Robust Communication

Alistair Barton*

First Version: August 16, 2022
This Version: January 15, 2025
Click here to access current version

Abstract

I study finite-state sender-receiver games with nearly independent preferences. Recent literature shows persuasive communication is possible when sender preferences are public and state-independent. I demonstrate such communication is fragile to perturbations introducing privacy to the sender's utility, but may become robust under slightly state-dependent perturbations. Developing a novel representation of equilibria as communication graphs, I show equilibria become robust if and only if their communication graph exhibits specific geometries consistent with the sender's state-dependence. I apply this result to show when empathy improves/*impedes* communication, and when money-burning benefits senders with state-dependent preferences, despite not benefiting senders with state-independent preferences.

In many settings, agents with plausibly independent preferences seek to influence one another. Salespeople incentivized by a commission steer consumers toward more expensive purchases, regardless of whether those purchases satisfy the consumers' needs. Stockbrokers receive brokerage fees determined by the trade volume of their client, independent of the wisdom of such trades. Politicians campaign for votes regardless of their ability or intention to advance their constituents' interests.

In such environments, communication seems intuitively non-credible. However, recent literature (Chakraborty and Harbaugh 2010; Lipnowski and Ravid 2020) has characterized how persuasion is possible in the canonical sender-receiver game when the sender's preference is strictly independent of their private information — so-called 'transparent' preferences.

*New York University, Email: alistair.barton@nyu.edu
I would like to thank Dilip Abreu and Joyee Deb for feedback and guidance on this project, I am grateful to Elliot Lipnowski, Basil Williams, Robin Lenoir, Doron Ravid, Erik Madsen, Debraj Ray, and Johannes Hörner for helpful comments.

Transparent preferences, however, are a knife-edge scenario, realistically a sender's preference will be affected by *some* private information: salespeople may have private information about their compensation schemes; stockbrokers may have idiosyncratic biases towards buying or selling; the value a politician places on winning an election will not be known to the public. In these cases the sender's preference is no longer transparent, but remains statistically independent of the receiver's preference: whether or not an action benefits the receiver is irrelevant to the sender's valuation.

The starting point for this paper is a fragility result: persuasion is impossible between parties with independent preferences if the sender's preference is even slightly private.[1] *Since perfect knowledge of another's preference is implausible, these earlier-studied equilibria do not describe realistic communication between parties with independent preferences.*

In this paper, we instead interpret these fragile equilibria as approximations of the communication that occurs when the sender's preference is *slightly* state-dependent. Examples include a salesperson who feels slight sympathy for the buyer; a broker who is slightly lying averse when making their recommendation; a politician who places value on being true to their principles.

This interpretation does not apply to all equilibria: some equilibria cannot be approximated for any type of state-dependence; similarly, for some types of state-dependence persuasion is impossible. Our main result is a precise characterization of the relationship between an equilibrium and the types of state-dependence that preserve it: given an equilibrium, what types of state-dependencies are necessary to make it robust; and given a state-dependence, which equilibria can be robustly approximated.

Formally, we consider the canonical sender-receiver game where the sender possesses two dimensions of private information, denoted (θ, ω). Only the first component, θ, is relevant to the receiver; this represents the *state* in our model. The second component, ω, introduces idiosyncratic randomness/privacy to the sender's preference. We consider sender utilities $u^{\varepsilon V}$ composed of a transparent utility u^0 and a perturbation V:

$$u^{\varepsilon V}(a, m|\theta, \omega) = u^0(a, m) + \varepsilon V(a, m|\theta, \omega) \tag{1}$$

where m is the sender's message — our model thus nests cheap talk, money burning, and signalling.[2] We assume θ and a belong to finite sets.

When $\varepsilon = 0$ the sender's preference is transparent (u^0), and the equilibria of this model are the equilibria studied by Lipnowski–Ravid. We call these *candidate equilibria*. Transparent preferences mean that the sender has no incentive for 'honesty' in candidate equilibria: the sender must be indifferent between *all* equilibrium messages, even those messages that are off-path in their

[1] Diehl and Kuzmics 2021 prove an early result specific to the Chakraborty–Harbaugh model. Steg et al. 2023 is another closely related paper. Both are discussed in the literature review.

