*27*). The dependent variable was the participant's *identification* of a dot as blue or not blue. The independent between-participants variable was the participant's *condition* (stable or decreasing). The independent within-participants variables were (a) the dot's RGB value or what we will call its *objective color* (which ranged from 0% blue to 99.6% blue) and (b) the *trial number* (which ranged from 1 to 1000). We included condition, trial number, and objective color (and all interactions between them) as fixed effects in our model. We included as random effects

(a) intercepts for participants (who may have entered our study with different thresholds) and (b) slopes for trial number. The inclusion of random intercepts significantly improved model fit relative to the baseline model, $\chi^2(2) = 494.59$, $p < 0.001$, as did the inclusion of random slopes, $\chi^2(2) = 127.66$, $p < 0.001$. Additionally, the inclusion of the three-way interaction between condition, trial number, and objective threateningness significantly improved model fit, $\chi^2(1) = 48.34$, $p < 0.001$. The generalized linear mixed model revealed that a Condition X Objective Color X Trial Number interaction predicted participants' identifications, $b = 12.50$, $SE = 1.75$, $z = 7.14$, $p < 0.001$, *95% CI* [8.85, 16.09], $R^2_{GLMM(c)} = 0.88$. (All reported 95% confidence intervals are the result of a bootstrapping procedure using 1000 bootstrap samples).

Methods for Study 2

Overview. In Study 2, we replicated the procedure for Study 1, except that instead of telling participants in the decreasing condition that the prevalence of blue dots *might change* over trials, we told them that the prevalence of blue dots *would decrease* over trials.

Sample. Participants were 43 students at Harvard University (10 males, 31 females, $M_{age} = 20.4$ years, SD = 2.1 years) who received either money or course credit in exchange for their participation. Two female participants who were given incorrect study materials were excluded, as was one male participant who disregarded experimental instructions and one male participant who reported being colorblind. This left 39 participants in the data set.

Procedures. The method for Study 2 was identical to the method for Study 1 except that before the study began, participants were explicitly told what would happen to the prevalence of blue dots during the study. Participants in the decreasing condition were told: "As the study goes on, blue dots are going to become less common. In other words, you will see fewer of them over time." Participants in the stable condition were told: "As the study goes on, blue dots are not going to become more or less common. In other words, you will see the same amount of them over time."

Analyses and Results. Did the decrease in the prevalence of blue dots cause participants' concepts of *blue* to expand even when they were explicitly told that the prevalence of blue dots would decrease? To find out, we fit a binomial generalized linear mixed model to our data in R (*26*) using the lme4 package (*27*). The dependent variable was the participant's *identification* of a dot as blue or not blue. The independent between-participants variable was the participant's *condition* (stable or decreasing). The independent within-participants variables were (a) the dot's RGB value or what we will call its *objective color* (which ranged from 0% blue to 99.6% blue) and (b) the *trial number* (which ranged from 1 to 1000). We included condition, trial number, and objective color (and all interactions between them) as fixed effects in our model. We included as random effects (a) intercepts for participants (who may have entered our study with different thresholds) and (b) slopes for trial number. The inclusion of random intercepts significantly improved model fit relative to the baseline model, $\chi^2(2) = 692.36$, $p < 0.001$, as did the inclusion of random slopes, $\chi^2(2) = 229.86$, $p < 0.001$. Additionally, the inclusion of the three-way interaction between condition, trial number, and objective color significantly improved model fit, $\chi^2(1) = 117.91$, $p < 0.001$. The generalized linear

mixed model revealed that a Condition X Objective Color X Trial Number interaction predicted participants' identifications, $b = 21.74$, $SE = 1.55$, $z = 14.00$, *95% CI* [17.83, 25.77], $R^2_{GLMM(c)} = 0.93$. All reported 95% confidence intervals are the result of a bootstrapping procedure using 1000 bootstrap samples.

Figure S2 shows the percentage of dots of each color that participants in each condition identified as blue on the initial trials (1-200) and the final trials (800-1000). The positive slope of all curves indicates that in both conditions, participants' identifications were highly correlated with the dot's position on the color spectrum. But the two panels differ in an important way. The two curves panel A are nearly perfectly superimposed, indicating that participants in the stable condition were just as likely to identify a dot as blue when it appeared on a final trial as when it appeared on an initial trial. But the two curves in panel B are offset in the middle, indicating that participants in the decreasing condition were more likely to identify dots from the middle of the color spectrum as blue when those dots appeared on a final trial than when they appeared on an initial trial. In short, when blue dots became less prevalent, participants identified as blue some dots that they had earlier identified as not blue, and they did this even when they were explicitly warned about the decrease in prevalence.

Methods for Study 3

Overview. In Study 3, we replicated the procedure for Study 1, except that this time a third of the participants in the decreasing condition were *explicitly instructed* not to change their identifications of dots over the course of the study ("Do your best to respond the same way if you see it again later in the study"), and another third were given

the same explicit instruction and also offered a *monetary reward* for following it ("We will be awarding a bonus of $10 to the five most consistent participants in this study").

Sample. Participants were 92 students at Harvard University (34 males, 57 females, M_{age} = 18.4 years, SD = 2.1 years) who received course credit in exchange for their participation. One female participant who was interrupted during the experimental session was excluded, leaving 91 participants in the data set.

