FACULTAD DE CIENCIAS EMPRESARIALES Y ECONOMIA

Serie de documentos de trabajo del Departamento de Economía / Department of Economics Working Papers Series

Attitude Polarization: Theory and Evidence

Jean-Pierre BenoÓt
London Business School

Juan Dubra
Universidad de Montevideo

July 22, 2014

Attitude Polarization: Theory and Evidence*

Jean-Pierre Benoît
London Business School

Juan Dubra
Universidad de Montevideo

July 22, 2014

Abstract

Numerous experiments have demonstrated the possibility of attitude polarization. For instance, Lord, Ross & Leper (1979) found that death penalty advocates became more convinced of the deterrent effect of the death penalty while opponents become more convinced of the lack of a deterrent effect, after being presented with the same studies. However, there is an unclear understanding of just what these experiments show and what their implications are. We argue that attitude polarization is consistent with an unbiased evaluation of evidence. Moreover, attitude polarization is even to be expected under many circumstances, in particular those under which experiments are conducted. We also undertake a critical re-examination of several well-known papers.

Keywords: Attitude Polarization; Confirmation Bias; Bayesian Decision Making.
Journal of Economic Literature Classification Numbers: D11, D12, D82, D83

Take two individuals with priors p and q over Θ and f and g over A. The first individual's prior over $\Omega = \Theta \times A$ is the product $p \times f$ and the second has prior $q \times g$.

Take a signal s that has probability $h_{\theta a}(s)$ in state $(\theta, a) \in \Omega$. For $\Theta = \{\theta_1, ..., \theta_n\}$ and $\theta_i < \theta_{i+1}$ for all i. Recall that $sgn(x)$, the sign function, is $1, 0$ or -1 according as $x > 0$, $x = 0$ or $x < 0$. We say that s is **unambiguous** if for all θ_i, θ_j and all a and $\overline{a}$, $sgn\left(h_{\theta_j a}(s) - h_{\theta_i a}(s)\right) = sgn\left(h_{\theta_j \overline{a}}(s) - h_{\theta_i \overline{a}}(s)\right)$. The property says that s is unambiguous if θ_j is more likely than θ_i after s, given a, then the same must be true after a different $\overline{a}$.

We say that there is polarization (after s) if $p(\cdot \mid s) \succeq p \succeq q \succeq q(\cdot \mid s)$ (where $p(\cdot \mid s)$ is the marginal of the posterior over Ω after s).

The following Theorem provides a characterization of what it means for a signal to be unambiguous, and what are its consequences. It is a generalization of Baliga et. al.

Theorem 1 *If signal s is unambiguous, there is no polarization, otherwise there are p, q (with $p = q$) and f, g such that polarization occurs.*

*We thank Gabriel Illanes and Oleg Rubanov. We also thank Vijay Krishna, Wolfgang Pesendorfer, Debraj Ray and Jana Rodríguez-Hertz for valuable comments.

Proof. Suppose that after some unambiguous s there is polarization. In that case, $p(\theta_n \mid s) \geq p(\theta_n)$ and $q(\theta_1) \leq q(\theta_1 \mid s)$. That is,

$$\begin{aligned} p(\theta_n \mid s) &= \frac{p(\theta_n)\sum_a f(a)\, h_{\theta_n a}(s)}{\sum_\theta \sum_a p(\theta) f(a)\, h_{\theta a}(s)} \geq p(\theta_n) \Leftrightarrow \textstyle\sum_a f(a)\, h_{\theta_n a}(s) \geq \sum_\theta p(\theta) \sum_a f(a)\, h_{\theta a}(s) \\ \textstyle\sum_a g(a)\, h_{\theta_1 a}(s) &\geq \textstyle\sum_\theta q(\theta) \sum_a g(a)\, h_{\theta a}(s) \end{aligned}$$

Similarly, from $p(\theta_1 \mid s) \leq p(\theta_1)$ and $q(\theta_n) \geq q(\theta_n \mid s)$ we obtain

$$\begin{aligned} \textstyle\sum_a f(a)\, h_{\theta_1 a}(s) &\leq \textstyle\sum_\theta p(\theta) \sum_a f(a)\, h_{\theta a}(s) \\ \textstyle\sum_a g(a)\, h_{\theta_n a}(s) &\leq \textstyle\sum_\theta q(\theta) \sum_a g(a)\, h_{\theta a}(s) \end{aligned}$$

From the four inequalities

$$\begin{aligned} \textstyle\sum_a f(a)\, h_{\theta_n a}(s) &\geq \textstyle\sum_\theta p(\theta) \sum_a f(a)\, h_{\theta a}(s) \geq \sum_a f(a)\, h_{\theta_1 a}(s) \\ \textstyle\sum_a g(a)\, h_{\theta_1 a}(s) &\geq \textstyle\sum_\theta q(\theta) \sum_a g(a)\, h_{\theta a}(s) \geq \sum_a g(a)\, h_{\theta_n a}(s) \end{aligned} \tag{1}$$

