When Do Extraordinary Claims Give Extraordinary Evidence?

Robin Hanson*
Department of Economics
George Mason University†

March 2007

Abstract

Extraordinary claims require extraordinary evidence. But on uninteresting topics, surprising claims usually *are* surprising evidence; we rarely make claims without sufficient evidence. On interesting topics, however, we can have interests in exaggerating or downplaying our evidence, and our actions often deviate from our interests. In a simple model of noisy humans reporting on extraordinary evidence, we find that extraordinary claims from low noise people are extraordinary evidence, but such claims from high noise people are not; their claims are more likely unusual noise than unusual truth. When people are organized into a reporting chain, noise levels grow exponentially with chain length; long chains seem incapable of communicating extraordinary evidence.

Introduction

People who make surprising claims are often told "extraordinary claims require extraordinary evidence." This maxim, however, seems to neglect the fact that for someone known to be reliable, in the sense that he would only make a surprising claim if he had actually seen surprising evidence, his extraordinary claim would *be* extraordinary evidence to others.

For example, you would assign a very low prior probability to your friend telling you that she met her 5'2" tall second cousin last Tuesday at 8:47am at 11 feet northwest of the smaller statue in a certain square. Nevertheless, after she made such a claim to you, you

*For comments I thank Stuart Armstrong, Hal Finney, James Miller, Lubos Motl, Mark Resnick, Barkley Rosser, John Thacker, Eliezer Yudkowsky, students in my Spring 2007 Graduate Industrial Organization class, and commentors on Jan. 18, 21, and Mar. 27 2007 posts at *OvercomingBias.com*. For financial support, I thank the Center for Study of Public Choice and the Mercatus Center.

†*rhanson@gmu.edu* http://hanson.gmu.edu 703-993-2326 FAX: 703-993-2323 MS 1D3, Carow Hall, Fairfax VA 22030

would likely place a high probability that the event happened as claimed. On topics like this, most people are fairly reliable.

For many kinds of more interesting topics, however, people can be less reliable. Not only do we not always do exactly what is in our best interest, but on such topics we often have interests that bias what we say. Furthermore, most communications we could receive from people making extraordinary claims are filtered through several levels of intermediaries, each staffed by these same unreliable humans. With such distortions, extraordinary claims need not rise to the level of extraordinary evidence.

In this note we analyze a simple model of this situation. We consider a single parameter, such as the money a business venture will make, the energy a device could release, or the expected number of people to be killed in a looming disaster. We assume a power law distribution for this parameter, so that large values become both especially unlikely as well as interesting, i.e., "extraordinary."

In the model, a person gets a noisy signal about this parameter, after which he updates his beliefs. We find that his median estimate should be lower than his signal, and that this difference should be linear in his signal variance. This corrects for the fact that with higher signal variance, high signals are more often due to signal error, instead of high truth. With enough signal noise, a truly extraordinary parameter value almost never shows itself clearly in an extraordinary signal.

If expertise and context is required to interpret a signal, then a person cannot simply pass on his signal "word for word" to others; he must instead summarize what his signal implies about the parameter of interest. This process of summarizing, however, also allows room for error and distortion.

We assume that each person does not always act optimally, though he is more likely to take actions which better achieve his interest. We also assume that while each person has a strong incentive to make forecasts which are likely to be correct, he also has a small bias incentive to either exaggerate or downplay the parameter value. Only he knows the exact strength of this bias; observers know only the distribution from which his bias is drawn.

In our society there are many intermediaries between those who make extraordinary claims and decision makers who might react to such claims. For example, if a government employee makes an extraordinary discovery, news of that discovery will be filtered through many levels of middle management, each of whom reserves the right to interpret and "spin" that news to adapt to context.

Consider an academic example. Nature might reveal extraordinary evidence to a research assistant, who then describes it to his project leader. The leader submits a paper describing these results to journals, where referees report their evaluation to editors, who then decide whether to publish. A journal or laboratory publicist may then suggest the publication to news reporters, who submit articles to news editors, who choose what readers see.

We will show that distortions introduced by these many intermediaries accumulate and increase effective signal variance exponentially. While the extraordinary claim of a single person could count for a great deal of extraordinary evidence, a chain of a half dozen of the same sort of people can effectively eliminate that evidence.

We now define the model, and then describe its solution and illustrate with an example.

