

Supplemental Material A

For “Talking the talk, not walking the walk: The co-evolution of overconfidence and loss aversion”

A1. Proof of Proposition 1a

Proposition 1a: *With a continuous, bounded distribution of types, there exists a unique separating equilibrium, which takes the form of a smooth profile of self-belief in which all types but the very worst are overconfident, if the net gain from increasing self-belief—that is, the benefit of being perceived as more capable minus the associated decision cost—is greater for the more able than for the less able (the single-crossing condition).*

Proof:

The proof proceeds in three steps: verifying standard existence conditions for the interior of the type space, formalizing the boundary condition and off-path beliefs, and proving strict overconfidence for all types above the lower bound.

Step 1: Standard Existence Conditions for Interior Types

The sender’s objective is to maximize $W(a, b) = u(a, x(b)) - C(a, b)$, where $x(b)$ is the receiver’s equilibrium action schedule. Following Mailath (1987), a unique, fully separating, and strictly monotonic signaling schedule $b(a)$ exists if the payoff function satisfies standard regularity conditions and the single-crossing property. Our model satisfies these requirements: $u_x > 0$, $u_{xx} \leq 0$, and $C_{bb} > 0$. Crucially, the willingness to trade off signal costs for better receiver action, defined as $Z \equiv -C_b/u_x$, is strictly increasing in type a . This single-crossing condition ($Z_a > 0$) ensures that higher types have a comparative advantage in signaling, which is sufficient to yield a strictly monotonic separating schedule for all interior types.

Step 2: The Boundary Condition and Off-Path Beliefs

Mailath’s (1987) existence proof requires the marginal cost of signaling to be strictly non-zero ($C_b \neq 0$) everywhere. However, in our model, the lowest-ability type ($a = \underline{a}$) has no lower types to separate from and thus minimizes costs by choosing truthful beliefs, $b(\underline{a}) = \underline{a}$. By definition, the marginal cost of inflating beliefs at the truth is zero, meaning $C_b(\underline{a}, \underline{a}) = 0$. This does not violate the existence of a separating equilibrium, but rather defines its initial value boundary condition. As demonstrated by Riley (1979, pp. 340-341), when the marginal signaling cost evaluates to zero at the lower bound, the equilibrium separating schedule must exhibit a vertical tangent at that bound. Specifically, evaluating the First-Order Condition at the lower bound yields:

$$u_x(\underline{a}, x(b))x'(b) = C_b(\underline{a}, \underline{a}) = 0$$

Because $u_x > 0$, it must be that $x'(b) = 0$ precisely at $a = \underline{a}$. Since the sender’s separating schedule $b(a)$ is the inverse of the receiver’s inference mapping, $x'(b) = 0$ implies that $db/da \rightarrow \infty$ as $a \rightarrow \underline{a}$. This initial vertical launch

imposes sufficient decision error costs to successfully deter the lowest type from mimicking marginally higher types.

Finally, for any off-path signal $b < \underline{a}$, the D1 criterion requires the receiver to assign beliefs to the type most eager to deviate. Because the single-crossing condition implies higher types find downward deviations strictly more costly, the belief is assigned to the lowest type, $p(b) = \underline{a}$. This yields the lowest possible receiver action while strictly increasing decision costs, ensuring no type has a profitable deviation below the separating schedule.¹

Step 3: Proof of Strict Overconfidence ($b > a$)

Having established a strictly monotonic separating schedule $b(a)$ initialized at $b(\underline{a}) = \underline{a}$, we now prove that $b > a$ for all $a > \underline{a}$. In a fully separating equilibrium, the receiver's action must strictly improve as the signal increases, meaning $x'(b) > 0$. For any interior type $a > \underline{a}$, the sender's First-Order Condition is:

$$C_b(a, b) = u_x(a, x(b))x'(b)$$

Because the marginal benefit of receiver action is strictly positive ($u_x > 0$) and the schedule is strictly monotonic ($x'(b) > 0$), the right side of the equation is strictly positive. Therefore, it must be that:

$$C_b(a, b) > 0$$

By assumption, the cost of decision error $C(a, b)$ is strictly convex in b ($C_{bb} > 0$) with a unique global minimum at realistic beliefs, meaning $C_b(a, a) = 0$. The only way for the marginal cost $C_b(a, b)$ to be strictly positive is if the chosen belief strictly exceeds true ability. Therefore, $b(a) > a$ for all $a > \underline{a}$, confirming that all types except the very worst must exhibit strict equilibrium overconfidence.