Procedures. The method for Study 3 was virtually identical to the method for Study 1 except for two things. First, we added two new conditions. Whereas participants in the *stable condition* and the *decreasing* condition were given the same instructions as they were given in Study 1, participants in the new conditions were given different instructions. Specifically, participants in the new *decreasing+instruction* condition were told that once they had identified a dot as blue or not blue "you should do your best to respond the same way if you see it again later in the study." Participants in the new *decreasing+instruction+incentive* condition were told the same thing, and in addition, they were also told that "as an incentive, we will be awarding a bonus of $10 to the five most consistent participants in this study." The second change to the method of Study 1 is that we reduced the number of trials from 1000 to 800. As such, the signal prevalence in the stable condition was 50% on trials 1-800, and the signal prevalence in the decreasing condition, the decreasing+instruction condition, and the decreasing+instruction+incentive condition was 50% on trials 1-200; 40% on trials 201-250; 28% on trials 251-300; 16% on trials 301-350; and 6% on trials 351-800.

Analyses and Results. Did the decrease in the prevalence of blue dots cause participants' concepts of *blue* to expand even when they were instructed, or instructed

and incentivized, not to let this happen? To find out, we fit a binomial generalized linear mixed model to our data in R (*26*) using the lme4 package (*27*). The dependent variable was the participant's *identification* of a dot as blue or not blue. The independent between-participants variable was the participant's *condition* (stable or decreasing). The independent within-participants variables were (a) the dot's RGB value or what we will call its *objective color* (which ranged from 0% blue to 100% blue) and (b) the *trial number* (which ranged from 1 to 800). We included condition, trial number, and objective color (and all interactions between them) as fixed effects in our model. We included as random effects (a) intercepts for participants (who may have entered our study with different thresholds) and (b) slopes for trial number. The inclusion of random intercepts significantly improved model fit relative to the baseline model, $\chi^2(2) = 1084.00$, $p < 0.001$, as did the inclusion of random slopes, $\chi^2(2) = 500.29$, $p < 0.001$. Additionally, the inclusion of the three-way interaction between condition, trial number, and objective color significantly improved model fit, $\chi^2(3) = 234.53$, $p < 0.001$. The generalized linear mixed model revealed that a Condition X Objective Color X Trial Number interaction predicted participants' identifications. Specifically, the stable prevalence condition differed significantly from the decreasing prevalence condition, $b = 21.98$, $SE = 0.67$, $z = 32.8$, $p < 0.001$, *95% CI* [18.44, 25.49], $R^2_{GLMM(c)} = 0.92$, the *decreasing* + instruction condition, $b = 27.84$, $SE = 1.48$, $z = 18.8$, $p < 0.001$, *95% CI* [23.72, 31.88], and the *decreasing* + *instruction* + *incentive* condition, $b = 15.34$, $SE = 1.29$, $z = 11.9$, $p < 0.001$, *95% CI* [12.13, 18.38]. The *decreasing* + *instruction* condition also differed significantly from the decreasing prevalence condition, $b = -5.86$, $SE = 0.71$, $z = -8.3$, $p < 0.001$, *95% CI* [-9.78, -1.91], as well as from the *decreasing* + *instruction* + *incentive* condition, $b =$ -

12.50, *SE* = 1.09, *z* = -11.5, p < 0.001, *95% CI* [-16.14, -8.85]. Finally, the *decreasing + instruction + incentive* condition differed significantly from the decreasing prevalence condition, *b* = 6.64, *SE* = 0.67, *z* = 9.9, p < 0.001, *95% CI* [3.56, 9.94]. (All reported 95% confidence intervals are the result of a bootstrapping procedure using 1000 bootstrap samples, and all reported p-values are adjusted for multiple comparisons using the Holm correction).

Figure S3 shows the percentage of dots of each color that participants in each condition identified as blue on the initial trials (1-200) and the final trials (600-800). The positive slope of all curves indicates that in all conditions, participants' identifications were highly correlated with the dot's position on the color spectrum. But the panels differ in an important way. The two curves in panel A are nearly perfectly superimposed, indicating that participants in the stable condition were just as likely to identify a dot as blue when it appeared on a final trial as when it appeared on an initial trial. But in each of the other panels, the two curves are offset in the middle, indicating that participants in the three decreasing conditions were more likely to identify dots from the middle of the color spectrum as blue when those dots appeared on a final trial than when they appeared on an initial trial. In short, when blue dots became less prevalent, participants identified as blue some dots that they had earlier identified as not blue, and they did this even when they had been instructed and incentivized not to let that happen.

Methods for Study 4

Overview. In Study 4, we replicated the procedure for Study 1, except that in Study 4 we decreased the prevalence of blue dots *gradually* for some participants (as we did in the previous studies) and *abruptly* for others.

Sample. Participants were 37 students at Harvard University (12 males, 25 females, M_{age} = 19.4 years, SD = 1.5 years) who received either money or course credit in exchange for their participation.

Procedures. The method for Study 4 was virtually identical to the method for Study 1 except for two things. First, we reduced the number of trials from 1000 to 800. Second, we added a new condition. For participants in the *stable condition,* the signal prevalence on trials 1-800 was 50%. This condition was the same as the stable condition in Study 3. For participants in the *gradually decreasing* condition, the signal prevalence was 50% on trials 1-200; 40% on trials 201-250; 28% on trials 251-300; 16% on trials 301-350; and 6% on trials 351-800. This condition was the same as the decreasing condition in Study 3. For participants in the new *abruptly decreasing* condition, the signal prevalence was 50% on trials 1-200, and 6% on trials 201-800.