Model

Consider a one dimensional spectrum of possibilities x, such as how many people will be killed in a looming disaster. Let us assume that x is distributed according to a power law[1] $P(x > \hat{x}) \propto \hat{x}^{-\beta}$, so that large values of x are "extraordinary." That is, large values are especially unlikely, but also especially important.

Let us transform the description of our parameter of interest from x to $y = \log(x)$, which is then distributed with density $\pi(y) \propto e^{-\beta y}$. We will assume that nature gives someone a noisy signal s that is normally distributed in y, and so log-normally distributed in x. That is, an agent has access to a signal $s = y + \epsilon$, where $\epsilon \sim \mathrm{N}(0, \sigma^2)$.

A perfectly rational, honest, and clear agent would, up seeing signal s, simply report to others either his signal s or his resulting posterior beliefs $P(y|s)$ calculated according to Bayes' rule $P(y|s) \propto P(s|y)\pi(y)$. A real human's report, however, may be distorted in two ways. First, a real human may have incentives that do not always and exactly favor honesty. Second, humans do not always take the action that is exactly best according their incentives.

To model these two distortions, let us first assume that observers know the exact value of σ^2, but not the exact signal value s. So the agent can choose a value of s' to report instead of nature's actual signal s. Second, assume observers will eventually know the exact parameter value y, call it y^*, and will reward the agent for having reported a signal s' that assigns a high probability, $P(y^*|s')$, to the actual[2] observed value y^*.

Specifically, the agent's utility from reporting s' is

$$U(s', b) = b\, s' + \ln(P(y^*|s')).$$

So the agent's payoff is the sum of a *logarithmic scoring rule* term $\log(P(y^*|s'))$, which by itself would induce an honest report $s' = s$ (Good, 1952; Winkler, 1969), and a distorting bias term $b\, s'$, which reduces honesty. The value of bias parameter b is known only to the agent; observers know only that bias b is drawn from a normal prior distribution $\lambda(b)$ given by $b = \bar{b} + \eta$, with $\eta \sim \mathrm{N}(0, \theta^2)$.

Third, we assume the agent's actions cannot be predicted exactly from knowing signal s and his bias b. Instead, he has a higher chance of taking actions that give him a higher expected payoff, as in

$$P(s'|s, b) \propto \exp(\mathrm{E}[U(s', b)|s, b]/r),$$

where r describes the agent's degree of irrationality. This is the "quantal response" behavioral assumption now popular in game theory (McKelvey & Palfrey, 1995). As $r \to 0$, the

[1]The distribution over x must cut off at some lower limit $\underline{x}$, but we assume this cutoff is well below the "extraordinary" values we are considering.

[2]Since the probability of any exact value y is zero, imagine the range of possible y is broken into a finite number of very small ranges of width dy.

agent is almost certain to take the exactly optimal action, but with larger r the more the agent will deviate more from this optimum.

Just as the agent can use Bayes' rule to turn a signal distribution $P(s|y)$ into a posterior $P(y|s)$, observers who hear the agent's report s' can treat it as a signal of nature's signal s, and hence of the parameter y, by computing

$$P(s'|s) = \int P(s'|s,b)\lambda(b)\,db,$$

$$P(s'|y) = \int P(s'|s)P(s|y)\,ds.$$

Observers can then turn this signal distribution $P(s'|y)$ into a posterior $P(y|s')$.

Now consider a chain of N agents, each reporting to the next. Nature chooses a signal s_0 about y, and then each agent i observes a signal report s_{i-1} chosen by the previous agent. Each agent can have differing values for $\bar{b}_i, \theta_i, r_i$, as long as these values are commonly known. An observer of any signal s_n can compute a posterior via

$$P(y|s_n) \propto \pi(y) \int P(s_0|y) \prod_{i=1}^{n} P(s_i|s_{i-1}) \prod_{i=0}^{n-1} ds_i.$$

If all agents were perfectly honest and rational, with all $\theta_i = r_i = 0$, then we would have $s_N = s_0$ and extraordinary claims s_N at the end of the chain could be valid extraordinary evidence of an extraordinary signal s_0. But how much bias and irrationality, across how long a chain, does it take until extraordinary end claims s_N say little about whether s_0 is extraordinary?