[2] Cheap talk occurs when the utility is independent of the message m, money burning is when there are costs associated with different messages m independent of the action a and state θ, signalling refers to models where the cost of a message m varies across states θ.

particular state. This strong reliance on indifference makes candidate equilibria tractable, but also leads to their fragility.

The idiosyncrasy ω serves the role of a robustness check in our model. To illustrate, consider a candidate equilibrium σ_0. A desirable property is that as $\varepsilon V \to 0$, we obtain equilibria $\sigma_{\varepsilon V}$ that approximate σ_0; thus, a slight miscalibration of the sender's preference does not significantly disturb the equilibrium. We find that persuasion is generally impossible when V is θ-independent. In this case, the sender's preference is only affected by their private idiosyncrasy ω, and thus only ω will determine their choice of message. The resulting messages are then uninformative about the state θ. The only equilibria that survive this robustness check resemble babbling.

We then consider the case of slightly state-dependent sender preferences. To analyze how this affects equilibria, we consider perturbations V lying within some fixed open set $\mathcal{O}$ of state-dependent perturbations. If a candidate equilibrium can be approximated with such perturbations, we say it is *$\mathcal{O}$-robust*. The set $\mathcal{O}$ captures the types of state-dependence that can be introduced into the sender's preference while preserving communication (even if the sender's preference is slightly idiosyncratic).

The main result of this paper is a precise characterization of $\mathcal{O}$-robustness. A core and novel element of our analysis is the *communication graph*, a bipartite graph that describes the sender's strategy. Its vertices are the states and messages, with edges connecting states with each message sent from that state with positive probability. Such a graph is a coarse representation of the equilibrium, as it omits the probability with which each message is sent; nevertheless, it contains sufficient data to study $\mathcal{O}$-robustness.

In Section 3, we show that under weak (generic) conditions, the communication graphs corresponding to candidate equilibria must be connected and/or contain a cycle.

Studying the incentive problem of the sender, we find that a candidate equilibrium is $\mathcal{O}$-robust, for *some* open set $\mathcal{O}$, iff its communication graph is acyclic. Generically then, such communication graphs are acyclic and connected — ie. tree graphs. Notably, this precludes separating equilibria.

We then characterize the sets $\mathcal{O}$ for which a candidate equilibrium is $\mathcal{O}$-robust. This is a joint condition on the communication graph and perturbations that we call *graph monotonicity*.[3] Any state-dependence that satisfies this monotonicity condition for a given communication graph thus preserves the associated candidate equilibrium.

In Section 5, we consider the implications of this work, focusing on the natural modification of empathy. Our results show that empathy is effective at preserving communication in models with *quantitative* actions (formally defined later), where the sender seeks to influence the receiver to take 'extreme' actions. An example is a broker seeking to persuade an investor to make trades, either buying or selling large quantities of an asset.

[3]Graph monotonicity is closely related to Mechanism Design's 'cyclic monotonicity'. We omit the 'cyclic' adjective to avoid confusion with the acyclicity of communication graphs. The relationship is further discussed in Appendix B.3.

However empathy can be an impediment to communication if the receiver's actions are more *qualitative*. An example is an insurance broker persuading a consumer to buy policies, offering nested levels of coverage, and trying to persuade the consumer to purchase the more expensive options. In such settings, persuasive candidate equilibria exist, but cheap talk may be unpersuasive if the sender is empathetic.

We use this example to show how money burning can be a useful communication technology for the sender, despite never benefiting a sender with transparent preferences. This is because money-burning allows the sender to persuade the receiver using distinct communication graphs, more compatible with their preference. In our example with qualitatively different insurance policies this allows the sender to discretely increase their equilibrium utility.

Lastly, we consider two signalling perturbations. The first perturbs the sender's preference toward lying aversion, which only preserves a few simple geometries of communication. Our last perturbation preserves all acyclic candidate equilibria. This perturbation is interpreted as a low-credibility disclosure technology — it corresponds to perturbing the transparent preference model toward a model of verifiable disclosure.