Analyses and Results. Did the decrease in the prevalence of blue dots cause participants' concept of *blue* to expand even when the decrease occurred abruptly? To find out, we fit a binomial generalized linear mixed model to our data in R (*26*) using the lme4 package (*27*). The dependent variable was the participant's *identification* of a dot as blue or not blue. The independent between-participants variable was the participant's *condition* (stable, gradually decreasing, or abruptly decreasing). The independent within-participants variables were (a) the dot's RGB value or what we will call its *objective color* (which ranged from 0% blue to 100% blue) and (b) the *trial number* (which ranged

from 1 to 800). We included condition, trial number, and objective color (and all interactions between them) as fixed effects in our model. We included as random effects (a) intercepts for participants (who may have entered our study with different thresholds) and (b) slopes for trial number. The inclusion of random intercepts significantly improved model fit relative to the baseline model, $\chi^2(2) = 234.49$, $p < 0.001$, as did the inclusion of random slopes, $\chi^2(2) = 48.32$, $p < 0.001$. Additionally, the inclusion of the three-way interaction between condition, trial number, and objective color significantly improved model fit, $\chi^2(1) = 72.52$, $p < 0.001$. The generalized linear mixed model revealed that a Condition X Objective Color X Trial Number interaction predicted participants' identifications. Specifically, the stable prevalence condition differed significantly from both the gradually decreasing prevalence condition, $b = 15.92$, $SE = 1.19$, $z = 13.3$, $p < 0.001$, *95% CI* [*11.63*, 20.36]], $R^2_{GLMM(c)} = 0.89$, as well as from the abruptly decreasing prevalence condition, $b = 15.26$, $SE = 0.56$, $z = 27.3$, $p < 0.001$, *95% CI* [11.04, 19.45]. However, the gradually and abruptly decreasing prevalence conditions did not differ significantly from one another, $b = 0.66$, $SE = 1.19$, $z = 0.6$, $p = 0.58$, *95% CI* [-3.96, 5.34]. (All reported 95% confidence intervals are the result of a bootstrapping procedure using 1000 bootstrap samples, and all reported p-values are adjusted for multiple comparisons using the Holm correction).

Figure S4 shows the percentage of dots of each color that participants in each condition identified as blue on the initial trials (1-200) and the final trials (600-800). The positive slope of all curves indicates that in both conditions, participants' identifications were highly correlated with the dot's position on the color spectrum. But panels B and C differ from panel A in an important way. The two curves in panel A are nearly perfectly

superimposed, indicating that participants in the stable condition were just as likely to identify a dot as blue when it appeared on a final trial as when it appeared on an initial trial. But the two curves in panels B and C are offset in the middle, indicating that participants in the two decreasing conditions were more likely to identify dots from the middle of the color spectrum as blue when those dots appeared on a final trial than when they appeared on an initial trial. In short, when blue dots became less prevalent, participants identified as blue some dots that they had earlier identified as not blue, and they did this even when the decrease in prevalence happened abruptly.

Methods for Study 5

Overview. In Study 5, we replicated the procedure for Study 1, except that in Study 5, instead of *decreasing* the prevalence of blue dots in the experimental condition, we *increased* their prevalence.

Sample. Participants were 23 students at Harvard University (11 males, 12 females, M_{age} = 22.1 years, SD = 2.5 years) who received course credit in exchange for their participation. One female participant did not follow the experimenter's instructions during the study and her data were excluded, leaving 22 participants in the data set.

Procedures. The method for Study 5 was virtually identical to the method for Study 1 except that we replaced the decreasing condition with an *increasing condition*. The signal prevalence in the increasing condition was 6% on trials 1-200; 16% on trials 201-250; 28% on trials 251-300; 40% on trials 301-350; and 50% on trials 351-1000.

Analyses and Results. Did the increase in the prevalence of blue dots cause participants' concepts of *blue* to contract (rather than to expand)? To find out, we fit a

binomial generalized linear mixed model to our data in R (*26*) using the lme4 package (*27*). The dependent variable was the participant's *identification* of a dot as blue or not blue. The independent between-participants variable was the participant's *condition* (stable or increasing). The independent within-participants variables were (a) the dot's RGB value or what we will call its *objective color* (which ranged from 0% blue to 100% blue) and (b) the *trial number* (which ranged from 1 to 1000). We included condition, trial number, and objective color (and all interactions between them) as fixed effects in our model. We included as random effects (a) intercepts for participants (who may have entered our study with different thresholds) and (b) slopes for trial number. The inclusion of random slopes significantly improved model fit relative to the baseline model, $\chi^2(2) = 49.57$, $p < 0.001$, as did the inclusion of random intercepts, $\chi^2(2) = 386.15$, $p < 0.001$. Additionally, the inclusion of the three-way interaction between condition, trial number, and objective color significantly improved model fit, $\chi^2(1) = 15.12$, $p < 0.001$. The generalized linear mixed model revealed that a Condition X Objective Color X Trial Number interaction predicted participants' identifications, $b = -8.13$, $SE = 1.40$, $z = -5.83$, *95% CI* [-12.31, -4.05], $R^2_{GLMM(c)} = 0.89$. (All reported 95% confidence intervals are the result of a bootstrapping procedure using 1000 bootstrap samples).

Figure S5 shows the percentage of dots of each color that participants in each condition identified as blue on the initial trials (1-200) and the final trials (800-1000). The positive slope of all curves indicates that in both conditions, participants' identifications were highly correlated with the dot's position on the color spectrum. But the two panels differ in an important way. The two curves in panel A are nearly perfectly superimposed, indicating that participants in the stable condition were just as likely to

identify a dot as blue when it appeared on a final trial as when it appeared on an initial trial. But the two curves in panel B are offset in the middle, indicating that participants in the increasing condition were less likely to identify dots from the middle of the color spectrum as blue when those dots appeared on a final trial than when they appeared on an initial trial. In short, when blue dots became more prevalent, participants identified as not blue some dots that they had earlier identified as blue.

Methods for Study 6

Overview. In Study 6, we showed participants a series of computer-generated human faces on a computer screen and asked them to determine whether the person they saw (hereinafter referred to as *the target*) was a threat or was not a threat. Over the course of many trials, we decreased the prevalence of threatening targets for some participants. We predicted that these participants would respond to the decreasing prevalence of threatening targets by identifying some targets as threats whom they had previously identified as non-threats.