Solution

Using Bayes' rule, $P(y|s) \propto P(s|y)\pi(y)$, and completing the square gives us $y \sim P(y|s) = \mathrm{N}(z, \sigma^2)$, where the median and mean y value is

$$z = s - \beta\sigma^2.$$

The median x is $\exp(z)$, while the mean x is $\exp(s + \sigma^2(0.5 - \beta))$.

The correction $\beta\sigma^2$ in the mean z accounts for the fact that high values of y are unlikely, making high signals s likely to result instead from unusually positive errors ϵ. The fact that it is σ^2 in this correction, and not σ, implies that there are two very different regimes. When $\beta\sigma \ll 1$, the correction is only a small fraction of σ, and matters little. But when $\beta\sigma \gg 1$, the correction is much larger than σ, and negates much of the impact of an apparently "extraordinary" signal s.

Since $P(y|s)$ is a normal distribution, its logarithm is quadratic, making utility $U(s', b)$ a quadratic function of s'. Given a normal distribution, the expectation of a quadratic function is also quadratic, and so up to a constant we have

$$\mathrm{E}[U(s',b)|s,b] = -\frac{(s'-\hat{s})^2}{2\sigma^2},$$

where $\hat{s} \equiv s + b\sigma^2$. The quantal response assumption $P(s'|s,b) \propto \exp(\mathrm{E}[U(s',b)|s]/r)$ exponentiates this quadratic, so that behavior is normally distributed as $s' = \hat{s} + \varepsilon$ where $\varepsilon \sim \mathrm{N}(0, r\sigma^2)$. From the point of view of an observer, we can substitute and get

$$s' = y + \epsilon + (\bar{b} + \eta)\sigma^2 + \varepsilon,$$

giving $\bar{s}' \equiv \mathrm{E}[s'] = y + \bar{b}\sigma^2$, and $\mathrm{E}[(s'-\bar{s}')^2] = \sigma^2(1 + r + \theta^2)$. Note that mean bias $\bar{b}$ can be completely corrected for, and does not influence the variance of s'.

For N chained agents, report s_n is distributed as $P(s_n|y) = \mathrm{N}(y + \bar{b}_n\sigma^2_{n-1}, \sigma^2_n)$, where

$$\sigma^2_n = \sigma^2_0 \prod_{i=1}^{n}(1 + r_i + \theta^2_i)$$

and σ_0 is the noise in nature's signal s_0 of y. Our posterior on y after report s_N should be

$$P(y|s_N) = \mathrm{N}(s_N - (\beta + \bar{b}_N)\sigma^2_N, \sigma^2_N).$$

Notice that the total error variance here goes as a *product* of independent error variances. In many systems total variance goes as only the *sum* of the variances of independence sources.

Example

To illustrate, imagine a type of disaster where the number of deaths x was distributed as $P(x > \hat{x}) \propto \hat{x}^{-1}$ (Thus[3] $\beta = 1$.) Large disasters of this type would cause as much damage on average as small ones, and so be important to consider.

Imagine that in truth an upcoming disaster would, if not prevented, be capable of killing $x = 6\mathrm{E}9$ ($\equiv 6 \times 10^9$) humans, i.e., all six billion of us. And imagine that John receives a (log-normally distributed) signal from nature about this upcoming disaster, a signal John knows has an error of about one order of magnitude. That is, 68% of the time nature's signal s will fall in the one standard deviation range (SR) of $[0.1x, 10x] = [6\mathrm{E}8, 6\mathrm{E}10]$. (So $\sigma = \ln(10)$.) Let us continue to track a SR of outcomes, within which reality lies 68% of the time.

After applying Bayes rule, the SR for the best median estimate of deaths based on nature's signal is [3E6, 3E8]. Based on this, we expect John to underestimate the harm. Nevertheless, John clearly has extraordinary evidence, which would justify an extraordinary claim; this disaster is well worth investigating, to see if prevention is possible.

[3]Many kinds of disasters do seem to follow such a power law. For tornados and terrorist attacks $\beta = 1.4$, while for earthquakes and wars $\beta = 0.41$ (Hanson, 2007a).

John will report to Mary. John will in effect forecast a probability distribution over deaths, and be rewarded for assigning a high probability to the actual number of deaths.[4] John, however, also has an incentive to exaggerate or downplay this disaster. Mary does not know John's exact bias incentive. But Mary does knows that 68% of the time John has more than three times as much to gain by raising the probability he assigns to what actually happens by 1%, as he might gain from raising or lowering his stated median estimate of x by 1%. (So $\theta = 1/3$.)

