PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Research

Cite this article: Quillien T. 2019 Universal modesty in signal-burying games. *Proc. R. Soc. B* **286**: 20190985.
http://dx.doi.org/10.1098/rspb.2019.0985

Received: 26 April 2019
Accepted: 10 June 2019

Subject Category:
Behaviour

Subject Areas:
behaviour, evolution, cognition

Keywords:
evolutionary game theory, signalling, agent-based model

Author for correspondence:
Tadeg Quillien
e-mail: tadeg.quillien@gmail.com

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.4544108.

THE ROYAL SOCIETY PUBLISHING

Universal modesty in signal-burying games

Tadeg Quillien

Center for Evolutionary Psychology, Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA 93106-9660, USA

TQ, 0000-0003-3632-7729

Why would individuals hide positive information about themselves? Evolutionary game theorists have recently developed the signal-burying game as a simple model to shed light on this puzzle. They have shown that the game has an equilibrium where some agents are better off deliberately reducing the visibility of the signal by which they broadcast their positive traits. However, this equilibrium also features individuals who fully broadcast their positive traits. Here, we show that the signal-burying framework can also explain modesty norms that everyone adheres to: the game contains an equilibrium where all agents who send a signal voluntarily reduce its conspicuousness. Surprisingly, the stability of the two kinds of equilibria rely on very different principles. The equilibrium where some agents brag is stable because of costly signalling dynamics. By contrast, the universal modesty equilibrium exists because buried signals contain probabilistic information about a sender's type, and receivers make optimal use of this information. In the latter equilibrium, burying a signal can be understood as a handicap which makes the signal more honest, but honesty is not achieved through standard costly signalling dynamics.

1. Introduction

In Stanley Kubrick's classic movie Dr Strangelove, the eponymous scientist is surprised to learn that the Soviets did not advertise their deterrence device: 'Of course, the whole point of a Doomsday Machine is lost, if you keep it a secret! Why didn't you tell the world?' Behavioural scientists are in a same state of perplexity with respect to many aspects of human behaviour. Philanthropists often make massive donations under cover of anonymity, and observers consider such anonymous giving more virtuous [1]. Artists hide some of the meaning of their work as Easter eggs for spectators to discover, novelists publish some of their books under pseudonyms and fashionistas sometimes buy expensive designer clothes that do not feature the brand logo. More generally, people are reluctant to advertise their positive traits too openly. These phenomena seem to run contrary to standard evolutionary and economic explanations of human behaviour, which hold that generosity, artistic pursuits, and conspicuous consumption function in large part to attract and impress cooperation and romantic partners [2–7]. Since the whole point about prowesses and good deeds is lost if we keep them a secret, why don't we tell the world?

In a recent paper, Hoffman *et al.* [8] argue that modesty sometimes does make functional sense. They introduce a simple formal model, the signal-burying game, and find equilibria of the game where some individuals are better off reducing the conspicuousness of their broadcasts of positive traits. Their main insight is that the act of deliberately making a signal hard to note is itself a signal: it conveys the message that you are confident that people will find out anyway, or that you do not really care about reaching out to those who would miss the hidden message.

At the equilibrium that Hoffman *et al.* describe, only a fraction of individuals who send a signal reduce its conspicuousness; by necessity, the equilibrium also features individuals who openly advertise their positive traits. Here, we show that their game-theoretic framework also has the potential to explain situations where a modesty norm is observed by everyone. We first

describe the setting of the signal-burying game. Then we present the original equilibrium found by Hoffman *et al.*, followed by the new pooling equilibrium.

2. Model

The signal-burying game pairs a sender with a receiver; the sender tries to convince the receiver to interact with him: if a receiver agrees to interact with the sender, the latter receives a pay-off of 1. The receiver, however, only wants to interact with certain types of senders, but a sender's type is only known to the sender. Senders can be of high, medium, or low quality. Likewise, receivers vary in their selectivity. Strong receivers only want to interact with high senders: they get a pay-off of 1 when interacting with a high sender, and a pay-off of -1 when interacting with a medium or low sender. Weak receivers are less selective: they get a pay-off of 1 when interacting with a high or medium sender, and -1 with a low sender. If the receiver declines to interact, both players get a null pay-off. The sender is a high, medium, or low type with probability p_h, p_m, or p_l, respectively. The receiver is strong or weak with probability q_s or q_w. These probability distributions are common knowledge, but a player's type is only known to that player.[1]

Senders can send a signal, if they wish. They have three options: send a buried signal, a clear signal, or remain silent. By assumption, signalling is prohibitively costly for low types, but is free for high and medium types. Therefore, by sending a signal, a high or medium type can broadcast positive information about himself (namely, that he is not a low type). If a sender sends a buried signal, that signal is only revealed to the receiver with probability r_h (if it was sent by a high sender) or r_m (if it was sent by a medium sender). If it is detected, the receiver knows that the signal was buried.