Sample. Participants were 49 students at Harvard University (28 male, 20 female, and 1 gender unspecified, M_{age} = 20.8 years, SD = 2.0 years) who received either money or course credit in exchange for their participation. One male participant reported having a form of prosopagnosia (face blindness), and his data were excluded, leaving 48 participants in the data set.

Procedures. Upon arrival at the laboratory, participants were escorted to a room equipped with a computer display and keyboard, and they remained there for the duration

of the study. Participants were told that a series of target persons would appear on the screen, one at a time, and that their task was to decide whether each target person was or was not a threat, and to indicate their decision by pressing one of two keys on the keyboard that were respectively labeled "threat" and "no threat." On each trial, a computer-generated image of a target person's face appeared on a solid gray background. In previous research, Todorov and colleagues (*28*, *29*) used a computational model to randomly generate a set of faces, and they then had human participants rate the threateningness of each face. We took the faces from this set that had been rated as most and least threatening, and we then used Fantamorph (*30*) to incrementally morph these faces into one another to produce a continuum of 60 computer-generated faces with expressions that ranged from not very threatening to very threatening. Sample faces are shown in Figure S6.

Although the threateningness of a face is inherently subjective, for the sake of consistency we refer to the mean rating of each target as its *objective threateningness*. Each target appeared on the screen for 500 milliseconds and was then replaced by a question mark, which remained on the screen until participants pressed one of the response keys. Participants were told that there would be 800 trials divided into 16 blocks, and that the prevalence of threatening targets might vary over blocks. Participants completed 10 practice trials to ensure that they understood the procedure, and then completed 800 test trials. To help participants remain attentive, we allowed them to take a break every 50 trials.

We created two conditions by dividing the target continuum into two halves that we will refer to as the "no threat continuum" and the "threat continuum." Half the

participants were randomly assigned to the *stable* condition. In this condition, we determined the threateningness of the target shown on each trial by randomly sampling the two continua with equal probability. We will refer to the probability that a target was sampled from the threat continuum as the *signal prevalence*. In the stable condition, the signal prevalence on trials 1-800 was 50%. The remaining participants were assigned to the *decreasing* condition. In this condition, we sampled the two continua with unequal probability on some trials. Specifically, in the decreasing condition, the signal prevalence was 50% on trials 1-200; 40% on trials 201-250; 28% on trials 251-300; 16% on trials 301-350; and 6% on trials 351- 800.

Analyses and Results. Did the decrease in the prevalence of threatening targets cause participants' concepts of *threat* to expand*?* To find out, we fit a binomial generalized linear mixed model to our data in R (*26*) using the lme4 package (*27*). The dependent variable was the participant's *identification* of a target as threatening or not threatening. The independent between-participants variable was the participant's *condition* (stable or decreasing). The independent within-participants variables were (a) the target's position on the continuum or what we will call its *objective threateningness* (which ranged from 0% threatening to 100% threatening) and (b) the *trial number* (which ranged from 1 to 800). We included condition, trial number, and objective threateningness (and all interactions between them) as fixed effects in our model. We included as random effects (a) intercepts for participants (who may have entered our study with different thresholds) and (b) slopes for trial number. The inclusion of random intercepts significantly improved model fit relative to the baseline model, $\chi^2(2) = 649.03$, $p < 0.001$, as did the inclusion of random slopes, $\chi^2(2) = 974.24$, $p < 0.001$. Additionally,

the inclusion of the three-way interaction between condition, trial number, and objective threateningness significantly improved model fit, $\chi^2(1) = 32.24$, $p < 0.001$. The generalized linear mixed model revealed that a Condition X Objective Threateningness X Trial Number interaction predicted participants' identifications, $b = 4.84$, $SE = 0.86$, $z = 5.61$, *95% CI* [3.12, 6.53], $R^2_{GLMM(c)} = 0.75$.

Methods for Study 7

Overview. In Study 7, we asked participants to play the role of a reviewer on an IRB. We showed participants a series of proposals for scientific studies and asked them to decide whether researchers should be prohibited from conducting the study or should be allowed to conduct the study. The proposals varied in their ethicality. Over the course of many trials, we decreased the prevalence of unethical proposals for some participants. We predicted that these participants would respond to the decrease in the prevalence of unethical proposals by rejecting some proposals that were ethically identical to those they had previously accepted.

Whereas colors and computer-generated faces vary on physical continua that can be measured on a ratio scale, ethicality can at best be measured on an ordinal scale. As such, the materials and procedures for Study 7 differed somewhat from the materials and procedures used in our previous studies.

Materials. We wrote 381 short proposals for scientific experiments involving human participants. The proposals contained between 5 and 37 words (M = 25.34 words). We used our own judgment to preliminarily classify each proposal as either ethical, ambiguous, or unethical. We then recruited 361 U. S. residents (198 male, 161 female, 2

gender unspecified) via Amazon Mechanical Turk and asked them to read and rate a subset of these proposals. We will refer to these participants as *the raters*. Raters were told that (a) the proposals described experiments that were designed to be conducted with adults who had volunteered to take part in exchange for money; (b) all the studies described in the proposals were research on human behavior; (c) when scientists lie to participants either before or during a study, they always tell those participants the truth when the study is over; and (d) participants are always free to withdraw from a study at any time.