However, if $h_{\theta_n a}(s) > h_{\theta_1 a}(s)$ for any a, by s unambiguous the same must be true for all a, and would therefore imply $\sum_a f(a)\, h_{\theta_n a}(s) > \sum_a f(a)\, h_{\theta_1 a}(s)$ and $\sum_a g(a)\, h_{\theta_n a}(s) > \sum_a g(a)\, h_{\theta_1 a}(s)$, which is a contradiction (similarly if $h_{\theta_n a}(s) < h_{\theta_1 a}(s)$ for any s). Hence, we must have $h_{\theta_n a}(s) = h_{\theta_1 a}(s)$ for all a. This, in turn implies (in equation (1) the first and third terms are equal)

$$\begin{aligned} \textstyle\sum_a f(a)\, h_{\theta_n a}(s) &= \textstyle\sum_\theta p(\theta) \sum_a f(a)\, h_{\theta a}(s) = \sum_a f(a)\, h_{\theta_1 a}(s) \\ \textstyle\sum_a g(a)\, h_{\theta_1 a}(s) &= \textstyle\sum_\theta q(\theta) \sum_a g(a)\, h_{\theta a}(s) = \sum_a g(a)\, h_{\theta_n a}(s). \end{aligned}$$

This implies $p(\theta_n \mid s) = p(\theta_n)$, $q(\theta_n \mid s) = q(\theta_n)$, $p(\theta_1 \mid s) = p(\theta_1)$ and $q(\theta_1 \mid s) = q(\theta_1)$.

Assume now as an induction step that for $i = 1, 2, ..., j, n-j, ..., n$ we have $p(\theta_i \mid s) = p(\theta_i)$ and $q(\theta_i \mid s) = q(\theta_i)$. One can repeat the steps above to obtain the result for $j+1$ and $n-j-1$. This concludes the proof.

To show polarization assume s is not unambiguous, so that there exist $H, L \in \Theta$ and $h, l \in A$ such that $\frac{g^\Theta_{Hh}(s)}{g^\Theta_{Lh}(s)} \geq 1 \geq \frac{g^\Theta_{Hl}(s)}{g^\Theta_{Ll}(s)}$ with one inequality strict. Set

$$\begin{array}{ccc} & \text{Probability of each state in } \Omega & \\ A \downarrow \Theta \rightarrow & H & L \\ h & wz & (1-w)\, z \\ l & w(1-z) & (1-w)(1-z) \end{array} \quad \text{and} \quad \begin{array}{ccc} & \text{Ancillary distribution } g^A & \\ t \downarrow E \rightarrow & (H,h) \text{ or } (L,h) & (H,l) \text{ or } (L,l) \\ t_h & 1 & 0 \\ t_l & 0 & 1 \end{array}.$$

to obtain

$$p(H \mid t_h, s) = \frac{g^\Theta_{Hh} zw}{g^\Theta_{Hh} zw + g^\Theta_{Lh} z(1-w)} > w = p(H \mid t_h) \Leftrightarrow g^\Theta_{Hh} > g^\Theta_{Lh}.$$

Similarly, $p(H \mid t_l, s) < w \Leftrightarrow \frac{g^\Theta_{Hl}(1-z)w}{g^\Theta_{Hl}(1-z)w + g^\Theta_{Ll}(1-z)(1-w)} < w \Leftrightarrow g^\Theta_{Hl} < g^\Theta_{Ll}$. Since one of the two inequalities is strict, we obtain polarization. In this case it obtains with g^A depending only on a and p such that Θ and A are independent. ■

Suppose types are $\theta \sim N(0,1)$ and that $A = \{R, P\}$. If Hannah is Poor, the signal is $\theta + \varepsilon$ where $\varepsilon \sim N(0,1)$, if Hannah is Rich, $\varepsilon \sim N(0,\sigma^2)$ for $\sigma < 1$.

Individual 1 thinks the probability of R is $r > \frac{1}{2}$ and individual 2 thinks it is $q < \frac{1}{2}$.

Fix any signal s. We have $g^{\Theta}_{\theta_j a}(s) > g^{\Theta}_{\theta_i a}(s) \Leftrightarrow |s-\theta_j| < |s-\theta_i|$, which implies that $g^{\Theta}_{\theta_j \bar{a}}(s) > g^{\Theta}_{\theta_i \bar{a}}(s)$ for all other $\bar{a}$. Hence, s is unambiguous.

In the previous proof, we need to check where we use that p and q have common support.

When priors are not independent, an unambiguous signal may lead to polarization.

Example 1 *Consider the following two prior beliefs (where the prior beliefs of truth are* $\frac{11}{16} = 0.6875$ *and* $\frac{5}{8} = 0.625$*), and the unambiguous signal* C

$$\begin{array}{cc} \frac{1}{2} & \frac{1}{4} \\ \frac{3}{16} & \frac{1}{16} \end{array} \quad \begin{array}{cc} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{8} & \frac{1}{8} \end{array} \quad \textit{and } C \textit{ with likelihoods} \quad \begin{array}{cc} \frac{1}{10} & \frac{1}{20} \\ \frac{1}{2} & \frac{9}{20} \end{array}$$

The idea is that in both cases the signal will increase the posterior in each ancillary state, but since the signal indicates that bottom state (L or "free") is so likely, in the second case you are assigning a lot more weight to the "low" original distribution $\left(\frac{1}{2}, \frac{1}{2}\right)$ *(that is the distribution of* $\left(\frac{1}{8}, \frac{1}{8}\right)$ *conditional on the bottom state). The posteriors of Truth are*

$$\frac{\frac{1}{10}\frac{1}{2} + \frac{1}{2}\frac{3}{16}}{\frac{1}{10}\frac{1}{2} + \frac{1}{2}\frac{3}{16} + \frac{1}{20}\frac{1}{4} + \frac{9}{20}\frac{1}{16}} = \frac{46}{59} > \frac{11}{16}$$

$$\frac{\frac{1}{10}\frac{1}{2} + \frac{1}{2}\frac{1}{8}}{\frac{1}{10}\frac{1}{2} + \frac{1}{2}\frac{1}{8} + \frac{1}{20}\frac{1}{4} + \frac{9}{20}\frac{1}{8}} = \frac{18}{29} < \frac{5}{8}$$

So, bottom line, the characterization theorem is false with general beliefs. In particular, an unambiguous signal can still generate polarization.