¹While the D1 criterion provides the standard game-theoretic mechanism to restrict off-path beliefs to the lowest type $\underline{a}$, it relies on strong epistemic assumptions regarding forward induction and intentionality (Banks and Sobel, 1987). An arguably more natural justification for this boundary condition arises from evolutionary game theory (e.g., Grafen, 1990). In an evolutionary signaling context, receivers do not engage in complex out-of-equilibrium inference; rather, they possess evolved stimulus-response thresholds calibrated to the naturally occurring population. Consequently, a signal falling below the equilibrium minimum, $b < \underline{a}$, fails to trigger a heightened response, defaulting automatically to the baseline treatment $x(\underline{a})$. This biological thresholding delivers the boundary penalty required to sustain the separating equilibrium without demanding hyper-rationality from the players.

A2. Derivation of the specific cost function

If the project is accepted, the probability of success in state s is

$$p = sa$$

The expected payoff in state s

$$pH - (1 - p)L = sa(H + L) - L$$

so the optimal cutoff state, S , has expected payoff zero,

$$S = \frac{L}{a(H + L)}$$

If all states above S are accepted, expected payoff from optimal acceptance is

$$\begin{aligned} P &= \int_S^1 (sa(H + L) - L) ds = 0.5(1 - S)(Ha - 2L + La + Hs a + Ls a) = \\ &= \frac{0.5}{a(H + L)} (Ha - L + La)(Ha - 2L + La + L) \end{aligned}$$

With biases, the cutoff state is

$$S' = \frac{\lambda L}{b(H + \lambda L)}$$

making the expected payoff

$$P' = \frac{0.5}{b(H + \lambda L)} (Hb - L\lambda + Lb\lambda) \left(Ha - 2L + La + \frac{a\lambda L}{b(H + \lambda L)} (H + L) \right)$$

The cost of bias with arbitrary λ is

$$c = P - P'$$

Because the introduction of loss aversion alters the cost of an agent's expected decision error, when λ is present we designate the augmented cost function as $c(a, b, \lambda)$ rather than $C(a, b)$.

$$\frac{\partial c}{\partial b} = c_b = \frac{\lambda L^2}{b^3} \frac{b(H + \lambda L) - \lambda a(H + L)}{(H + \lambda L)^2} > 0. \text{ If } \lambda = 1, c_b = \frac{L^2}{b^3} \frac{(b - a)}{H + L}$$

The loss aversion that fully offsets b is

$$\lambda^* = \frac{bH}{a(H + L) - bL}$$

$$c_{bb} = -\frac{L^2}{b^4} \frac{\lambda(2b(H + \lambda L) - 3a\lambda(H + L))}{(H + \lambda L)^2} > 0 \text{ when } 2b(H + \lambda L) < 3a\lambda(H + L)$$

$$c_{ba} = -\frac{L^2}{b^3} \frac{\lambda^2}{(H + \lambda L)^2} (H + L) < 0$$

$$c_{\lambda} = -H \frac{L^2(b(H + \lambda L) - \lambda a(H + L))}{b^2(H + \lambda L)^3} < 0 \text{ when } b > \frac{\lambda a(H + L)}{H + \lambda L} \text{ which holds for } \lambda = 1$$

when $b > a$.

For Section A6, *Visibility of loss aversion*, note that

$$\frac{c_{\lambda}}{c_b} = -\frac{Hb}{\lambda(H + \lambda L)}$$

so this ratio is independent of ability, a .

A3. Derivation of the plot of belief against ability

A3.1 Uncapped Case

We wish to compute and plot the separating equilibrium signaling schedule $b(a)$ for the specific parametric case outlined in the main text. The parameters are $H = L = 2$ and $u(a, x) = x$, with ability uniformly distributed on the interval $[0.1, 0.8]$.