The interesting feature of the game resides in the fact that both high and medium types are able to send a signal, and therefore a receiver cannot distinguish between these two types if she receives a clear signal. We thus make the final assumption that medium senders outnumber high senders in the population ($p_h < p_m$), such that when both types send clear signals, a strong receiver would get a negative expected pay-off (namely $p_h - p_m$) for interacting with a player sending a clear signal. The main insight of Hoffman *et al.* is that high types can distinguish themselves from medium types, and thereby attract strong receivers, by making their signal harder to detect. We describe their result in the next section; in a later section we show the existence of another kind of burying equilibrium.

3. Standard burying equilibrium

Under certain conditions, there exists an equilibrium where high senders have incentives to send buried signals, while medium senders are better off sending clear signals [8]. As a result, only high senders send buried signals; therefore, burying reliably conveys sender quality. Then, strong receivers only accept buried signals, while weak receivers accept both kinds of signals (nobody accepts silent senders). The equilibrium exists when:

$$r_h > q_w \tag{3.1}$$

and

$$r_m < q_w. \tag{3.2}$$

The first condition says that high senders are better off sending a buried signal. Their buried signal is detected with probability r_h, and when this happens the receiver accepts the interaction. Their clear signal is always detected, but only weak receivers (who make up a proportion q_w of the population) accept the interaction. By a similar logic, the second condition says that medium senders are better off sending a clear signal. Note that (3.1) and (3.2) imply that $r_h > r_m$: buried signals from high types are more likely to be detected than those from medium types. This could represent, for instance, the fact that among people who make charitable donations, some are well-connected enough that they could more easily arrange for their identity to leak if they chose to donate anonymously.

A strong receiver who deviates from the strategy by accepting clear signals is worse off, because only medium senders emit clear signals; a strong receiver who deviates by refusing all signals is worse off, because she forgoes profitable interactions with high senders. Both receiver types refrain from accepting silent senders provided that a sender from which no signal is detected is more likely to be a low type than a high type:

$$p_h(1 - r_h) < p_l. \tag{3.3}$$

If a weak receiver switches to accepting only buried signals, she just loses profitable interactions with medium senders. If (3.1)–(3.3) hold, the strategy profile is a strict Nash equilibrium, and therefore an evolutionarily stable strategy (ESS) [9].

Here, receivers can reliably distinguish between the three types of senders thanks to costly signalling dynamics. According to costly signalling theory, signals can be reliable when the potential benefit of sending the signal, minus its cost, is higher for honest signallers than for liars [10]. In the most straightforward case, signal costs enforce honesty by causing complete *separation* between senders: all individuals with a given trait send the signal, while all individuals without the trait refrain from signalling (because they cannot afford to).

The standard burying equilibrium plays this trick twice. Separation between low senders and other senders is achieved because signalling is prohibitively costly for low types only. Separation between high and medium types is slightly unconventional, yet still relies on costly signalling dynamics. By choosing to bury his signal, a sender pays an opportunity cost: he loses some profitable interactions with the weak receivers who would have accepted a clear signal but fail to detect the buried signal. Because $r_h > r_m$, this opportunity cost is greater for a medium sender. When inequalities (3.1) and (3.2) above hold, this opportunity cost is enough to prevent medium, but not high types, from burying.

We show in the appendix that when (3.1) and (3.2) hold, the standard burying strategy can invade a non-signalling strategy.

4. Pooling burying equilibrium

At the standard burying equilibrium, only high senders bury their signal, leaving medium senders to openly brag about their positive traits. Here, we show that the framework developed by Hoffman *et al.* is also able to account for norms of modesty that are upheld by everyone in the community:

the signal-burying game has an equilibrium where every individual who sends a signal buries this signal.

To see why this is possible, consider the fact that receivers can take advantage of probabilistic information contained in a buried signal. Assume that the buried signals sent by high types are more likely to be detected than the ones sent by medium types; then a significant proportion of the buried signals detected by receivers will come from high types. Under the right conditions, it may, therefore, be in the interest of strong receivers to accept buried signals even in a population where both high and medium types bury their signal. This happens when:

$$p_h r_h > p_m r_m. \tag{4.1}$$

This is because, of all the senders a strong receiver can be paired with, a proportion p_h will be high types; of these, a proportion r_h will see their buried signal detected by the receiver (left-hand side of the inequation). Similarly, a proportion p_m will be medium types; of these, a proportion r_m will see their buried signal detected by the receiver (right-hand side). Therefore, of all senders a receiver can be paired with, a proportion $p_h r_h$ will be high types whose buried signal is detected, and a proportion $p_m r_m$ will be medium types whose buried signal is detected. Because interacting with a medium type is equally as bad as interacting with a high type is good, strong receivers should accept buried signals when doing so guarantees them to meet, on average, more high than medium senders, i.e. when (4.1) holds.[2]

Strong receivers reject silent senders when the absence of a signal is more likely to come from medium or low types than high types:

$$p_h(1 - r_h) < p_m(1 - r_m) + p_l. \tag{4.2}$$

Because we assume that $p_h < p_m$, this condition is always met when (4.1) holds (because (4.1) and $p_h < p_m$ together imply that $r_h > r_m$).