1 Experts

Our intuition is that if people disagree about how likely the truth of the statement is, and they have observed more or less the same signals, then it must be because they disagree about the likelihood of some ancillary state; then, when they are shown more information like the previous one, those differences in beliefs make them further polarize. The following theorem, proves exactly this intuition, and confirms a finding in the experimental literature, that "experts" are more likely to polarize than people who do not know much about the issue.

People have prior (a probability distribution over the set Ω, where P, Q, R, T are numbers that add to 1):

Prior over Ω

	useful	not useful
selection	P	Q
free	R	T

We assume from the outset that the prior is independent: $\frac{P}{Q} = \frac{R}{T}$ (the prior can be written as the product of a distribution on $\{u, n\} \times \{s, f\}$.

Individuals observe a sample of signals "about" u or n. Let $\mathcal{S}$ be the (finite) sample space of signals S. For each signal $S_i \in \mathcal{S}$ we write its likelihood as

$$\begin{array}{ccc} \multicolumn{3}{c}{\text{Likelihood of } S_i{:}J_\omega(S_i)} \\ & \text{useful} & \text{not useful} \\ \text{selection} & p_i & q_i \\ \text{free} & r_i & t_i \end{array} \tag{2}$$

In addition to the information S in $\mathcal{S}$ they observe signals about s or f, where the individual observes the signal $\sigma \in (0,1)$ with a density

	Probability of σ
selection: in states us or ns signals drawn from	π
free: in states uf or nf signals drawn from	ρ

and $\frac{\pi(\sigma)}{\rho(\sigma)}$ increasing in σ. We assume additionally that $\sigma \to 1$ is completely informative about s ($\lim_{\sigma \to 1} \frac{\pi(\sigma)}{\rho(\sigma)} = \infty$) and $\sigma \to 0$ is completely informative about f ($\lim_{\sigma \to 0} \frac{\pi(\sigma)}{\rho(\sigma)} = 0$).

After they observe this information, they observe a Common signal C with likelihoods

$$\begin{array}{ccc} \multicolumn{3}{c}{\text{Likelihoods of } C \text{ for ech } \omega\in\Omega} \\ & \text{useful} & \text{not useful} \\ \text{selection} & p & q \\ \text{free} & r & t \end{array} . \tag{3}$$

We postulate the following assumption about signals S:

Assumption 1. Weak Ambiguity (WA). Signal S_i satisfies Weak A1 if $p_i t_i > q_i r_i$.
Ambiguity (A). We say that C is ambiguous if $p > q$ and $t > r$.

Theorem 2 *Take two people who have observed the same S (say, two experts who know the whole "body of evidence" about an issue). Assume the prior is independent. We know C must satisfy ambiguity and we are told that C is a typical signal, so we also assume S satisfies weak ambiguity.*

There exists v_S such that $P(u \mid S,\sigma,C) > P(u \mid S,\sigma) \Leftrightarrow P(u \mid S,\sigma) > v_S$.

Proof. Step 1. Individual increases belief after C iff high σ; define cutoff σ_B. For any probability distribution B over Ω we have $B(u \mid C,\sigma) > B(u \mid \sigma)$ (for C satisfying Ambiguity) iff

$$\frac{pB(us \mid \sigma) + rB(uf \mid \sigma)}{qB(ns \mid \sigma) + tB(nf \mid \sigma)} > \frac{B(us \mid \sigma) + B(uf \mid \sigma)}{B(ns \mid \sigma) + B(nf \mid \sigma)} \Leftrightarrow$$

$$\left(pB(us)\frac{\pi(\sigma)}{\rho(\sigma)} + rB(uf)\right)\left(B(ns)\frac{\pi(\sigma)}{\rho(\sigma)} + B(nf)\right) > \left(B(us)\frac{\pi(\sigma)}{\rho(\sigma)} + B(uf)\right)\left(qB(ns)\frac{\pi(\sigma)}{\rho(\sigma)} + tB\right.$$

Letting

$$f(\sigma) \equiv B(ns)B(us)\frac{\pi(\sigma)}{\rho(\sigma)}(p-q) + B(us)B(nf)p - B(ns)B(uf)q - B(us)B(nf)t + B(uf)B(ns)r$$

equation (4) can be written as

$$\frac{\pi(\sigma)}{\rho(\sigma)} f(\sigma) > B(uf)\, B(nf)\,(t-r)$$

We have that $f(\sigma)$ is increasing in σ. As $\sigma \to 0$, $f(\sigma)$ converges to a constant, so the lhs converges to $0 < B(uf)\, B(nf)\,(t-r)$. As $\sigma \to 1$, $\frac{\pi(\sigma)}{\rho(\sigma)} f(\sigma) \to \infty$. Since $\frac{\pi(\sigma)}{\rho(\sigma)}$ and $f(\sigma)$ are increasing, there exists a unique $\sigma_B \in (0,1)$ such that $B(u \mid C, \sigma) > B(u \mid \sigma) \Leftrightarrow \sigma > \sigma_B$. For such a σ_B, $B(u \mid C, \sigma_B) = B(u \mid \sigma_B) \equiv \mu_B$. $\boxdot$