In a fully separating equilibrium, receivers correctly infer the sender's true ability, meaning the receiver's action is

$$x(b) = a(b).$$

For interior types, the sender's signaling schedule is defined by the First-Order Condition

$$C_b(a, b) = u_x(a, x(b)) x'(b).$$

Given $u(a, x) = x$, the marginal benefit of receiver action is

$$u_x = 1.$$

Substituting $x'(b) = a'(b)$, the condition simplifies to

$$a'(b) = C_b(a, b).$$

From our specific cost function with $\lambda = 1$, $H = 2$, and $L = 2$, the marginal cost of the signal is

$$C_b = \frac{b - a}{b^3}.$$

Therefore, the separating equilibrium is characterized by the following first-order ordinary differential equation (ODE) representing the inverse signaling schedule:

$$a'(b) = \frac{b - a}{b^3}.$$

A3.1.1 Boundary Condition and Numerical Method

To generate the equilibrium schedule, we treat this ODE as an Initial Value Problem (IVP). Because the lowest-ability type does not need to separate from anyone lower, they choose accurate beliefs, providing the initial boundary condition

$$a(0.1) = 0.1.$$

Evaluating the ODE at this boundary yields

$$a'(0.1) = 0.$$

Because $a'(b)$ is the inverse of the sender's strategy $b'(a)$, this corresponds to a vertical tangent ($db/da \rightarrow \infty$) at the lower bound, cleanly initializing the separating schedule.

Rather than finding roots type-by-type, the plot in Figure 1 is generated by numerically integrating this differential equation forward. Using a standard numerical ODE solver (such as a Runge-Kutta method), we integrate

$$a'(b) = \frac{b - a}{b^3}$$

starting from $b = 0.1$ and $a = 0.1$.

We continue the numerical integration over increasing values of b until the inferred ability $a(b)$ reaches the upper bound of the type space,

$$a = 0.8.$$

A3.1.2 Plotting

Once the integration is complete, the resulting trajectory is plotted with the true ability $a(b)$ on the horizontal axis and the chosen signal b on the vertical axis. The resulting curve $b(a)$ is strictly increasing and satisfies $b > a$ for all interior types, yielding the smooth overconfidence schedule depicted in Figure 1.

A3.2 Capped Model

Types are $a \in [\underline{a}, \bar{a}]$. The sender chooses a signal $b \geq 0$. The receiver chooses an action $x(b)$ increasing in b .

A3.2.1 Payoffs

Sender payoff is

$$U_S(a, b) = V(a, x(b)) - C(b, a),$$

where $C(b, a)$ is minimized at $b = a$.

A3.2.2 Capped Receiver Valuation

Receiver action has a cap in sender valuation. We represent this as a piecewise function:

$$V(a, x) = \begin{cases} x, & \text{for } x < x^* \\ x^*, & \text{for } x \geq x^* \end{cases}$$

Thus the marginal value of receiver action is

$$V_x = \begin{cases} 1, & \text{for } x < x^* \\ 0, & \text{for } x \geq x^* \end{cases}$$

A3.2.3 Uncapped Benchmark

Without the cap, the Spence equilibrium is fully separating:

$$b_{\text{old}}(a) \text{ strictly increasing and satisfying } b_{\text{old}}(a) > a.$$

Define the threshold type a_1 by

$$x(b_{\text{old}}(a_1)) = x^*,$$

and let

$$b_{\text{cap}} = b_{\text{old}}(a_1).$$

A3.2.4 Equilibrium with the Cap

When x^* binds, the sender's marginal benefit from increasing b becomes zero. The sender then solves:

$$\min_b C(b, a) \quad \text{subject to } x(b) \geq x^*.$$

Since $C(b, a)$ is minimized at $b = a$:

- If $a < b_{\text{cap}}$, then $b = a$ yields $x(b) < x^*$ and is not feasible. The cheapest feasible signal is $b = b_{\text{cap}}$.
- If $a \geq b_{\text{cap}}$, then $b = a$ satisfies $b = a \geq b_{\text{cap}}$ and thus induces $x(b) \geq x^*$. Hence $b = a$ is the cost-minimizing feasible signal.

Thus the equilibrium signalling strategy is

$$b^*(a) = \begin{cases} b(a), & a \leq a_1, \\ b_{\text{cap}}, & a_1 < a < b_{\text{cap}}, \\ a, & a \geq b_{\text{cap}}. \end{cases}$$

A3.2.5 Interpretation

1. **Low types:** behave exactly as in the uncapped separating equilibrium.
2. **Intermediate types:** pool at b_{cap} , the minimum signal inducing x^* .
3. **High types:** revert to truth-telling $b = a$, because it is both feasible and cost-minimizing.

The signal schedule therefore increases for low types, is flat for intermediate types, and increases again along the 45° line for high types.