Assume that strong receivers only accept buried signals, and weak receivers accept buried and clear signals. Depending on whether weak receivers also have an incentive to accept silent senders, we have two variants of the pooling burying equilibrium. Weak receivers accept silent senders when the latter are more likely to be medium or high types than low types:

$$p_h(1 - r_h) + p_m(1 - r_m) > p_l. \tag{4.3}$$

When this is the case, there is no opportunity cost to burying, so medium and high types always bury their signal. Otherwise, when weak receivers do not accept silent senders, high and medium senders bury if

$$r_h > q_w \tag{4.4}$$

and

$$r_m > q_w. \tag{4.5}$$

In sum, when (4.3) holds, burying is Nash if (4.1) holds. When (4.3) does not hold, burying is Nash provided that (4.1), (4.4), and (4.5) hold. In the first case, weak receivers accept senders who send a buried signal or stay silent, while in the second case weak receivers accept senders who send buried signals but reject silent senders. In both cases, all receivers are indifferent between accepting and rejecting clear signals; strong receivers accept buried signals and reject silent senders; and medium and high senders send buried signals. We show in the appendix that when (4.1), (4.4), and (4.5) hold, the pooling burying strategy can invade a population of non-signallers.

Note that the pooling burying strategy profile is not ESS: since no sender is sending clear signals, mutant strong receivers who accept clear signals can invade the population via genetic drift. Nonetheless, the pooling equilibrium can be an actual outcome of evolution. We show this in two ways. First, we conduct individual-based simulations, which show that invasions by clear-signalling strategies occur but are never long-lasting, such that, when the pooling burying strategy is Nash, the population spends most of the time at the pooling equilibrium. Second, we study a slightly modified version of the game where players can make mistakes, and show that the pooling strategy is ESS in this modified game.

5. Individual-based simulations

Analytical considerations show that the pooling burying strategy can be invaded, via neutral drift, by mutant strong receivers who accept clear signals. When this happens, high and medium senders are then better off sending a clear signal (since all receivers accept them). If $p_h < p_m$, accepting clear signals is detrimental to strong receivers, so they switch to accepting only buried signals. Provided that the incentives of medium and high senders favour burying, they go back to sending buried signals, and the population is at the pooling burying equilibrium again.

Therefore, when the conditions favouring the stability of the pooling burying strategy hold, analytical modelling predicts that the population will alternate between clear-signalling and burying strategies. In order to see whether one strategy would dominate this cycle, we performed individual-based computer simulations (see appendix for methods). Looking at populations in the last 1000 generations of each simulation, we find that for the simulations within the region of parameter space obeying (4.1), (4.4), and (4.5) and $p_h < p_m$, an average 87% of medium senders, and 93% of high senders, send a buried signal (figure 1). Figure 2 presents a representative trajectory of the evolutionary dynamics. We see that although a typical population oscillates between clear-signalling and burying strategies, the population spends most of the time at the latter equilibrium. The result makes intuitive sense: in a population playing the clear-signalling strategy, strong receivers who accept clear signals are strictly selected against, so the pooling burying strategy rapidly invades. By contrast, in a population playing the pooling burying strategy, only neutral drift can increase the frequency of strong receivers who accept clear signals, so such invasions are rare.

6. Analytical model with mistakes

Strategies that are not evolutionarily stable can become so when the game is modified to take into account the possibility that individuals make mistakes, or that some individuals are unable to take a given action [11,12]. In the kinds of situations modelled by the signal-burying game, it is reasonable to assume that individuals who intend to bury their signal sometimes fail to do so. This is likely to occur, notably because individuals will try to steer a