From Step 1, there exists a σ_S such that $P(u \mid S, C, \sigma_S) = P(u \mid S, \sigma_S)$ and $P(u \mid S, C, \sigma) > P(u \mid S, \sigma)$ if and only if $\sigma > \sigma_S$. Define v_S as $v_S = P(u \mid S, \sigma_S)$. Then, from Lemma 1 we know if S is weakly ambiguous beliefs $P(u \mid S, \sigma)$ are increasing in σ, so that $P(u \mid S, \sigma) > v_S \Leftrightarrow \sigma > \sigma_S \Leftrightarrow P(u \mid S, C, \sigma) > P(u \mid S, \sigma)$. ■

Lemma 1 *Suppose the prior is independent. S_i satisfies WA if and only if posteriors of u increase with σ. In particular, $p_i t_i > q_i r_i \Leftrightarrow P(u \mid S, \sigma)$ is strictly increasing in σ (and $<$ iff strictly decreasing). The σ_e for which $P(u \mid S, \sigma_e) = P(u \mid S)$ is defined by $\frac{\pi(\sigma_e)}{\rho(\sigma_e)} = 1$.*

Proof. We have that for a signal S with likelihoods p_i, q_i, t_i, r_i

$$P(u \mid \sigma, S) = P(us \mid \sigma, S) + P(uf \mid \sigma, S) = \frac{P\frac{\pi(\sigma)}{\rho(\sigma)} p_i + R r_i}{P\frac{\pi(\sigma)}{\rho(\sigma)} p_i + R r_i + T t_i + Q\frac{\pi(\sigma)}{\rho(\sigma)} q_i}$$

which increases in σ iff the following expression increases in σ

$$X = \frac{\frac{\pi(\sigma)}{\rho(\sigma)} P p_i + R r_i}{T t_i + \frac{\pi(\sigma)}{\rho(\sigma)} Q q_i}.$$

The derivative of this expression wrt $\frac{\pi(\sigma)}{\rho(\sigma)}$ is

$$\frac{dX}{d\frac{\pi(\sigma)}{\rho(\sigma)}} = \frac{P p_i T t_i - Q q_i R r_i}{\left(T t_i + Q q_i \frac{\pi(\sigma)}{\rho(\sigma)}\right)^2} > 0 \Leftrightarrow p_i P t_i T > q_i Q r_i R. \tag{5}$$

When $\frac{\pi(\sigma)}{\rho(\sigma)} = 1$ we get

$$P(u \mid \sigma, S) = \frac{P\frac{\pi(\sigma)}{\rho(\sigma)} p_i + R r_i}{P\frac{\pi(\sigma)}{\rho(\sigma)} p_i + R r_i + T t_i + Q\frac{\pi(\sigma)}{\rho(\sigma)} q_i} = \frac{P p_i + R r_i}{P p_i + R r_i + T t_i + Q q_i} = P(u \mid S).$$

■

2 Different signals: counterexample

The question is then whether our results go through when people have observed different signals. That is not necessarily the case. We now present an example to illustrate.

Consider the following signals.

$$\overset{s_1}{\begin{matrix} \frac{3}{7}+\varepsilon & \frac{3}{7}-\varepsilon \\ \frac{2}{7}+\varepsilon & \frac{4}{7}-\varepsilon \end{matrix}} \text{ and } \overset{s_2}{\begin{matrix} \frac{4}{7}-\varepsilon & \frac{2}{7}+\varepsilon \\ \frac{3}{7}-\varepsilon & \frac{3}{7}+\varepsilon \end{matrix}} \text{ and } \overset{s_3}{\begin{matrix} 0 & \frac{2}{7} \\ \frac{2}{7} & 0 \end{matrix}} \text{ and } \overset{C}{\begin{matrix} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} \end{matrix}}$$

The prior is uniform, and people receive signals σ about Selection or not according to distributions π (when the state is selection) and ρ (when it is no selection).

Notice first that signals s_1 and s_2 do not affect the belief in Selection (which I call H sometimes):

$$P(H \mid \sigma, s_1) = \frac{\frac{3}{7}\pi\frac{1}{4}+\frac{3}{7}\pi\frac{1}{4}}{\frac{3}{7}\pi\frac{1}{4}+\frac{3}{7}\pi\frac{1}{4}+\frac{2}{7}\rho\frac{1}{4}+\frac{4}{7}\rho\frac{1}{4}} = \frac{\pi}{\pi+\rho} = \frac{\pi\frac{1}{4}+\pi\frac{1}{4}}{\pi\frac{1}{4}+\pi\frac{1}{4}+\rho\frac{1}{4}+\rho\frac{1}{4}} = P(H \mid \sigma)$$

(and similarly for s_2; the trick was having the rows add up to the same number).

With s_1 "no one" believes in T with probability larger than $\frac{1}{2}$, because you have to be "certain" that the state is Selection.

With s_2 the opposite is true: all believe in T with probability greater than $\frac{1}{2}$. I don't consider s_3 because it has probability 0 in state TH.