Related Literature

This work primarily lies at the intersection of cheap talk communication models and the Harsanyi purification literature.

The original cheap talk model (Crawford and Sobel 1982) specifically studies the case where the state/action space is $\Theta = \mathcal{A} = [0, 1]$, the sender's preference is state-dependent, and the receiver has a unique preferred action for every belief. The existence of informative equilibria is a question of the parametric 'alignment' between the sender and receiver preferences — loosely resembling our notion of empathy. This requires state-dependence in the sender's preference: if the sender has a strict, state-independent preference (e.g. higher actions are better), then no information can be communicated in equilibrium — if two messages led to different actions, the sender would always want to deviate to send the message corresponding to the preferred action.

Chakraborty and Harbaugh 2010 are the first (to my knowledge) to study the intriguing possibility of communication with state-independent preferences. They consider models with multi-dimensional state/action space $\Theta = \mathcal{A} \subseteq \mathbb{R}^n$ $(n > 1)$ where the receiver has quadratic preferences, and show that there exists an informative equilibrium for *any* transparent sender preference. Their argument uses hyperplanes to partition the state space into sets whose induced actions the sender is indifferent between. This requires a multi-dimensional space, as these partitions are obtained by continuously rotating a dividing hyperplane. The sender having a public preference is essential — Diehl and Kuzmics 2021 show that these equilibria vanish under slight uncertainty in the sender's preference, as Harsanyi's purification fails to apply (discussed more below). Our fragility theorem adapts this result to general finite-action environments.

The model with a public sender preference is extended to general state spaces by Lipnowski and

Ravid 2020. They adopt a belief-based approach (as pioneered by Kamenica and Gentzkow 2011), showing that communication can be understood in terms of the quasiconcave and quasiconvex envelopes of the sender's indirect utility over receiver beliefs. This shows that equilibria with communication exist iff there exists some Blackwell experiment that guarantees the sender a better/worse *ex post* payoff than babbling. This result applies to models with finite state- and action-spaces, which is the environment we focus on.

Independent of the current paper, Steg et al. 2023 consider a similar question of which Lipnowski–Ravid equilibria can be robustly approximated in binary-state models. Their notion of robustness subtly differs from our own, as they study perturbations to a slightly state-dependent sender preference (with no idiosyncrasy). The resulting equilibria may be sensitive to the specific state-dependent preference that is being perturbed.[4] In contrast, we require equilibria to be approximated from an open set of perturbations. This slightly stronger condition allows us to (1) obtain equilibria that are robust to slight/idiosyncratic changes in the direction of state-dependence, and (2) more tractably analyze models with multiple states.

Our notion of robustness is derived from the notion Harsanyi purification, first developed in Harsanyi 1973 asks when mixed-strategy equilibria can be interpreted, not as players randomizing, but as players choosing pure strategies based on small idiosyncrasies in their preference. This idiosyncrasy results in players almost always having a strict best response to the (seemingly) mixed strategies of other players.

Harsanyi shows that, for *generic* utility functions in finite normal form games, any equilibrium of the game without idiosyncrasies can be approximated by equilibria of the game with slight idiosyncrasies (as a lower hemicontinuity result).

This result may fail for specific player utilities — one such case is communication games with transparent preferences, another is cheap talk models in general. In Section 2 we discuss the reason why purification fails in these situations (more severely in the first case), and how adding state dependence allows purification to apply. We develop novel techniques to approach this problem in extensive-form games with incomplete information.

1 Model

We consider the canonical sender-receiver game where the sender begins by observing their private information then chooses a message to send to the receiver. Upon reception, the receiver chooses an action, determining the utility obtained by both parties.

The sender's private information (θ, ω) is drawn from a two dimensional probability space

[4] In the notation of eq. 1, Steg et al. 2023 studies equilibria that are approximated as perturbations converge to a given V when ε is a fixed small number — sensitivity occurs when the approximation error is constant as $\varepsilon \to 0$. We essentially reverse the order of limits, studying equilibria that can be approximated as $\varepsilon \to 0$ whenever V is in an open set $\mathcal{O}$. These notions of robustness are compared in more detail in Appendix B.2.