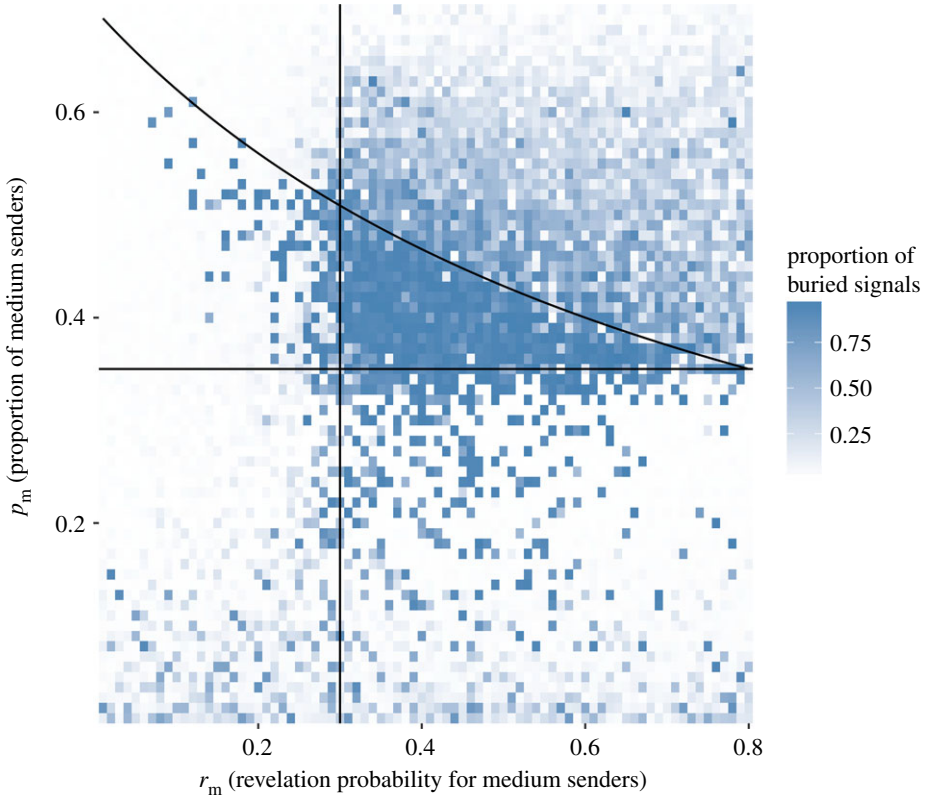

Figure 1. Average proportion of medium senders sending buried signals in the last 1000 generations of each simulation, as a function of p_m and r_m. Each pixel corresponds to one simulation. Other parameter values were fixed at $r_h = 0.8$, $r_l = 0.4$, $q_w = 0.3$, $q_s = 0.7$, $p_l = 0.3$, $p_h = 1 - p_l - p_m$. Between the three black lines is the region of parameter space obeying (4.1), (4.4), and (4.5) and $p_h < p_m$. (Online version in colour.)

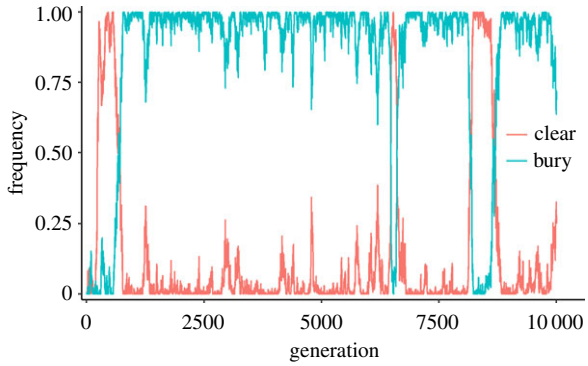

Figure 2. Frequency of medium senders who send a clear (red line) or buried signal (blue line) over time in a representative simulation run. Here, $r_m = 0.4$, and $p_m = 0.4$. Other parameter values were $r_h = 0.8$, $r_l = 0.4$, $q_w = 0.3$, $q_s = 0.7$, $p_l = 0.3$, $p_h = 0.3$. (Online version in colour.)

middle ground between making their signal too hard to detect (which would make them forgo too many interactions) and making it too obvious (in which case observers will stop considering the signal as 'buried' at all). Sometimes they will err on the side of making the signal detectable enough, and end up making the signal too obvious.

In the appendix, we formalize this argument by studying a game where the trade-off between detectability and discreetness is explicitly modelled. Here, we simply analyse a game similar to the one described in the previous sections, with one modification: with probability ϵ, a player who wishes to send a buried signal sends a clear signal instead. An alternative, but mathematically equivalent model would be to assume that ϵ represents a proportion of senders who are unable, for non-heritable reasons, to send buried signals.

In this game, the condition for strong receivers to accept buried signals, under the pooling burying strategy, is the same as in the basic game: $p_h r_h > p_m r_m$ (inequation (4.1)). Since medium and high types have the same error rate, introducing mistakes does not change the diagnosticity of buried signals.