Setting $\varepsilon = 0$ for simplicity, we now find the cutoffs for σ such that after the common signal C individuals increase their beliefs.

$$\begin{aligned} P(T \mid s_i, C, \sigma) &= \frac{pp_i\pi(\sigma)\frac{1}{4}+rr_i\rho(\sigma)\frac{1}{4}}{pp_i\pi(\sigma)\frac{1}{4}+rr_i\rho(\sigma)\frac{1}{4}+qq_i\pi(\sigma)\frac{1}{4}+tt_i\rho(\sigma)\frac{1}{4}} \\ P(T \mid s_i, C, \sigma) &= \frac{p_i\pi(\sigma)\frac{1}{4}+r_i\rho(\sigma)\frac{1}{4}}{p_i\pi(\sigma)\frac{1}{4}+r_i\rho(\sigma)\frac{1}{4}+q_i\pi(\sigma)\frac{1}{4}+t_i\rho(\sigma)\frac{1}{4}} \end{aligned}$$

so

$$\begin{aligned} P(T \mid s_1, C, \sigma) &= \frac{\frac{1}{2}\frac{3}{7}\frac{\pi(\sigma)}{\rho(\sigma)}+\frac{1}{4}\frac{2}{7}}{\frac{1}{2}\frac{3}{7}\frac{\pi(\sigma)}{\rho(\sigma)}+\frac{1}{4}\frac{2}{7}+\frac{1}{4}\frac{3}{7}\frac{\pi(\sigma)}{\rho(\sigma)}+\frac{1}{2}\frac{4}{7}} \text{ and } P(T \mid s_1, \sigma) = \frac{\frac{3}{7}\frac{\pi(\sigma)}{\rho(\sigma)}+\frac{2}{7}}{\frac{3}{7}\frac{\pi(\sigma)}{\rho(\sigma)}+\frac{2}{7}+\frac{3}{7}\frac{\pi(\sigma)}{\rho(\sigma)}+\frac{4}{7}} \\ P(T \mid s_2, C, \sigma) &= \frac{\frac{1}{2}\frac{4}{7}\frac{\pi(\sigma)}{\rho(\sigma)}+\frac{1}{4}\frac{3}{7}}{\frac{1}{2}\frac{4}{7}\frac{\pi(\sigma)}{\rho(\sigma)}+\frac{1}{4}\frac{3}{7}+\frac{1}{4}\frac{2}{7}\frac{\pi(\sigma)}{\rho(\sigma)}+\frac{1}{2}\frac{3}{7}} \text{ and } P(T \mid s_2, C, \sigma) = \frac{\frac{4}{7}\frac{\pi(\sigma)}{\rho(\sigma)}+\frac{3}{7}}{\frac{4}{7}\frac{\pi(\sigma)}{\rho(\sigma)}+\frac{3}{7}+\frac{2}{7}\frac{\pi(\sigma)}{\rho(\sigma)}+\frac{3}{7}} \end{aligned}$$

Letting $\frac{\pi(\sigma)}{\rho(\sigma)} = x$, we find the x that solves $P(T \mid s_i, C, \sigma) = P(T \mid s_i, \sigma)$:

$$\begin{aligned} \frac{\frac{1}{2}\frac{3}{7}x+\frac{1}{4}\frac{2}{7}}{\frac{1}{2}\frac{3}{7}x+\frac{1}{4}\frac{2}{7}+\frac{1}{4}\frac{3}{7}x+\frac{1}{2}\frac{4}{7}} &= \frac{\frac{3}{7}x+\frac{2}{7}}{\frac{3}{7}x+\frac{2}{7}+\frac{3}{7}x+\frac{4}{7}} \Leftrightarrow x_1 = \frac{2}{3}\sqrt{2} = 0.942\,81 \\ \frac{\frac{1}{2}\frac{4}{7}x+\frac{1}{4}\frac{3}{7}}{\frac{1}{2}\frac{4}{7}x+\frac{1}{4}\frac{3}{7}+\frac{1}{4}\frac{2}{7}x+\frac{1}{2}\frac{3}{7}} &= \frac{\frac{4}{7}+\frac{3}{7}}{\frac{4}{7}x+\frac{3}{7}+\frac{2}{7}x+\frac{3}{7}} \Leftrightarrow x_2 = \frac{1}{24}\sqrt{541}+\frac{1}{24} = 1.010\,8 \end{aligned}$$

We then have:

- those who believe in T with probability greater than $\frac{1}{2}$ are those who observe s_2 (no one who received s_1), who have a probability of $\frac{4}{7}$; of those, those with $\frac{\pi(\sigma)}{\rho(\sigma)} > 1.01$ increase their belief.

- those who believe in T with probability less than $\frac{1}{2}$ are those who observe s_1, who have a probability of $\frac{3}{7}$; of those, all who have $\frac{\pi(\sigma)}{\rho(\sigma)} > 0.94$ increase their beliefs.

Hence, the proportion of those who increase their belief is less for those who believe in T highly, than for those who do not believe in T : the probability of σ with $\frac{\pi(\sigma)}{\rho(\sigma)} > 1.01$ is lower than the prob of σ with $\frac{\pi(\sigma)}{\rho(\sigma)} > 0.94$

3 Fixes

There are two reasons why the previous example doesn't work. The first is quite simple: it is not really a counterexample to our intuition, since there is not enough variation in the belief in Selection (or in H). The following theorem shows that when there is enough variation in that belief, then there is polarization.

People have prior (a probability distribution over the set Ω, where $a, b \in (0,1)$):

Prior over Ω

	True	False
High	ab	$a(1-b)$
Low	$(1-a)b$	$(1-a)(1-b)$

There is a set of signals $\mathcal{S}$ and a collection of likelihood functions f_ω for $\omega \in \Omega$ such that $f_\omega(S)$ is the probability that signal $S \in \mathcal{S}$ will happen in state ω. For each signal S_i we generally let $p_i = f_{TH}(S_i)$, $q_i = f_{FH}(S_i)$, $r_i = f_{TL}(S_i)$ and $t_i = f_{FL}(S_i)$.