John has incentives that reward him, and he tries to choose the report that gives him the greatest expected reward. But John does not alway choose well. If one report would assign 1% more probability to what actually happens than another report, then John is 9% more likely to choose the first (better) report. (So $r = 1/9$.) These two sources of error, a bias incentive and choice errors, each add about 5% to the noise in John's report; eliminating either source while doubling the other gives about the same effect.

The net result is that John's report has 10% more noise than the signal nature gave him. While nature's signal to John had a SR of $[0.1x, 10x]$, the signal that is John's report to Mary has a SR of $[0.079x, 12.7x]$. And while the SR of the median deaths estimate based on nature's signal is [3E6, 3E8], the same SR given John's report is [7.3E5, 1.2E8]. This is somewhat muted, but still reasonably extraordinary evidence, and pretty big news.

Imagine that Mary cannot take relevant action directly. Instead, imagine a reporting chain of John to Mary to Fred to Lisa to Bill to Pam to Joe, where only Joe can make a relevant decision. So between Joe and nature is a chain of seven people, each reporting to the next. Imagine further that each is exactly as noisy as John, and that everyone knows all the noise levels.

Given these assumptions, Pam's report is about twice as noisy as nature's signal. While nature's signal to John had a SR of $[0.1x, 10x]$, the signal that is Pam's report to Joe has a SR of $[0.0099x, 101x]$. And while the SR of the median deaths estimate based on nature's signal was [3E6, 3E8], the same SR given Pam's report is the much less extraordinary [0.032, 330].

Even Pam's report should seem noteworthy, however, as the SR of *mean* deaths estimate is [1.5E4, 6.3E7]. But if there were fourteen people in the reporting chain, or if each person were twice as noisy, even the SR of mean deaths estimate would be the quite negligible [1E-13, 1.5E-5]; even a two standard deviation positive signal would only give a mean estimate of 0.16 deaths. The initially extraordinary evidence would have been completely washed out in the noise of human error in the reporting chain.

Conclusion

Extraordinary claims do require extraordinary evidence, but on uninteresting topics we can usually count on people to reliably communicate their exposure to a priori unlikely evidence. On interesting topics, however, people become less reliable; they often have interests in exaggerating or downplaying the implications of their evidence, and their actions often deviate

[4] Like most disaster experts, John in effect "bets on death."

from their interests.

We have considered a simple model of noisy humans reporting extraordinary evidence. When a person's noise is low, his extraordinary claim *is* extraordinary evidence, but when his noise is high, extraordinary claims are more likely to be due to unusual noise than unusual truth; the extraordinary evidence is washed out. Relatively low noise people who are organized into a reporting chain are equivalent to a single high noise person; such reporting chains are simply not capable of communicating extraordinary evidence.

This model has only considered two sources of noise in communication. Other possible complications include uncertainty about error variances, thicker than Gaussian error tails, a lack of available incentives tied to actual outcomes, lower bounds to parameter distributions, and upper bounds to interesting values. Models that included these noise sources would presumably result in even noisier communication, and thus a faster washing out of extraordinary evidence.

On the other hand, perhaps we need only the first few levels of reporting need discretion to adapt a claim to context; reports at further levels could be something like "Fred said that our best estimate of disaster deaths is one million." Another alternative might be to use prediction markets to shorten reporting chains; the person who saw nature's signal trades in the market, anyone who wants to correct that market price for context does so via trades, and everyone else is referred to the market price (Hanson, 2007b).

References

Good, I. J. (1952). Rational Decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *14*(1), 107–114.

Hanson, R. (2007a). Catastrophe, Social Collapse, and Human Extinction. In Bostrom, N., & Cirkovic, M. (Eds.), *Global Catastrophic Risks*. Oxford University Press.

Hanson, R. (2007b). Insider Trading and Prediction Markets. *Journal of Law, Economics, and Policy*. to appear.

McKelvey, R. D., & Palfrey, T. (1995). Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior*, *10*, 6–38.

Winkler, R. L. (1969). Scoring Rules and the Evaluation of Probability Assessors. *Journal of the American Statistical Association*, *64*(327), 1073–1078.