Strong receivers reject silent senders provided that

$$p_h(1 - r_h) < p_m(1 - r_m) + \frac{p_l}{1 - \epsilon}.$$

Just as in the no-mistakes model, this condition is always met provided that (4.1) and $p_h < p_m$ are true.

Weak receivers reject silent senders if:

$$p_h(1 - r_h) + p_m(1 - r_m) < \frac{p_l}{1 - \epsilon}. \tag{6.1}$$

When (6.1) does not hold, high and medium senders always bury. Otherwise, they have an incentive to bury provided that:

$$(1 - \epsilon)r_h + \epsilon q_w > q_w \tag{6.2}$$

and

$$(1 - \epsilon)r_m + \epsilon q_w > q_w. \tag{6.3}$$

These inequalities reduce to $r_h > q_w$ and $r_m > q_w$, which are simply inequalities (4.4) and (4.5).

Medium and high types are equally likely to send a clear signal by mistake; therefore, because $p_h < p_m$, a strong receiver should not accept clear signals, and the clear-signalling strategy cannot invade by neutral drift. Therefore, if the pooling burying strategy is a Nash equilibrium, it is also ESS.

7. Discussion

We have shown that, in addition to the burying equilibrium identified by Hoffman *et al.*, the signal-burying game has a 'pooling burying equilibrium' where all individuals who send a signal bury that signal.

It follows that the game can be used to model a wide range of phenomena where people do not advertise their positive attributes as openly as they could. The standard burying equilibrium has great explanatory power to explain instances of 'elitist' modesty, where senders of inconspicuous signals refrain from bragging because they do not want to be confused with more plebeian signallers–as when sophisticated members of the upper class wear logo-less designer clothes so as to differentiate themselves from 'nouveaux riches' [8]. But most norms of self-effacement are more pervasive–intuitively, people with positive characteristics should take every opportunity they can to display them, yet people very rarely openly brag about how much money they make, or spontaneously list their scholarly achievements, to individuals they have just met.[3] Indeed, we would view such behaviour as odd, and draw negative inferences about the individual. The pooling burying equilibrium explains this by a logic similar to that which governs 'full disclosure' phenomena in animal signalling. In a famous example, the pitch of a toad's croak is a good indicator of its size, so toads vocalize in order to intimidate their rivals [15]–but surprisingly, even small toads vocalize, giving away their small size. This is because not vocalizing at all is the worst signal a toad could send about his formidability: it would be an implicit confession that the animal has strictly nothing good to show [16]. In the context of signal burying, this argument is turned on its head: subtlety in signalling is itself a signal (i.e. deliberate burying is the equivalent of croaking, not of staying silent), and even individuals who are not adept at demonstrating their positive attributes in a subtle manner may find themselves inclined to do so. As an example, even someone ill-at-ease with the socially acceptable ways of demonstrating one's intelligence (subtle cultural references, clever jokes, etc.) might still be better off not constantly mentioning that he has a college degree. Plain bragging about his education would give away the fact that he is not smart enough to advertise his mental agility indirectly.

A remarkable outcome of the present analysis is that the two burying equilibria are stable thanks to completely different mechanisms.

The pooling burying strategy is stable because strong receivers make optimal use of the probabilistic information contained in the signal. When buried signals sent by high types are more likely to be detected than those sent by medium types, most detected signals come from high senders (or at least, the proportion of detected signals that come from high senders is greater than the ratio of high to medium senders in the population), which makes burying status a cue to sender quality.

As a result, both medium and high senders have an incentive to bury their signal in order to attract strong receivers. By doing so, they trade-off the conspicuousness of their signal against its inferential content: they are willing to reduce the number of receivers they reach, in order to present themselves in the best light. One may describe senders who play this strategy as imposing a handicap on themselves. Deliberately diminishing the visibility of their signal hurts their chances of interacting with weak receivers, who are just looking for evidence that they are not low types. Furthermore, the magnitude of this handicap must be larger for medium than for high types ($r_m < r_h$) for the strategy to be stable. Therefore, the situation bears a superficial similarity to the Handicap principle theorized by Zahavi [4] and later formalized by Grafen [17], whereby costly signals can promote communication when interests conflict.

However, this similarity is deceptive. Standard costly signalling models rely on a pay-off differential between sender types; namely, the marginal net pay-off of sending the signal must be higher for honest senders than dishonest individuals [10]. This pay-off differential makes the signal reliable: honest signallers can better afford to send the signal, therefore they are more likely to do so—as a result, signals disproportionately come from honest types. By contrast, at the pooling burying equilibrium, even though medium senders earn lower expected pay-offs than high senders, both types of senders are equally likely to signal: therefore, the pay-off differential is not what makes the signal reliable. Rather, receivers directly exploit the information provided by the fact that buried signals from high senders are more easily revealed.