In addition to these signals, individuals also receive one of two signals $\{h, l\}$ about the ancillary state, where the probability of signal h is given by

$$P_\omega(h) = \begin{cases} \pi & \text{if } \omega = TH, FH \\ \rho & \text{if } \omega = TL, FL \end{cases} \quad \text{for } \pi > \rho.$$

We are interested in the informativeness of signals h or l : how are the beliefs about H or L affected by the signals. Thus, we analyze

$$P(H \mid S, h) = P(TH \mid h) + P(FH \mid h) = \frac{p\pi ab + q\pi a(1-b)}{p\pi ab + q\pi a(1-b) + r\rho(1-a)b + t\rho(1-a)(1-b)}. \tag{6}$$

This posterior is a monotone function (which converges to 1 as $\pi, 1-\rho \to 1$) of

$$\frac{a}{1-a}\frac{\pi}{\rho}\frac{pb + q(1-b)}{rb + t(1-b)}.$$

and similarly, $P(H \mid S, l)$ is a monotone (decreasing) function of

$$\frac{a}{1-a}\frac{1-\pi}{1-\rho}\frac{pb + q(1-b)}{rb + t(1-b)}.$$

Therefore, signals about the ancillary issue are more informative when π increases and ρ decreases.

For the common signal, we assume that its likelihood in each state is

Likelihood of C for each state in Ω

	True	False
High	P	Q
Low	R	T

for $P > Q$, $T > R$ (that is C is ambiguous).

We are interested in the following two quantities: the proportion of people with beliefs greater than v who increase their beliefs (after C), and the proportion of people with beliefs less than v who increase their beliefs; we want to show

$$\frac{\sum P(S_i)\, P\left(a > a_v^i, a_C^i\right)}{\sum P(S_i)\, P\left(a > a_v^i\right)} > \frac{\sum P(S_i)\, P\left(a_v^i > a > a_C^i\right)}{\sum P(S_i)\, P\left(a < a_v^i\right)} \tag{7}$$

and we want to show that the expression on the left is greater than that on the right.

Theorem 3 *If there is enough variation in the beliefs about the ancillary issue (if π is sufficiently large and ρ sufficiently low), and people have observed ambiguous signals, then the proportion of people whose beliefs are larger than $b = P(T)$ and increase them after observing C is larger than the proportion of people whose beliefs are lower than b and increase them after observing C. In particular, for π and $1-\rho$ large enough, all those above v increase and all those below b decrease their beliefs after C.*

Proof. First notice that ambiguity of S implies that $P(T \mid H, S) > b > P(T \mid L, S)$. Next, we know that if π is large enough and ρ is low enough, continuity of beliefs in π and ρ ensure that all those who observe h will have beliefs larger than b:

$$P(T \mid S, h) = \frac{p\pi ab + r\rho(1-a)b}{p\pi ab + r\rho(1-a)b + q\pi a(1-b) + t\rho(1-a)(1-b)} \underset{\pi,1-\rho\to 1}{\rightarrow} \frac{pb}{pb + q(1-b)} = P(T \mid H, S) >$$

Finally, note that the posterior belief after h, C converges (when $\pi, 1-\rho \to 1$) to the belief after C in state H:

$$P(T \mid S, h, C) = \frac{Pp\pi ab + Rr\rho(1-a)b}{Pp\pi ab + Rr\rho(1-a)b + Qq\pi a(1-b) + Tt\rho(1-a)(1-b)} \underset{\pi,1-\rho\to 1}{\rightarrow} \frac{Ppb}{Ppb + Qq(1-b)} =$$

which is larger than $\frac{pb}{pb+q(1-b)} = P(T \mid H, S)$. Hence, for $\pi, 1-\rho$ close to 1 we obtain $P(T \mid S, h, C) > P(T \mid S, h)$.

We conclude that those who observe signal h, if π is large enough and ρ small enough, have beliefs larger than b and increase their beliefs after C. Those who observe signal l have a belief lower than b and decrease their beliefs. ■

In the previous result, there are only two signals, and "enough variation", but that is an extreme case for the more general case that there are many signals, and those that are more informative have enough probability.

But even without the "enough variation", there is something else that makes the previous example in Section 2 really not a counterexample: the signals are not really ambiguous. Consider for example signal s_1 with likelihoods:

$$\begin{matrix} \frac{3}{7}+\varepsilon & \frac{3}{7}-\varepsilon \\ \frac{2}{7}+\varepsilon & \frac{4}{7}-\varepsilon \end{matrix}.$$

It is not really ambiguous, since it is basically "bad news": it is neutral when the state is H, and bad news in state L. We therefore introduce the notion that there must be some symmetry in that if the signal is good news in one state, it must be "comparably" bad news in the other state. With likelihoods as in (2):

Weak Ambiguity. Signal S_i satisfies Weak Ambiguity if $p_i t_i > q_i r_i$.

We now repeat some of the previous material, and show that when signals are weakly symmetric, polarization holds.

People have prior (a probability distribution over the set Ω):

$$\begin{array}{ccc} & \multicolumn{2}{c}{\text{Prior over } \Omega} \\ & \text{T} & \text{F} \\ \text{H} & ya & y(1-a) \\ \text{L} & (1-y)a & (1-y)(1-a) \end{array}$$

and they observe a sample of signals "about" T or F. Let $\mathcal{S}$ be the (finite) sample space of signals S. For each signal $S \in \mathcal{S}$ we write its likelihood as

$$\begin{array}{ccc} & \multicolumn{2}{c}{\text{Likelihood of } S_i\text{:}J_\omega(S_i)} \\ & \text{useful} & \text{not useful} \\ \text{selection} & p & q \\ \text{free} & r & t \end{array} \tag{8}$$

In addition to the information S in $\mathcal{S}$ they observe signals about s or f, where the individual observes the signal $\sigma \in (0,1)$ with a density

	Probability of σ
selection: in states us or ns signals drawn from	π
free: in states uf or nf signals drawn from	ρ

and $\frac{\pi(\sigma)}{\rho(\sigma)}$ increasing in σ (we might get rid of this which adds nothing). We assume additionally that $\sigma \to 1$ is completely informative about s ($\lim_{\sigma\to 1} \frac{\pi(\sigma)}{\rho(\sigma)} = \infty$) and $\sigma \to 0$ is completely informative about f ($\lim_{\sigma\to 0} \frac{\pi(\sigma)}{\rho(\sigma)} = 0$) **this is just to simplify**.