Each rater was paid $1 to read and rate 76 proposals. We divided the 381 proposals into a set of 15 proposals that were seen by all raters (the constant set) and a set of 366 proposals that were seen by a subset of raters (the variable set). Specifically, the 366 proposals were divided into 6 sets of 61 proposals (the variable sets), and each rater saw one of these 6 variable sets as well as the constant set of 15 proposals. Twenty-one of the proposals in each of the variable sets had been preliminarily classified as ethical, 23 had been preliminarily classified as ambiguous, and 17 had been preliminarily classified as unethical. The 61 proposals in each of the variable sets were presented in random order, and after the 20th and 40th, and 61st proposals we included a "catch question" to ensure that raters were reading carefully (viz., "If you're actually reading this question, please select the number 3 as your response. Thank you for reading all the questions carefully"). Each rater first saw one of the 6 variable sets of 61 proposals, and then saw the 15 proposals in the constant set. After seeing each proposal, raters were asked the question "Should this experiment be allowed to be conducted?" which they answered using a 7-point Likert scale whose endpoints were anchored with the phrases "Definitely

not" (1) and "Definitely" (7). Raters spent between 3.18 and 53.72 minutes (M = 16.09 min) making their ratings. After they did so, raters completed several other measures including a Turing test (e.g., "If you're reading this, type the word *banana*"), and supplied demographic information.

We excluded the ratings of two male and three female raters who failed the Turing test, and then computed the mean rating of each proposal. Despite the fact that participants' ratings were inherently subjective, for the sake of consistency we will refer to the mean of each proposal's ratings as its *objective ethicality.* Each rater saw 76 proposals. Fifteen of these proposals (the constant set) were seen by all raters, which allowed us to estimate how much the complete pool of raters agreed with regard to judgments of ethicality. Inter-rater reliability was quite high (Cronbach's alpha = .85), indicating that raters were in very close agreement about the objective ethicality of the proposals. We used each proposal's objective ethicality to classify it as a member of one of three categories. To ensure that we had a sufficient number of proposals in each of these categories, we classified proposals whose objective ethicality was greater than 6 and less than or equal to 7 as *ethical*; proposals whose objective ethicality was greater than 4 and less than or equal to 6 as *ambiguous*; and proposals whose objective ethicality was less than or equal to 4 and greater than or equal to 1 as *unethical*. We then selected the proposals in each of the three categories whose objective ethicality ratings had the lowest standard deviations. Specifically, we selected 113 *ethical proposals* (e.g., "Participants will make a list of the cities they would most like to visit around the world, and write about what they would do in each one"), 80 *ambiguous proposals* ("Participants will be given a plant and told that it is a natural remedy for itching. In

reality, it will cause itching. Their reaction will be recorded"), and 80 *unethical proposals* (e.g., "Participants will be asked to lick a frozen piece of human fecal matter. Afterwards, they will be given mouthwash. The amount of mouthwash used will be measured"). These 273 proposals were used as materials in Study 7.

Sample. Participants in Study 7 were 84 students at Harvard University (16 male, 66 female, 2 gender unspecified, M_{age} = 20.73 years, SD = 2.8 years) who received either money or course credit for their participation.

Procedures. Upon arrival at the laboratory, participants were escorted to a room equipped with a computer display and keyboard, and they remained there for the duration of the study. Participants were told that a series of proposals for scientific studies would appear on the screen, one at a time, and that their task was to decide whether researchers should or should not be allowed to conduct each study. They were asked to indicate their decision about each proposal by pressing one of two keys on the keyboard that were respectively labeled "approve" and "reject." On each trial, participants read one of 273 proposals. Each proposal appeared on the screen and remained there until participants pressed one of the response keys. Participants were told that there would be 240 trials divided into 10 blocks, and that the ethicality of the proposals might vary over blocks. Participants completed one practice trial to ensure that they understood the procedure, and then completed 240 test trials. To help participants remain attentive, we allowed them to take a break every 24 trials.

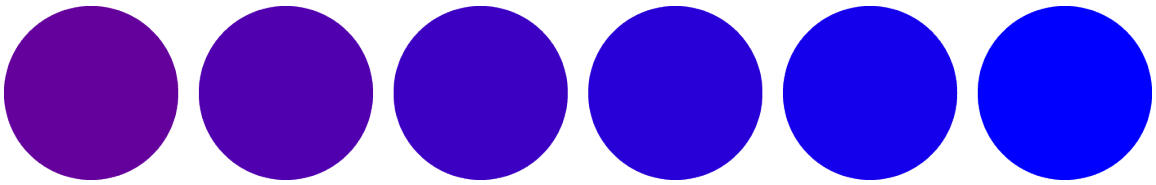
We created two conditions. Half the participants were randomly assigned to the *stable* condition. In this condition, we determined the ethicality of the proposal on each trial by randomly sampling the three ethicality categories (ethical, ambiguous, and

unethical) with equal probability. We will refer to the probability that a proposal was sampled from the unethical category as the *signal prevalence*. In the stable condition, the signal prevalence on trials 1-240 was 33.3%. In the *decreasing* condition, we sampled the three categories with unequal probability on some trials. Specifically, in the decreasing condition, the signal prevalence was 33.3% on trials 1-96; 25% on trials 97-120; 16.6% on trials 121-144; 8.3% on trials 145-168; and 4.12% on trials 169-240.

Analyses and Results. Did the decrease in the prevalence of unethical proposals cause participants' concepts of *unethical* to expand? To find out, we fit a binomial generalized linear mixed model to our data in R (*26*) using the lme4 package (*27*). The dependent variable was a binary measure of whether a proposal was accepted or rejected. The independent between-participants variable was the prevalence of unethical proposals (stable or decreasing), and our independent within-participants variables were (a) the trial number (which ranged from 1 to 240) and (b) the objective ethicality rating of each proposal, which were reverse-scored for analysis so that 7 = "This experiment should definitely not be allowed" and 1 = "This experiment should definitely be allowed". We included prevalence, trial number, and objective ethicality rating as fixed effects in our model, along with all interactions. We included as random effects (a) intercepts for participants (who may have entered our study with different thresholds) and (b) slopes for trial number. Model fit was significantly improved by both random slopes for trial, $\chi^2(2) = 63.69$, $p < 0.001$, and random intercepts for participants, $\chi^2(2) = 404.51$, $p < 0.001$. Additionally, the inclusion of the three-way interaction between condition, trial number, and objective ethicality rating significantly improved model fit, $\chi^2(1) = 24.71$, $p < 0.001$. The generalized linear mixed model revealed that a Prevalence X Objective Ethicality