The standard burying strategy does rely on costly signalling: burying is a reliable cue to high type because the opportunity cost of burying dissuades medium, but not high senders from burying. It is a separating equilibrium, where only high senders send a buried signal. By contrast, the existence of the pooling burying equilibrium shows that modesty can make a signal more convincing even in the absence of a separation between sender types.

8. Conclusion

The signal-burying game highlights a potential reason why individuals may not advertise positive information about themselves to the fullest extent: making a signal harder to notice can itself act as a signal. If senders with the most desirable qualities are also the most adept at conveying these qualities in a subtle fashion, then subtlety can reliably convey possession of these qualities. The present paper offers the additional insight that even the senders who are bad at skillfully obfuscating their message may find it worthwhile to do so. Doing otherwise would be an implicit confession that they are not skillful buriers, and therefore not high-quality individuals. In many cases, maintaining ambiguity about one's quality is worth the risk of not getting the message through.

Data accessibility. Data from the simulations are available as part of the electronic supplementary material.

Competing interests. I declare I have no competing interests.

Funding. No funding has been received for this article.

Acknowledgements. I thank Moshe Hoffman, Daniel Sznycer, Leda Cosmides, and two anonymous referees for helpful comments and suggestions.

Endnotes

[1]For simplicity of exposition, we omit some of the parameters used in the original model by Hoffman *et al.*

[2]We also present a derivation of (4.1) in a Bayesian format in the appendix.

[3]Such self-effacement norms are especially pronounced in some non-Western contexts, for instance, in East Asia [13,14].

Appendix A

(a) Derivation of inequality (4.1) using Bayes' rule

Let us assume that medium and high senders all send a buried signal. A receiver who detects a buried signal can infer the probability that the sender is a high type using Bayes' rule:

$$P(\text{high}|\text{buried}) = \frac{P(\text{buried}|\text{high})P(\text{high})}{P(\text{buried})},$$

where $P(\text{buried}|\text{high})$ is the probability that a sender sends a buried signal that gets detected, given that the sender is a high type. Since by assumption all high senders send a buried signal, this is equal to r_h. Therefore, we have:

$$P(\text{high}|\text{buried}) = \frac{r_h p_h}{r_h p_h + r_m p_m}.$$

A similar argument reveals that:

$$P(\text{medium}|\text{buried}) = \frac{r_m p_m}{r_h p_h + r_m p_m}.$$

A strong receiver should accept a buried signal when $P(\text{high}|\text{buried}) > P(\text{medium}|\text{buried})$, that is, when $r_h p_h > r_m p_m$.

(b) Invasion potential of burying strategies

A non-signalling strategy is a strategy where no sender ever signals, and receivers are predisposed to always reject senders, no matter which signal they receive.

Assume every agent initially plays the non-signalling strategy. Since no signals are sent, via genetic drift some receivers can eventually start accepting all signals. In particular, weak receivers who accept all signals can increase in frequency regardless of the strategies that drift introduces in the sender population, because selection against signalling in low senders prevents drift from increasing the frequency of low senders who signal.

As a result, selection starts favouring high and medium senders who send clear signals. Because $p_h < p_m$, strong receivers who accept clear signals are counter-selected, yet strong receivers who accept buried signals can increase in frequency via genetic drift.

If $r_h > q_w$, and $r_m < q_w$ (that is, if (3.1) and (3.2) hold), once the proportion of strong receivers who accept buried signals becomes high enough, there is an incentive for high senders (but not medium senders) to send buried signals, and the standard burying strategy invades.

If, instead, $r_h > q_w$ and $r_m > q_w$ (that is, if (4.4) and (4.5) hold), once the proportion of strong receivers who accept buried signals becomes high enough, there is an incentive for high senders and medium senders to send buried signals. Strong receivers keep accepting these buried signals if $p_h r_h > p_m r_m$ (condition (4.1)), and the pooling burying strategy invades.

(c) Individual-based simulations

Each simulation involved a population of 300 senders, and a population of 300 receivers, allowed to evolve for 10 000 generations. For each simulation, we recorded the average frequency of each allele during the last 1000 generations. At the beginning of each generation, each sender was paired randomly with a receiver. Then each pair played a signal-burying game, after which all agents died and reproduced asexually. Within each population, an agent's expected number of offspring was equal to the ratio of its total pay-off to the average pay-off in that population (pay-offs were first standardized such that the agent with the lowest pay-off had pay-off 1; this was achieved by subtracting, to the pay-off of each agent, the pay-off of the agent with the lowest pay-off, then adding 1 to the pay-off of each agent. This procedure ensured that no agent had a negative pay-off before the selection phase). The size of each population was kept constant at 300 agents during a simulation.