A4. Proof of Proposition 1b

Proposition 1b: *Assume a discrete state space of true abilities $a_1 < a_2 < \dots < a_N$. Consider a separating Perfect Bayesian Equilibrium (PBE) supported by out-of-equilibrium beliefs that satisfies the Intuitive Criterion (Cho & Kreps, 1987), uniquely isolating the least-cost separating outcome. Let b_i denote the equilibrium self-belief of type a_i , and let $p(b)$ denote the receiver's strictly increasing perceived ability schedule, such that $p(b_i) = a_i$. Type a_i exhibits strict equilibrium overconfidence ($b_i > a_i$) if and only if:*

$$u(a_{i-1}, a_i) - u(a_{i-1}, a_{i-1}) > C(a_{i-1}, a_i) - C(a_{i-1}, b_{i-1}^*)$$

Proof:

By the Intuitive Criterion, the unique surviving separating equilibrium is the least-cost separating (Riley, 1979) outcome. In this equilibrium, each type a_i (for $i > 1$) chooses the cost-minimizing signal $b_i \geq a_i$ that is just sufficient to deter the immediately lower type, a_{i-1} , from mimicking them. Therefore, the equilibrium signal b_i is determined by the binding local downward Incentive Compatibility (IC) constraint:

$$u(a_{i-1}, a_i) - C(a_{i-1}, b_i) \leq W^*(a_{i-1})$$

We prove the equivalence in two steps.

Step 1: Sufficiency

Assume the condition holds: $u(a_{i-1}, a_i) - C(a_{i-1}, a_i) > W^*(a_{i-1})$.

This inequality implies that if type a_i were to choose their cost-minimizing truthful signal ($b_i = a_i$), type a_{i-1} would achieve a strictly higher payoff by mimicking type a_i than by playing their own equilibrium strategy. Thus, $b_i = a_i$ violates the downward IC constraint.

To restore incentive compatibility, type a_i must inflate their signal to increase the decision error cost for the mimicking type. Because $C(a, b)$ is strictly convex in b and minimized at $b = a$, the marginal cost of signaling $C_b(a_{i-1}, b)$ is strictly positive for all $b \geq a_i > a_{i-1}$. Therefore, to satisfy the IC constraint $C(a_{i-1}, b_i) \geq u(a_{i-1}, a_i) - W^*(a_{i-1})$, the signal b_i must be strictly increased beyond a_i . Hence, $b_i > a_i$.

Step 2: Necessity

Assume type a_i exhibits strict equilibrium overconfidence, meaning $b_i > a_i$.

In the least-cost separating equilibrium, type a_i strictly prefers to minimize their signaling cost $C(a_i, b_i)$. Because $C(a_i, b)$ is uniquely minimized at $b = a_i$, type a_i will only choose a signal $b_i > a_i$ if forced to do so by a binding IC constraint. If the proposed condition did not hold—that is, if $u(a_{i-1}, a_i) - C(a_{i-1}, a_i) \leq W^*(a_{i-1})$ —then type a_{i-1} would not profitably mimic type a_i even if type a_i simply told the truth. In that case, the least-cost separating signal for type a_i would exactly equal their true ability, $b_i = a_i$, contradicting our assumption that $b_i > a_i$. Therefore, for $b_i > a_i$ to emerge in the least-cost separating outcome,

it must be that telling the truth strictly violates the downward IC constraint, which requires exactly that $u(a_{i-1}, a_i) - C(a_{i-1}, a_i) > W^*(a_{i-1})$.

A5. Proof λ cannot fully offset $b > 1$ (Lemma 1)

Continuous distribution

Consider an agent in state $i \in \{1, \dots, n\}$ presented with a project. There is a continuous set of potential projects indexed by underlying quality $\theta \in \Theta_i$. A project of quality θ yields a random monetary payoff $x \in \mathbb{R}$, drawn from the true cumulative distribution function $F_i(x | \theta)$. The agent is overconfident and instead perceives the payoff distribution to be $G_i(x | \theta)$, where G_i first-order stochastically dominates F_i . The agent applies loss aversion λ to negative payoffs.

The perceived expected utility of agent i from project θ is

$$U_i(\theta, \lambda) = \int_0^\infty x dG_i(x | \theta) + \lambda \int_{-\infty}^0 x dG_i(x | \theta).$$

The agent accepts the project if $U_i(\theta, \lambda) \geq 0$.