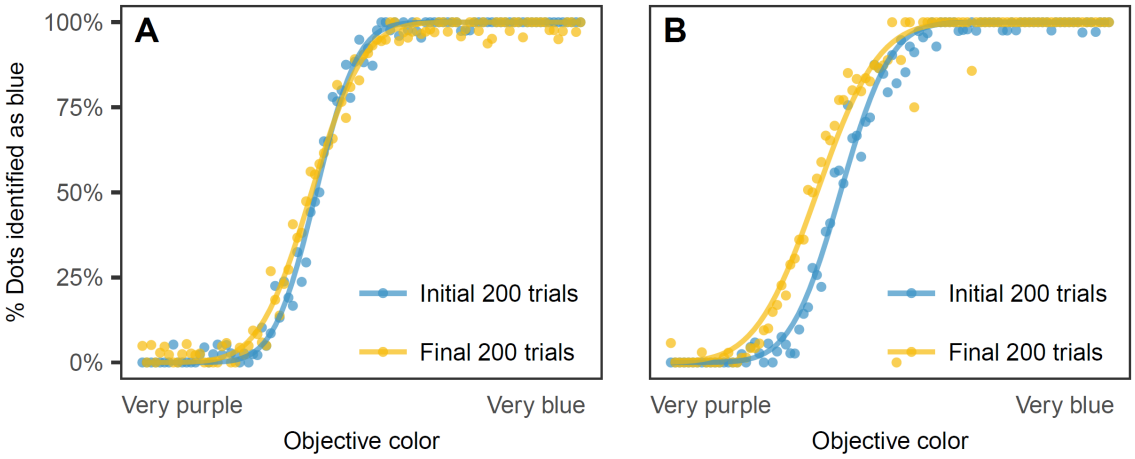
Rating X Trial Number interaction predicted participants' identifications, $b = 5.10$, $SE = 1.02$, $z = 4.98$, *95% CI* [3.09, 7.10], $R^2_{GLMM(c)} = 0.73$.

Fig. S1. Examples of Dots Used in Studies 1-5



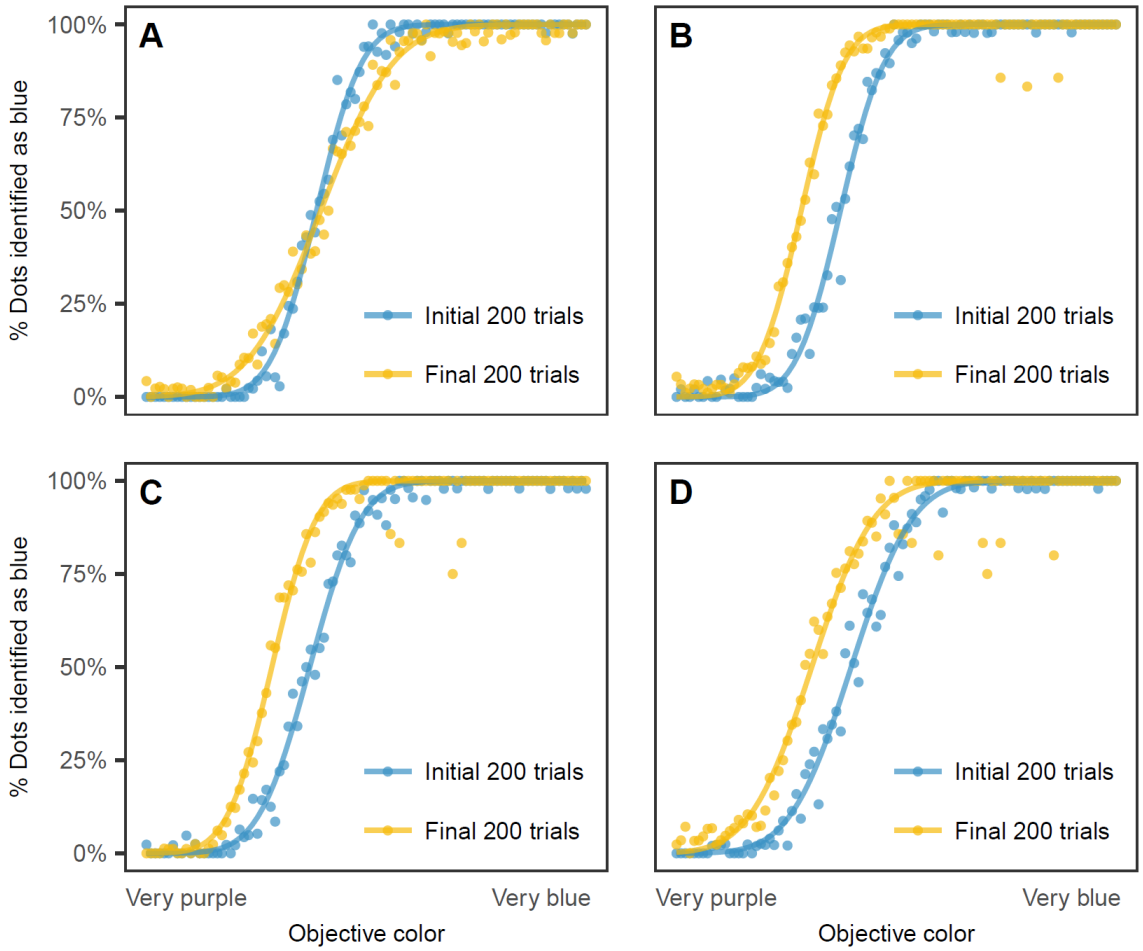
The color spectrum comprised 100 dots ranging from approximately RGB 100-0-155 (very purple) to RGB 0-0-255 (very blue) and this figure shows (from left to right) the 1st, 20th, 40th, 60th, 80th, and 100th dots. The three dots on the left are from the purple spectrum and the three dots on the right are from the blue spectrum.