$(\Theta \times \Omega, 2^{\Theta} \otimes \mathcal{F}_\omega, \mathbb{P} = \mu \otimes \mathbb{P}_\omega)$. The finite set Θ contains the receiver-relevant state distributed according to the prior μ — henceforth we will use **state** exclusively to refer to these objects — while Ω contains sender **idiosyncrasy** types distributed according to $\mathbb{P}_\omega$. This imposes independence between states and idiosyncrasies are independent. We assume that $(\Omega, \mathcal{F}_\omega, \mathbb{P}_\omega)$ is a rich probability space.

After seeing the state and their idiosyncrasy, the sender chooses a message from the message space M. We impose little *a priori* structure on M, only equipping it with the discrete topology. The reader may consider this message space to be some countable set of messages/signals.

The receiver observes this message (but not the state nor idiosyncrasy), then chooses from the set of mixed actions $\Delta\mathcal{A}$ where the underlying pure action set $\mathcal{A}$ is finite. We represent general actions in $\Delta\mathcal{A}$ by p to emphasize their possibly mixed nature, reserving a for actions that are restricted to be pure.

The sender and receiver have respective utility functions over pure actions

$$\begin{aligned} u^{\varepsilon V} &: (\mathcal{A} \times M) \times (\Theta \times \Omega) \to \mathbb{R} \\ u_R &: \mathcal{A} \times \Theta \to \mathbb{R}. \end{aligned}$$

With slight abuse of notation we also use $u^{\varepsilon V}, u_R$ to denote their von Neumann–Morgenstern extensions to mixed actions $p \in \Delta\mathcal{A}$.

It is convenient to adopt a compact designation for an action-message pair: $\pi \equiv (p, m) \in \Delta\mathcal{A} \times M$, or $\alpha \equiv (a, m) \in \mathcal{A} \times M$ when the action is constrained to be pure. We sometimes refer to such π as a message or an action to emphasize the relevant component.

We decompose the sender's preference into a **transparent** (ie. public, state-independent) component u^0, and a **perturbation** V:

$$u^{\varepsilon V}(\pi|\theta, \omega) = u^0(\pi) + \varepsilon V(\pi|\theta, \omega)$$

where ε controls how far the preference is from being transparent. We focus on the case where ε is small. Since state-dependence and idiosyncrasy both factor through the perturbation V, they should be understood as a second order concerns, reflecting slight biases of the sender.

We will say a sender preference is **idiosyncratic** if it depends on the idiosyncrasy $\omega \in \Omega$. Idiosyncratic perturbations will be denoted with a capital V, emphasizing their interpretation as a random variable. Non-idiosyncratic perturbations will be denoted v, these are elements of $\mathbb{R}^{\mathcal{A}\times M\times\Theta}$ and will be refered to as **modifications**.

While this paper focuses on cheap talk models, we allow the sender's utility to also feature message dependence. This allows us to extend our analysis to money-burning models (where u^0 depends on the message), and 'weak' signalling models (where V depends on the message). Readers may find it convenient to primarily consider cheap talk models in the technical sections of this paper, with the understanding that these techniques generally extend to broader settings.

Our interest is in Perfect Bayesian Equilibria of this sender-receiver game. Formally, these are triplets $\sigma \equiv (\mathcal{M}, \nu, \mathcal{P})$ composed of a sender strategy $\mathcal{M} : \Theta \times \Omega \to \Delta M$, a receiver posterior belief function $\nu : M \to \Delta\Theta$,[5] and a receiver action $\mathcal{P} : M \to \Delta\mathcal{A}$ such that

$$m \in \arg\max_{m' \in M} u^{\varepsilon V}(\mathcal{P}(m'), m' | \theta, \omega) \quad \text{for } m \in \text{supp}(\mathcal{M}(\theta, \omega)) \tag{2a}$$

$$\nu(\theta | m) = \mathbb{P}[\theta | m] \quad \text{for } m \in \text{supp}(\mathcal{M}) \tag{2b}$$

$$\mathcal{P}(m) \in \arg\max_{p \in \Delta\mathcal{A}} \int u_R(p|\theta)\, d\nu(\theta|m). \tag{2c}$$