Sender and receiver type were non-heritable, and determined at random, according to the probabilities p_h, p_m, p_l, q_s, and q_w, for each agent at the start of every generation. A sender's genotype consisted of three genes, determining the sender's behaviour for each possible type it could be (high, low, or medium); each gene could take either Bury, Clear, or Quiet as alternative alleles. A receiver's genotype consisted in 6 genes, determining the receiver's behaviour (Accept or Reject) for each combination of receiver's type (strong or weak) and type of signal received (buried signal, clear signal, no signal). Each agent inherited its parent's genotype, subject to mutation: with independent probability 0.002, each gene would undergo mutation, where mutation consisted of replacing the gene's current value with a random uniform draw over the set of possible alleles for this gene.

We conducted 5600 simulations, while varying the values of r_m (from 0.01 to 0.8) and p_m (from 0.01 to 0.7) in 0.01 increments. Other parameter values were fixed at $r_h = 0.8$, $r_l = 0.4$, $q_w = 0.3$, $q_s = 0.7$, $p_l = 0.3$, $p_h = 1 - p_l - p_m$. Low senders incurred a cost of 10 when sending a signal. Each simulation was initialized with a population where senders never signalled and receivers were predisposed to reject all senders, regardless of the signal they would receive. Simulation software is written in JavaScript; script and data for the simulations are available in the electronic supplementary material.

(d) Modelling the trade-off between detectability and discreetness

We study a signal-burying game identical to the basic version analysed in the main text, except that the probability of detection of a buried signal (r) is determined endogenously. Instead of making a discrete decision to bury or not, senders can decide how much to bury their signal, via a continuous variable $b \in [0, 1]$. A sender's signal is detected by the receiver with probability $r = 1 - \alpha b$. Additionally, whereas in the basic version of the game, a receiver who detects a signal always identifies the signal as buried, we now assume that there is a probability b that a detected signal is identified as having been buried.

Therefore, senders are faced with a trade-off: the deeper they bury a signal, the less likely it is to be detected, but the more likely it is to be identified as having been buried. We allow α to differ between high and medium senders. $\alpha_h < \alpha_m$, for instance, would mean that high senders are smarter than medium senders about how they hide their

signal. We assume that $p_h(1-r_h)+p_m(1-r_m)<p_l$, such that receivers have an incentive to reject silent senders.

Assume that strong receivers accept only buried signals, while weak receivers accept all signals. Then, the expected pay-off of a high sender is given by:

$$E(h)=r_h(b_h q_s+q_w),$$

because a proportion r_h of receivers detect the signal; of these, all weak receivers accept the signal, while only the strong receivers who identify the signal as buried accept it. This is equivalent to:

$$E(h)=(1-\alpha_h b_h)(b_h q_s+q_w).$$

The derivative of $E(h)$ with respect to b_h is

$$E(h)'=-2\alpha_h q_s b_h-\alpha_h q_w+q_s.$$

To find the value of b_h that maximizes $E(h)$, we set $E(h)'=0$ and solve for b_h. This yields:

$$b_h^*=\frac{q_s-\alpha_h q_w}{2\alpha_h q_s}. \tag{A 1}$$

If the right-hand side of the above equation is below 0 or above 1, then $b_h^*=0$ or $b_h^*=1$, respectively (since b_h is constrained to range between 0 and 1).

Similarly, the expected pay-off of a medium sender, $E(m)$, is maximized when

$$b_m^*=\frac{q_s-\alpha_m q_w}{2\alpha_m q_s}, \tag{A 2}$$

with the same constraint as above on the range of b_m.

When are strong receivers better off accepting buried signals? Of all the senders they encounter, a proportion p_h are high types; of these, a proportion r_h see their signal detected; of these signals, a proportion b_h^* are identified as being buried. Therefore, of all senders a receiver can be paired with, a proportion $p_h r_h b_h^*$ will be high types whose signal is detected and identified as buried; similarly, a proportion $p_m r_m b_m^*$ will be medium types whose signal is detected and identified as buried. Strong receivers are therefore better off accepting buried signals when

$$p_h r_h b_h^* > p_m r_m b_m^*. \tag{A 3}$$

For the burying equilibrium to be stable, it must also be the case that strong receivers are better off rejecting clear signals:

$$p_h r_h(1-b_h^*) < p_m r_m(1-b_m^*). \tag{A 4}$$

A final condition for burying to be a Nash equilibrium is that high senders bury their signal to some extent, that is $b_h^*>0$. From (A 1), we see that this happens when:

$$\alpha_h < \frac{q_s}{q_w}. \tag{A 5}$$

When (A3)–(A5) hold, the burying strategy is Nash. In order for it to be ESS, we also need some signals to be identified as clear at least a fraction of the time; otherwise selection on receivers' reaction to clear signals disappears. This condition is met when $b_m^*<1$; from (A2), this is equivalent to $\alpha_m > (q_s/2q_s+q_w)$.