Note that after observing a signal σ their beliefs become (for example, for TH), for $x = \frac{\pi y}{\pi y + \rho(1-y)}$

$$\frac{\pi y a}{\pi y a + \pi y (1-a) + \rho(1-y)a + \rho(1-y)(1-a)} = \frac{xa}{xa + x(1-a) + (1-x)a + (1-x)(1-a)} = xa$$

and similarly for the other states:

$$\begin{array}{ccc} & \text{T} & \text{F} \\ \text{H} & xa & x(1-a) \\ \text{L} & (1-x)a & (1-x)(1-a) \end{array} \tag{9}$$

So from now on, we assume everybody has a prior as in (9), with the same a for everybody, but a distribution of x, which is derived from the distribution of σ.

Suppose v is a belief in T that can be attained when signal S (with likelihoods $\frac{pq}{rt}$), a value between the beliefs when H is known and when L is known: $\frac{pa}{pa+q(1-a)} > v > \frac{ra}{ra+t(1-a)}$. Then, the cutoff $x = \frac{\pi y}{\pi y+\rho(1-y)}$ (that is, we are indirectly defining the cutoff σ) for which $P(T \mid S, x_S) = v$ is defined by

$$x_S = \frac{t_S \frac{v}{1-v}\frac{1-a}{a} - r_S}{p_S - r_S + \frac{v}{1-v}\frac{1-a}{a}(t_S - q_S)} \equiv \frac{t_S g - r_S}{p_S - g q_S + t_S g - r_S} \tag{10}$$

After they observe this information, they observe a Common signal C with likelihoods

$$\overset{\text{Likelihoods of } C \text{ for ech } \omega\in\Omega}{\begin{array}{ccc} & \text{T} & \text{F} \\ \text{H} & P & Q \\ \text{L} & R & T \end{array}} \quad . \tag{11}$$

We postulate the following assumption about signals S and C:

Weak Ambiguity. Signal S_i satisfies Weak Ambiguity if $p_i t_i > q_i r_i$.
Ambiguity. We say that C is ambiguous if $p > q$ and $t > r$.
Weak Symmetry. We say that S_i is weakly symmetric if $p_i = bq_i$ and $t_i = br_i$.

Theorem 4 *Assume S is WA, C is ambiguous, and both are weakly symmetric. Then, there is attitude polarization:*

$$P\{S,\sigma : P(T \mid S,\sigma,C) > P(T \mid S,\sigma) \mid P(T \mid S,\sigma) > P(T)\} > P\{S,\sigma : P(T \mid S,\sigma,C) > P(T \mid S,\sigma) \mid P$$

Proof. From equation (10), if S and C are weakly symmetric, and setting $v = a$

$$P(T \mid S,\sigma) \geq v \Leftrightarrow x_S^v \geq \frac{t_S \frac{v}{1-v}\frac{1-a}{a} - r_S}{p_S - r_S + \frac{v}{1-v}\frac{1-a}{a}(t_S - q_S)} = \frac{br - r}{bq - r + br - q} = \frac{r}{q+r}. \tag{12}$$

If S with likelihoods $\frac{pq}{qp}$and C with likelihoods $\frac{PQ}{QP}$ are weakly symmetric, $P(T \mid S,\sigma,C) > P(T \mid S,\sigma)$ happens iff

$$\begin{aligned} \frac{BQbqxa + Rra(1-x)}{BQbqxa + Rra(1-x) + Qqx(1-a) + BRbr(1-x)(1-a)} &> \frac{bqxa + ra(1-x)}{bqxa + ra(1-x) + qx(1-a) + br(1-} \\ (BQbqxa + Rra(1-x))(qx(1-a) + br(1-x)(1-a)) &> (bqxa + r(1-x)a)(Qqx(1-a) + BRb \\ (BQbqx + Rr(1-x))(qx + br(1-x)) &> (bqx + r(1-x))(Qqx + BRbr(1-x)) \end{aligned}$$

It is easy to check that $P(T \mid S,\sigma,C) > P(T \mid S,\sigma) \Leftrightarrow \sigma > \sigma_S^C \Leftrightarrow x > x_S^C \in (0,1)$ (there is a cutoff for x or σ such that the individual revises upwards iff his belief in H, prior to observing S is high enough).

Suppose $Q > R$, then an individual who believes in T exactly $v = a = P(T)$, revises upwards: plugging x_S from (12) in (13) we obtain

$$\begin{aligned} (BQbqr + Rrq)(qr + brq) &> (bqr + rq)(Qqr + BRbrq) \Leftrightarrow \\ (BQb + R)(1+b) &> (b+1)(Q + BRb) \Leftrightarrow BQb + R > Q + BRb \Leftrightarrow Q > R. \end{aligned}$$

This means that $x_S^v > x_S^C$ for all S. So all those who believe more than v (those that have $x > x_S^v$) also revise upward $x > x_S^v > x_S^C$. At the same time, all those with $x < x_S^C$ believe in T less than v, and revise downward after C. The inequality in the statement of the theorem then satisfied (as $1 > z$ for some positive z).