We denote the set of these equilibria by $\Sigma(u^{\varepsilon V}, u_R)$ if u_R is allowed to vary, and $\Sigma(u^{\varepsilon V})$ when u_R is fixed. The topology on equilibria is inherited from the distribution space $\Delta(\Theta \times M \times \Delta\mathcal{A})$: $\sigma_n \to \sigma$ if (1) the distribution of sender strategies $\int \mathcal{M}(\cdot, \omega)\, d\mathbb{P}_\omega$ converge, and (2) receiver strategies converge when restricted to on-path messages.

Definition 1. A **candidate equilibrium** σ_0 is an equilibrium of the model with transparent sender preference ($\varepsilon = 0$), ie. $\sigma_0 \in \Sigma(u^0, u_R)$.

Upper hemicontinuity ensures that candidate equilibria will be the limit of the equilibria $\sigma_\varepsilon \in \Sigma(u^{\varepsilon V}, u_R)$ in our model as $\varepsilon \to 0$.

We are particularly interested in non-trivial equilibria where the sender successfully persuades the receiver with positive probability. Formally:

Definition 2. An equilibrium σ is **persuasive** if there is positive probability that the receiver chooses an action that is not a best response to their prior belief:

$$\mathbb{P}\left[\mathcal{P}(\mathcal{M}(\theta, \omega)) \in \arg\max_{p \in \Delta\mathcal{A}} \int u_R(p|\theta)\, d\mu(\theta)\right] < 1.$$

We maintain the following assumption throughout the paper:

Assumption (S). *For any pair of distinct pure actions $a \neq a'$ and messages $m \neq m'$, we have $u^0(a, m) \neq u^0(a', m')$.*

Within cheap talk this is merely the assumption that u^0 is injective on the set of pure actions $\mathcal{A}$.

The content of this assumption is that the sender is never indifferent between messages that lead to different pure actions — this will make mixed actions a necessary feature of persuasive equilibria.

[5]For simplicity, we omit the belief over Ω, as sender idiosyncrasy is irrelevant to the receiver

1.1 Candidate Equilibria

We follow the belief-based approach of Lipnowski and Ravid 2020 to characterize candidate equilibria $\sigma_0 \in \Sigma(u^0, u_R)$.

With transparent sender preference u^0, eq. 2a becomes

$$m \in \arg\max_{m \in \Delta M} u^0(\mathcal{P}(m), m) \quad \text{for } m \in \text{supp}(\mathcal{M}). \tag{3}$$

That is every on-path message must result in the same expected sender-utility $\overline{u}$. Assumption (S) implies that if multiple messages attain this maximum, they cannot result in distinct pure actions. Consequently at most one pure action can feature in a candidate equilibrium.

The power of transparent preferences is that this constraint is state-independent. This means that, for fixed $\mathcal{P}$ satisfying eq. 3, the set of equilibrium posteriors $\mathcal{B}_\sigma := \text{supp}(\nu \circ \mathcal{M})$ is only constrained to satisfy

$$\mu \in \text{co}(\mathcal{B}_\sigma), \tag{4a}$$

where $\text{co}(\cdot)$ is the convex hull operation on sets.

The last constraint to check is receiver optimality. In particular, for a posterior belief ν induced by the message m, this constrains the sender-utility to lie in the following set

$$u^*(\nu, m) := \left\{ u^0(p, m); p \in \arg\max_{p' \in \Delta\mathcal{A}} \int u_R(p'|\theta)\, d\nu(\theta) \right\}.$$

The attainability (and incentive compatibility) of the sender-utility $\overline{u}$ as described by eq. 3 may then be rewritten as

$$\overline{u} \in \bigcap_{m \in \text{supp}(\mathcal{M})} u^*(\nu(m), m) \tag{4b}$$

$$\overline{u} \geq \max_{m \in M} \min_{\nu \in \Delta\Theta} \{u^*(\nu, m)\}. \tag{4c}$$

This says that the sender can be made indifferent over on-path messages, while preferring them to off-path messages. Together, eqs. 4abc characterize candidate equilibria.