Assume that, for each state i , $U_i(\theta, \lambda)$ is continuous in θ , strictly increasing in θ , and crosses zero exactly once. Then there exists a unique acceptance cutoff $\hat{\theta}_i(\lambda)$ such that

$$U_i(\hat{\theta}_i(\lambda), \lambda) = 0,$$

and the agent accepts precisely those projects with $\theta \geq \hat{\theta}_i(\lambda)$.

A risk-neutral agent with accurate beliefs evaluates a project of quality θ according to

$$w_i(\theta) = \int_{-\infty}^\infty x dF_i(x | \theta).$$

Assume $w_i(\theta)$ is strictly increasing in θ and crosses zero exactly once. Then the unique first-best cutoff θ_i^{FB} is defined by

$$w_i(\theta_i^{FB}) = 0.$$

Under accurate beliefs, the first-best rule is to accept all projects with $\theta \geq \theta_i^{FB}$.

For each state i , define welfare under a generic acceptance threshold t by

$$V_i(t) = \int_t^\infty w_i(\theta) P_i(\theta) d\theta,$$

where $P_i(\theta)$ is the density of available projects in state i . Because $w_i(\theta)$ is strictly increasing and satisfies $w_i(\theta_i^{FB}) = 0$, it follows that $V_i(t)$ is uniquely maximized at $t = \theta_i^{FB}$. To see this, note that

$$V_i'(t) = -w_i(t) P_i(t).$$

Hence $V_i'(t) > 0$ for $t < \theta_i^{FB}$, $V_i'(t) < 0$ for $t > \theta_i^{FB}$, and therefore $V_i(t)$ attains its unique maximum at $t = \theta_i^{FB}$.

If beliefs were accurate and $\lambda = 1$, then $G_i = F_i$ implies $\hat{\theta}_i(1) = \theta_i^{FB}$, so the agent would attain first-best welfare in every state.

Now suppose the agent is overconfident, so G_i first-order stochastically dominates F_i . Then, absent correction, the agent overvalues projects and typically sets too low a cutoff. To restore first-best behavior in a given state i , one would need a state-specific loss-aversion parameter λ_i^* satisfying

$$\widehat{\theta}_i(\lambda_i^*) = \theta_i^{FB}.$$

Substituting θ_i^{FB} into the indifference condition gives

$$\lambda_i^* = \frac{\int_0^\infty x dG_i(x \mid \theta_i^{FB})}{-\int_{-\infty}^0 x dG_i(x \mid \theta_i^{FB})}.$$

Assume there exist at least two states j and k such that

$$\lambda_j^* \neq \lambda_k^*.$$

That is, the loss-aversion parameter required to implement the first best is heterogeneous across states.

Suppose a single uniform parameter λ must be used in all states. Aggregate welfare is then

$$W(\lambda) = \sum_{i=1}^n V_i(\widehat{\theta}_i(\lambda)) = \sum_{i=1}^n \int_{\widehat{\theta}_i(\lambda)}^\infty w_i(\theta) P_i(\theta) d\theta.$$

If some uniform λ implemented the first best in every state, it would have to satisfy

$$\widehat{\theta}_i(\lambda) = \theta_i^{FB} \quad \text{for all } i = 1, \dots, n.$$

But by definition of λ_i^* , this requires

$$\lambda = \lambda_i^* \quad \text{for all } i = 1, \dots, n.$$

Hence a uniform λ can implement the first best in every state only if

$$\lambda_1^* = \lambda_2^* = \dots = \lambda_n^*.$$

This contradicts the assumption that $\lambda_j^* \neq \lambda_k^*$ for some j, k . Therefore, for every uniform λ , there exists at least one state m such that

$$\widehat{\theta}_m(\lambda) \neq \theta_m^{FB}.$$

Since $V_m(c)$ is uniquely maximized at $t = \theta_m^{FB}$, it follows that

$$V_m(\widehat{\theta}_m(\lambda)) < V_m(\theta_m^{FB}).$$

For every other state i , welfare cannot exceed its first-best value, so

$$V_i(\widehat{\theta}_i(\lambda)) \leq V_i(\theta_i^{FB}).$$

Summing across states gives

$$W(\lambda) = \sum_{i=1}^n V_i(\hat{\theta}_i(\lambda)) < \sum_{i=1}^n V_i(\theta_i^{FB}) =: W^{FB}.$$

Thus the best achievable outcome under a uniform loss-aversion parameter is strictly below the first-best outcome achievable under accurate beliefs. Q.E.D.