Fig. S2: Results for Study 2



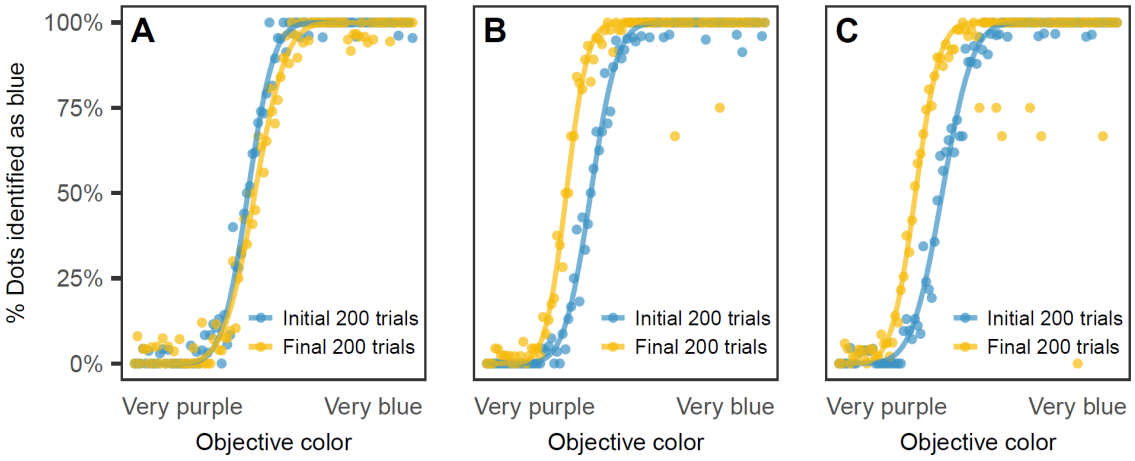
Panel A shows the stable prevalence with warning condition and panel B shows the decreasing prevalence with warning condition. The x-axes show the dot's objective color (i.e., its location on the spectrum) and the y-axes show the percentage of trials on which participants identified that dot as blue.

Fig. S3: Results for Study 3



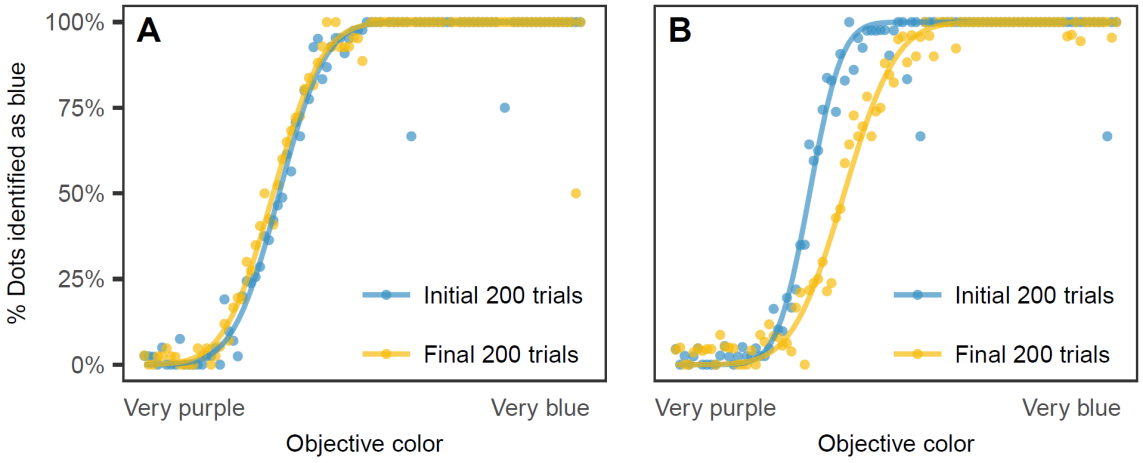
Panel A shows the stable prevalence condition, panel B shows the decreasing prevalence condition, panel C shows the decreasing prevalence + instruction condition, and panel D shows the decreasing prevalence + instruction + incentive condition. The x-axes show the dot's objective color (i.e., its location on the spectrum) and the y-axes show the percentage of trials on which participants identified that dot as blue.

Fig. S4: Results for Study 4



Panel A shows the stable prevalence condition, panel B shows the gradually decreasing prevalence condition, and panel C shows the abruptly decreasing prevalence condition. The x-axes show the dot's objective color (i.e., its location on the spectrum) and the y-axes show the percentage of trials on which participants identified that dot as blue.

Fig. S5: Results for Study 5



Panel A shows the stable prevalence condition, and panel B shows the increasing prevalence condition. The x-axes show the dot's objective color (i.e., its location on the spectrum) and the y-axes show the percentage of trials on which participants identified that dot as blue.

Fig. S6: Examples of Computer-Generated Faces Used in Study 6

The target person continuum ranged from 1 (not threatening) to 60 (very threatening) and this figure shows (from left to right) faces 1, 10, 20, 30, 40 50, and 60**.** The four target persons on the left are from the no threat continuum and the three target persons on the right are from the threat continuum.