Note that this burying equilibrium encompasses as special cases the standard burying equilibrium (with high types burying and medium types always sending clear signals) when $\alpha_h<(q_s/q_w)$ and $\alpha_m>(q_s/q_w)$, and the pooling equilibrium (with every agent who sends a signal burying that signal to some extent) when both α_m and α_h are smaller than (q_s/q_w).

As a numerical example, when $q_s=0.7$, $q_w=0.3$, $p_h=0.3$, $p_m=0.4$, $\alpha_h=0.5$, and $\alpha_m=0.8$, we have:

$$b_h^*=0.79, \quad r_h=0.61$$
$$b_m^*=0.41, \quad r_m=0.67.$$

Then $p_h r_h b_h^*=0.14$, $p_m r_m b_m^*=0.11$, $p_h r_h(1-b_h^*)=0.04$, $p_m r_m(1-b_m^*)=0.16$, thus inequalities (A3) and (A4) hold, and the burying strategy is ESS. Note that here, unlike the models in the main text, signals from high types happen to be slightly *less likely* to be detected than signals from medium types. Buried signals still function as a cue to high quality because signals from high types are less 'obvious': they are more likely to be identified as having been buried than signals from medium types. Here, burying evolves because high types are better at making apparent the encrypted nature of their message.

References

1. De Freitas J, DeScioli P, Thomas KA, Pinker S. 2018 Maimonides ladder: states of mutual knowledge and the perception of charitability. *J. Exp. Psychol. General* **148**, 158–173. (doi:/10.1037/xge0000507)
2. Trivers RL. 1971 The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57. (doi:10.1086/406755)
3. Spence M. 1973 Job market signaling. *Q. J. Econ.* **87**, 355–374. (doi:10.2307/1882010)
4. Zahavi A. 1975 Mate selection—a selection for a handicap. *J. Theor. Biol.* **53**, 205–214. (doi:10.1016/0022-5193(75)90111-3)
5. Miller G. 2000 *The mating mind: how sexual choice shaped the evolution of human nature*. New York City, USA: Anchor.
6. Barclay P. 2013 Strategies for cooperation in biological markets, especially for humans. *Evol. Human Behav.* **34**, 164–175. (doi:10.1016/j.evolhumbehav.2013.02.002)
7. Sznycer D *et al.* 2017 Cross-cultural regularities in the cognitive architecture of pride. *Proc. Natl Acad. Sci. USA* **114**, 1874–1879. (doi:10.1073/pnas.1614389114)
8. Hoffman M, Hilbe C, Nowak MA. 2018 The signal-burying game can explain why we obscure positive traits and good deeds. *Nat. Human Behav.* **2**, 397–404. (doi:10.1038/s41562-018-0354-z)
9. Maynard Smith J 1982 *Evolution and the theory of games*. Cambridge, UK: Cambridge University Press.
10. Higham JP. 2013 How does honest costly signaling work? *Behav. Ecol.* **25**, 8–11. (doi:10.1093/beheco/art097)
11. Boyd R. 1989 Mistakes allow evolutionary stability in the repeated prisoner's dilemma game. *J. Theor. Biol.* **136**, 47–56. (doi:10.1016/S0022-5193(89)80188-2)
12. Sherratt TN, Roberts G. 2001 The importance of phenotypic defectors in stabilizing reciprocal altruism. *Behav. Ecol.* **12**, 313–317. (doi:10.1093/beheco/12.3.313)

13. Heine SJ, Lehman DR, Markus HR, Kitayama S. 1999 Is there a universal need for positive self-regard? *Psychol. Rev.* **106**, 766–794. (doi:10.1037/0033-295X.106.4.766)
14. Eid M, Diener E. 2001 Norms for experiencing emotions in different cultures: inter-and intranational differences. *J. Pers. Soc. Psychol.* **81**, 869–885. (doi:10.1037/0022-3514.81.5.869)
15. Davies NB, Halliday TR. 1978 Deep croaks and fighting assessment in toads *Bufo bufo*. *Nature* **274**, 683–685. (doi:10.1038/274683a0)
16. Krebs JR, Dawkins R. 1984 Animal signals: mindreading and manipulation. In: *Behavioural Ecology: an evolutionary approach*. pp. 380–402. Blackwell Scientific Publication: Oxford, United Kingdom.
17. Grafen A. 1990 Biological signals as handicaps. *J. Theor. Biol.* **144**, 517–546. (doi:10.1016/S0022-5193(05)80088-8)