Discrete distribution

Now suppose the set of feasible project qualities is discrete. Let θ_i^{FB} again denote the first-best cutoff in state i , and let λ_i^* be the state-specific parameter satisfying

$$\hat{\theta}_i(\lambda_i^*) = \theta_i^{FB}.$$

If a uniform parameter λ is imposed and the λ_i^* are not all equal, then there exists at least one state m such that

$$\hat{\theta}_m(\lambda) \neq \theta_m^{FB}.$$

In a discrete environment, a distorted cutoff need not strictly lower welfare unless it changes the acceptance decision for at least one project with positive probability. Without further conditions, a uniform λ could theoretically implement the first-best in all states if the distorted cutoffs all happen to fall into 'empty' intervals between discrete project qualities across all states, yielding a weak rather than strict welfare loss. To rule out this mathematically possible but economically knife-edge scenario, we impose a standard genericity condition: we assume that in every state where $\hat{\theta}_i(\lambda) \neq \theta_i^{FB}$ there is positive probability mass on at least one project quality lying between the distorted cutoff and the first-best cutoff.

Then, in state m , the distorted cutoff causes at least one project with positive probability to be treated differently from the first-best rule. If $\hat{\theta}_m(\lambda) < \theta_m^{FB}$, the agent accepts some project with negative true expected value. If $\hat{\theta}_m(\lambda) > \theta_m^{FB}$, the agent rejects some project with positive true expected value. Either way, state- m welfare is strictly below first best. Therefore aggregate welfare under a uniform parameter is strictly below the first-best level. Q.E.D.

A6. Visibility of loss aversion

The analysis for λ visible without choice is for two types, α, β . If a β is to emulate an α it must match both its belief and its loss aversion. Now, for given b , an α no longer sets $C_\lambda = 0$. Minimum cost separation requires

$$u(a_\beta, a_\beta) - c(a_\beta, a_\beta, 1) = u(a_\beta, a_\alpha) - c(a_\beta, b_\alpha, \lambda_\alpha) \quad (1)$$

From (1), because β must match the choice of λ by α ,

$$\frac{db_\alpha}{d\lambda_\alpha} = \frac{-c_\lambda(a_\beta, b_\alpha, \lambda_\alpha)}{c_b(a_\beta, b_\alpha, \lambda_\alpha)}$$

Evaluated at the hidden λ equilibrium values, this derivative is positive as the cost minimizing λ is decreasing in a . The FOC for an interior maximum for λ_α is

$$c_\lambda(a_\alpha, b_\alpha, \lambda) + c_b(a_\alpha, b_\alpha, \lambda) \frac{db_\alpha}{d\lambda} = 0 \quad (2)$$

The first term of (2) is zero at the cost minimising λ for α as applies in the hidden equilibrium. The second term is positive at the hidden value equilibrium, thus the visible equilibrium cannot be the same as the hidden equilibrium. At an internal equilibrium the FOC is

$$\frac{c_\lambda(a_\beta, b_\alpha, \lambda_\alpha)}{c_b(a_\beta, b_\alpha, \lambda_\alpha)} = \frac{c_\lambda(a_\alpha, b_\alpha, \lambda_\alpha)}{c_b(a_\alpha, b_\alpha, \lambda_\alpha)} \quad (3)$$

As previously demonstrated, this ratio is independent of a . The implication is that given an arbitrary λ_α , with b_α then solved from (2), (3) holds, so all mixes of b and λ that solve (2) are equivalent equilibria. There is therefore no need for loss aversion in the visible case.

Finally, consider the case where the sender can voluntarily and verifiably disclose their loss aversion. As a benchmark, the "hidden" equilibrium analyzed in the main text assumes λ is endogenously chosen by the sender to minimize decision costs, but remains strictly unobservable to the receiver. Now suppose a verifiable disclosure technology exists: if type α voluntarily discloses their chosen λ_α , any type β attempting to emulate them must actually adopt and disclose λ_α . Can type α achieve a cheaper separating equilibrium (a lower b) by verifiably disclosing a λ that strategically deviates from its hidden cost-minimizing level? To do so, α would need to select a λ that increases the mimicking cost for β more than it increases the direct decision cost for α . However, because β is already further from its cost-minimizing preferences than α , shifting λ away from the optimal interior level imposes a higher marginal cost penalty on α than on β . Consequently, α cannot use the verifiable disclosure of a distorted λ to profitably lower the required signaling belief b . In equilibrium, α simply chooses the same cost-minimizing λ as in the unobservable case, and β 's indifference condition remains unchanged. Therefore, the voluntary, verifiable disclosure of loss aversion does not alter the separating equilibrium.