References and Notes

1. S. O. Lilienfeld, Microaggressions: Strong claims, inadequate evidence. *Perspect. Psychol. Sci.* **12**, 138–169 (2017). doi:10.1177/1745691616659391 Medline

2. N. Haslam, Concept creep: Psychology's expanding concepts of harm and pathology. *Psychol. Inq.* **27**, 1–17 (2016). doi:10.1080/1047840X.2016.1082418

3. A. Parducci, Category judgment: A range-frequency model. *Psychol. Rev.* **72**, 407–418 (1965). doi:10.1037/h0022602 Medline

4. H. Helson, *Adaptation-Level Theory* (Harper & Row, New York, 1964).

5. C. W. G. Clifford, M. A. Webster, G. B. Stanley, A. A. Stocker, A. Kohn, T. O. Sharpee, O. Schwartz, Visual adaptation: Neural, psychological and computational aspects. *Vision Res.* **47**, 3125–3131 (2007). doi:10.1016/j.visres.2007.08.023 Medline

6. C. Summerfield, F. P. de Lange, Expectation in perceptual decision making: Neural and computational mechanisms. *Nat. Rev. Neurosci.* **15**, 745–756 (2014). doi:10.1038/nrn3838 Medline

7. X. X. Wei, A. A. Stocker, A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. *Nat. Neurosci.* **18**, 1509–1517 (2015). doi:10.1038/nn.4105 Medline

8. N. Stewart, N. Chater, G. D. A. Brown, Decision by sampling. *Cognit. Psychol.* **53**, 1–26 (2006). doi:10.1016/j.cogpsych.2005.10.003 Medline

9. A. Pepitone, M. DiNubile, Contrast effects in judgments of crime severity and the punishment of criminal violators. *J. Pers. Soc. Psychol.* **33**, 448–459 (1976). doi:10.1037/0022-3514.33.4.448

10. G. Rodríguez, S. Blanco, Contrast effect on the perception of the severity of a criminal offence. *Anu. Psicol. Juridica* **26**, 107–113 (2016). doi:10.1016/j.apj.2016.02.001

11. D. L. Chen, T. J. Moskowitz, K. Shue, Decision making under the Gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *Q. J. Econ.* **131**, 1181–1242 (2016). doi:10.1093/qje/qjw017

12. U. Simonsohn, F. Gino, Daily horizons: Evidence of narrow bracketing in judgment from 10 years of M.B.A. admissions interviews. *Psychol. Sci.* **24**, 219–224 (2013). doi:10.1177/0956797612459762 Medline

13. S. Bhargava, R. Fisman, Contrast Effects in Sequential Decisions: Evidence from Speed Dating. *Rev. Econ. Stat.* **96**, 444–457 (2014). doi:10.1162/REST_a_00416

14. U. Simonsohn, New Yorkers Commute More Everywhere: Contrast Effects in the Field. *Rev. Econ. Stat.* **88**, 1–9 (2006).

15. U. Simonsohn, G. Loewenstein, Mistake #37: The effect of previously encountered prices on current housing demand. *Econ. J. (Lond.)* **116**, 175–199 (2006). doi:10.1111/j.1468-0297.2006.01052.x

16. L. Damisch, T. Mussweiler, H. Plessner, Olympic medals as fruits of comparison? Assimilation and contrast in sequential performance judgments. *J. Exp. Psychol. Appl.* **12**, 166–178 (2006). doi:10.1037/1076-898X.12.3.166 Medline

17. M. Carandini, D. J. Heeger, Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* **13**, 51–62 (2011). doi:10.1038/nrn3136 Medline

18. K. Louie, M. W. Khaw, P. W. Glimcher, Normalization is a general neural mechanism for context-dependent decision making. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 6139–6144 (2013). doi:10.1073/pnas.1217854110 Medline

19. A. Parducci, The relativism of absolute judgements. *Sci. Am.* **219**, 84–90 (1968). doi:10.1038/scientificamerican1268-84 Medline

20. J. M. Wolfe, T. S. Horowitz, N. M. Kenner, Cognitive psychology: Rare items often missed in visual searches. *Nature* **435**, 439–440 (2005). doi:10.1038/435439a Medline

21. M. C. Hout, S. C. Walenchok, S. D. Goldinger, J. M. Wolfe, Failures of perception in the low-prevalence effect: Evidence from active and passive visual search. *J. Exp. Psychol. Hum. Percept. Perform.* **41**, 977–994 (2015). doi:10.1037/xhp0000053 Medline

22. M. Roser, The short history of global living conditions and why it matters that we know it (2017); available at https://ourworldindata.org/a-history-of-global-living-conditions-in-5-charts.

23. S. Pinker, *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress* (Viking Press, New York, 2018).

24. YouGov Survey (2015), (available at https://d25d2506sfb94s.cloudfront.net/cumulus_uploads/document/z2knhgzguv/GB_Website.pdf).

25. D. Levari, Prevalence-induced concept change in human judgment. Zenodo, Version v1.0.0 (2018); 10.5281/zenodo.1219833.

26. Y. Chung, S. Rabe-Hesketh, V. Dorie, A. Gelman, J. Liu, A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika* **78**, 685–709 (2013). doi:10.1007/s11336-013-9328-2 Medline

27. R Core Team, *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna, Austria, 2016; http://www.r-project.org/).

28. D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015). doi:10.18637/jss.v067.i01

29. A. Todorov, N. N. Oosterhof, Modeling social perception of faces. *IEEE Signal Process. Mag.* **28**, 117–122 (2011). doi:10.1109/MSP.2010.940006

30. A. Todorov, R. Dotsch, J. M. Porter, N. N. Oosterhof, V. B. Falvello, Validation of data-driven computational models of social perception of faces. *Emotion* **13**, 724–738 (2013). doi:10.1037/a0032335 Medline

31. FantaMorph Abrosoft (Version 3.0) (2014); available at http://www.fantamorph.com/index.html.