If $Q < R$, an individual who believes in T exactly $v = P(T)$ revises downward, which means $x_S^v < x_S^C$. Then all those with beliefs in T lower than v revise downward, while those with $x > x_S^C$ believe in T more than v and revise upward, which establishes the inequality in the theorem (as $z > 0$, for some $z < 1$). ■

Note the following generalization, that needs to be checked.

Theorem 5 *Assume S is WA, C is ambiguous, and both are weakly symmetric. Then, there is attitude polarization: for any v,*

$$P\{S,\sigma : P(T \mid S,\sigma,C) > P(T \mid S,\sigma) \mid P(T \mid S,\sigma) > v\} > P\{S,\sigma : P(T \mid S,\sigma,C) > P(T \mid S,\sigma) \mid P(T \mid$$

Proof. Set $g = \frac{v}{1-v}\frac{1-a}{a}$ and plug $x_S = \frac{(bg-1)r}{bq-r+g(br-q)}$ from (12) in (**??**) to obtain

$$\begin{aligned}
(BQbqx + Rr(1-x))(qx + br(1-x)) &> (bqx + r(1-x))(Qqx + BRbr(1 \\
(BQbq(bg-1)r + Rr(b-g)q)(q(bg-1)r + br(b-g)q) &> (bq(bg-1)r + r(b-g)q)(Qq(b \\
(BQb(bg-1) + R(b-g))((bg-1) + b(b-g)) &> (b(bg-1) + (b-g))(Q(bg-1) \\
\left(b^2-1\right)\left(Bg(Q-R)b^2 + \left(R - BQ - Qg^2 + BRg^2\right)b + (Q-R)g\right) &> 0
\end{aligned}$$

Note that if $Q - R > 0$, the coefficient in the quadratic term on b is positive, as is the independent term, meaning to say that the equation is satisfied for all b (that is, for every signal S). This means that for all S, an individual who believes in T exactly v will revise upwards, as will all those with beliefs larger than v (Similarly, for $Q - R < 0$, the equation is satisfied for no b, which means that those who believe in T exactly v will revise downwards (while we know that those with high enough belief in T will revise upwards). ■

4 Plous: intuition for the basic step

A paper by Plous tests directly our intuition: he asks whether backup systems (checks) are important in nuclear power safety, or whether having a low rate of potential accidents is more important. He then checks that those who believe in backups are more likely (than those who believe low rates are important) to increase their belief that nuclear power is safe after receiving a report of another instance of a failure fixed by backup systems.

We now show that this intuition works in our model if we assume that the common signal C is symmetric.

Symmetric. We say that the signal C is symmetric if its likelihoods are $\substack{BQ,Q \\ Q,BQ}$.

Start with an independent prior,

	True	False
High	xz	$x(1-z)$
Low	$(1-x)z$	$(1-x)(1-z)$

where z is the same for all, but x is not (because they have observed different σs).

Their posteriors are then a constant times

	True	False
High	$bqxz$	$qx(1-z)$
Low	$r(1-x)z$	$br(1-x)(1-z)$

A person revises up after C (with $B > 1$) if and only if

$$\begin{aligned}\frac{BQbqxz + Rr(1-x)z}{Qqx(1-z) + BRbr(1-x)(1-z)} &> \frac{bqxz + r(1-x)z}{qx(1-z) + br(1-x)(1-z)} \Leftrightarrow \frac{BQbqx + Rr(1-x)}{Qqx + BRbr(1-x)} > \frac{bqx + r}{qx + br} \\ \frac{Bbqx + r(1-x)}{qx + Bbr(1-x)} &> \frac{bqx + r(1-x)}{qx + br(1-x)} \Leftrightarrow qx > r(1-x).\end{aligned}$$

We have that

$$\begin{aligned}P(H \mid S, \sigma) &> \frac{1}{2} \Leftrightarrow bqxz + qx(1-z) > r(1-x)z + br(1-x)(1-z) \Leftrightarrow qx(bz + 1 - z) > r(1-x)(z \\ \frac{qx}{r(1-x)} &> \frac{z + b(1-z)}{bz + 1 - z}\end{aligned}$$

Theorem 6 *Plous. Fix define* $\overline{B} = \left\{S, \sigma : P(H \mid S, \sigma) > \frac{1}{2}\right\}$ *and its complement* $\underline{B}$*. If S is WS, C is symmetric and ambiguous (and S is similar)*

$$P\left(S, \sigma : P(T \mid S, \sigma, C) > P(T \mid S, \sigma) \mid \overline{B}\right) > P(S, \sigma : P(T \mid S, \sigma, C) > P(T \mid S, \sigma) \mid \underline{B}) \tag{16}$$

That is, those with higher belief in H (in "selection") are more likely to update up after C.

Proof. If $z \geq \frac{1}{2}$, those who have $P(H \mid S, \sigma) \leq \frac{1}{2}$ have $qx \leq r(1-x)$ (by (15) and $b \geq 1$), so no one revises up (by 14), so the rhs of (16) is 0, while the rhs is positive (for x close to 1, $qx > r(1-x)$, which ensures that those individuals believe in H more than $\frac{1}{2}$ and revise up).

If $z < \frac{1}{2}$, those who have $P(H \mid S, \sigma) > \frac{1}{2}$ have $qx > r(1-x)$ (by (15) and $b \geq 1$), which ensures that all revise up (by 14), so the lhs of (16) is 1, while the rhs is less than 1 (for x close to 0, the individual believes in H less than $\frac{1}{2}$, and revises down). ■