A7. Non existence of a "biased" compensatory equilibrium.

The issue is whether in the absence of a reason to signal type, there can be an evolutionary stable equilibrium with $\lambda > 1, a > b$. That is, at such a configuration is there a local minimum of the cost function when there are two decision problems. The FOCs are

$$C_b = \frac{\lambda L^2 [b(H + \lambda L) - \lambda a(H + L)]}{b^3 (H + \lambda L)^2} + \frac{\lambda l^2 [b(h + \lambda l) - \lambda a(h + l)]}{b^3 (h + \lambda l)^2} = 0$$

$$C_\lambda = H \frac{L^2 [b(H + \lambda L) - \lambda a(H + L)]}{b^3 (H + \lambda L)^3} + h \frac{l^2 [b(h + \lambda l) - \lambda a(h + l)]}{b^3 (h + \lambda l)^3} = 0$$

These equations evidently hold when $b = a$ and $\lambda = 1$. At this point the Hessian is positive definite so a minimum of the cost function has been located. If there is to be another solution, it must satisfy

$$\frac{H}{(H + \lambda L)} = \frac{h}{(h + \lambda l)}$$

which is impossible unless $H/L = h/l$. So when the two decision tasks are not proportional, in effect not the same, there is no other local minimum except at $b = a$ and $\lambda = 1$, the unbiased equilibrium, which is therefore the unique evolutionary stable equilibrium.

A8. Proof that $\frac{dp^*}{dK} > 0$

Let the agent have initial wealth W and a strictly increasing, strictly concave utility function u . The project has two outcomes, either KH or $-KL$ with K a scaling factor and $H > 0$, $L > 0$, and $K > 0$. The probability of the success outcome, p , is drawn from some distribution.

The threshold probability p^* that makes the agent indifferent between taking the project or rejecting it and keeping wealth W satisfies:

$$p^*u(W + KH) + (1 - p^*)u(W - KL) = u(W) \quad (4)$$

Differentiating with respect to K yields:

$$\frac{dp^*}{dK} = \frac{(1 - p^*)Lu'(W - KL) - p^*Hu'(W + KH)}{u(W + KH) - u(W - KL)} \quad (5)$$

The denominator of (5) is strictly positive. Substituting p^* and $(1 - p^*)$ from (4), the numerator of (5) is strictly positive if:

$$[u(W + KH) - u(W)]Lu'(W - KL) > [u(W) - u(W - KL)]Hu'(W + KH) \quad (6)$$

From the strict concavity of u :

$$u(W) - u(W - KL) < KLu'(W - KL)$$

and

$$u(W + KH) - u(W) > KHu'(W + KH)$$

Multiplying the first inequality by $Hu'(W + KH)$ and the second by $Lu'(W - KL)$ gives:

$$[u(W + KH) - u(W)]Lu'(W - KL) > KHLu'(W + KH)u'(W - KL)$$

and

$$[u(W) - u(W - KL)]Hu'(W + KH) < KHLu'(W - KL)u'(W + KH)$$

Therefore, the inequality in (6) strictly holds, which implies that the numerator of (5) is positive, and so $\frac{dp^*}{dK} > 0$. Q.E.D.

A9. Noisy signals of own ability

We set out a stylized signaling model in which individuals receive a noisy private signal of their own ability before choosing their public self-belief. This framework yields a signaling equilibrium characterized by two distinct belief levels. Conditional on true ability, this generates a realistic dispersion of self-beliefs: while overconfidence prevails on average, some high-ability individuals receive low signals and exhibit realized underconfidence. We then apply our loss-aversion framework to explore how evolution calibrates decision-making preferences for these different belief pools.

Model Setup

Individuals have a true ability $a \in \{a_L, a_H\}$ with $a_L < a_H$. Before choosing a public self-belief $b \geq 0$, each individual receives a private binary signal $s \in \{h, l\}$. Signals are informative but imperfect:

$$Pr(s = h \mid a = a_H) = Pr(s = l \mid a = a_L) = q > \frac{1}{2}. \quad (7)$$

Individuals do not know their true type and must choose their belief b based only on their observed signal s .

Choosing b incurs a miscalibration cost $C(b, a)$, minimized at $b = a$. Receivers observe b , and their treatment of the sender is based on the expected ability of the pool choosing that specific b .

Overconfidence and Underconfidence

We consider a separating equilibrium based on private signals, resulting in two equilibrium self-belief levels: a low level b_L and a high level b_H with $b_H > b_L$.

1. The Overconfident Pool (b_H)

Individuals who receive the high signal ($s = h$) choose b_H . To deter low-signal agents from mimicking them, b_H must be inflated above the cost-minimizing posterior belief of the high-signal agents. Because b_H is pushed upward by this deterrence condition, the mean ability of individuals in this group lies strictly below b_H . This pool is systematically overconfident.

2. The Low-Belief Pool (b_L)

Individuals who receive the low signal ($s = l$) choose b_L . Because there is no signaling benefit to inflating b within this pool, they select b_L to minimize their expected miscalibration cost based on their low signal. Therefore b_L is a weighted average strictly between a_L and a_H .

The b_L pool contains two distinct sub-populations:

- True low-ability types (a_L) who correctly received the low signal. For them, $b_L > a_L$. They are mildly overconfident.
- True high-ability types (a_H) who unluckily received the low signal. For them, $b_L < a_H$. They suffer from realized underconfidence.

Loss Aversion

We now integrate loss aversion into the analysis.

In the high-belief pool (b_H), self-beliefs are systematically inflated above the cost-minimizing Bayesian posterior. While it is not strictly guaranteed that $b_H > a_H$ for the true high-ability types in this pool, the pool on average overestimates its probability of success relative to its information set, leading to excessively reckless decision-making (Type I errors). Evolution corrects this average recklessness by pairing the b_H strategy with loss aversion: $\lambda > 1$.

For the low-belief pool (b_L), the optimal preference parameter λ is ambiguous. Because the pool contains a mixture of true types, evolution must calibrate a single expected λ that balances two opposing errors.

- The overconfident low types (a_L) in the pool will be too reckless, accepting projects with negative expected value. To correct them, evolution pushes toward loss aversion ($\lambda > 1$).
- The underconfident high types (a_H) in the pool will be too timid, rejecting projects with positive expected value (Type II errors). To correct them, evolution pushes toward loss-loving preferences ($\lambda < 1$).

The final optimized λ for the b_L pool depends on the aggregate expected decision error of this mixed population. However, because the low-signal recipients constitute the bottom tier of the signaling game, they have no lower type to separate from and therefore do not strategically inflate their confidence. Because their self-beliefs, b_L , are calibrated to minimize expected miscalibration costs based strictly on their information set, the pool incurs no net directional bias on average. As a result, evolution has no incentive to endow this specific strategy with offsetting loss aversion (optimal $\lambda = 1$). Overall, the noisy signal model implies that those drawing the high signal are systematically overconfident and therefore offset by loss aversion ($\lambda > 1$). Conversely, for individuals drawing the low signal, the pool lacks loss aversion ($\lambda = 1$). Consequently, while the true low-ability types in this bottom pool are mildly overconfident without loss aversion, the true high-ability types in this pool uniquely exhibit realized underconfidence combined with no loss aversion.

While the two-signal case demonstrates that the lowest pool is loss-neutral, extending the state space to three or more signals (e.g., high, medium, low) allows underconfidence and loss aversion to coexist. In a three-signal model, the intermediate 'medium' pool must strategically inflate its confidence to separate from the 'low' pool, resulting in an average overconfidence that evolution offsets

with $\lambda > 1$. If a true high-ability individual receives a noisy medium signal, they will adopt this medium-pool strategy. Consequently, their expressed belief remains strictly below their true high ability (realized underconfidence), yet they inherit the loss aversion ($\lambda > 1$) optimized for their overconfident pool.

References

- Banks, J. S., & Sobel, J. (1987). Equilibrium selection in signaling games. *Econometrica*, 55(3), 647-661.
- Cho, I. K., & Kreps, D. M. (1987). Signaling games and stable equilibria. *The Quarterly Journal of Economics*, 102(2), 179-221.
- Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology*, 144(4), 517-546.
- Mailath, G. J. (1987). Incentive compatibility in signaling games with a continuum of types. *Econometrica*, 55(6), 1349-1365.
- Riley, J. G. (1979). Informational equilibrium. *Econometrica*, 47(2), 331-359.