Two notable transparent communication technologies are cheap talk (analyzed by Lipnowski and Ravid 2020), and money burning (analyzed in Appendix B.1). In the cheap talk case, the sender's preference is message-independent, so eq. 4c is trivial. The equilibrium can then be characterized by a set of beliefs $\mathcal{B}_\sigma \subseteq \Delta\Theta$ and a sender utility $\overline{u}$ satisfying eqs. 4a and 4b. Consequently, persuasive cheap talk equilibria exist whenever $|M| \geq |\Theta|$ and μ lies in the convex hull of posteriors that attain at least utility $\overline{u}$, as illustrated by the following example:

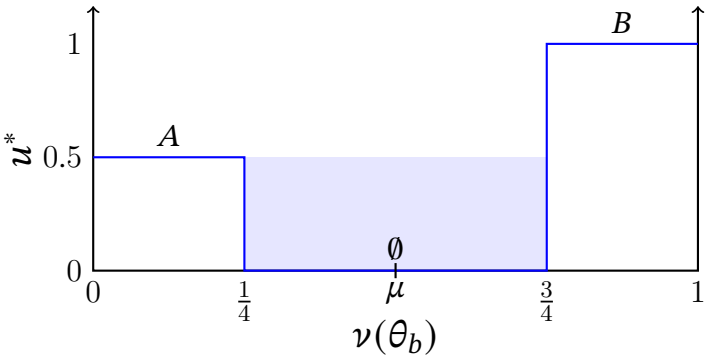


Figure 1: The sender's indirect utility in Example 1. The shaded region is the range of utilities attainable by cheap talk for a given prior.

Example 1 (Binary-state salesperson)**.** Suppose there is a salesperson trying to convince a consumer to make a purchase between two products A and B. There are two possible states of the world: in state $\theta = \theta_a$ product A is good while product B is bad, while in state $\theta = \theta_b$ product B is good while product A is bad. The consumer believes each state is equally likely: their prior belief is $\mu = \frac{1}{2}(\theta_a \oplus \theta_b)$. We represent a belief $\nu \in \Delta\{\theta_a, \theta_b\}$ by the probability $\nu(\theta_b)$ it puts on state θ_b.

The salesperson knows which product is of good quality, but only cares about their commision, which is higher for product B than for product A. Based on the salesperson's recommendation, the consumer can purchase either product, or walk away without making a purchase — we label their action by the product purchased $\mathcal{A} = \{A, B, \emptyset\}$, where $\emptyset$ means no purchase. This leads to the following utility tables for the receiver (consumer) and sender (salesperson) respectively:

$\theta \backslash a$	A	$\emptyset$	B
$u_R(\cdot\|\theta_a)$	1	3/4	0
$u_R(\cdot\|\theta_b)$	0	3/4	1

	A	$\emptyset$	B
u^0	1/2	0	1

The sender's indirect utility for this problem is illustrated in Fig. 1. Notably, full-revelation *ex post* dominates babbling. Attainable sender-utilities for a given prior are given by the shaded region above the prior. The sender can obtain utility $\overline{u}$ only if $\overline{u} \in [0, 1/2]$, since these are the only utilities supported on both sides of the prior belief.

This can be done by inducing the beliefs $\nu_0 = 1/4$ and $\nu_1 = 3/4$, convincing the receiver to choose $\mathcal{P}(\nu_0) = 2\overline{u}A \oplus (1 - 2\overline{u})\emptyset$ and $\mathcal{P}(\nu_1) = \overline{u}B \oplus (1 - \overline{u})\emptyset$ respectively. These posteriors correspond to the sender playing the mixed strategy $\mathcal{M}(\theta_a) = \frac{3}{4}m_0 \oplus \frac{1}{4}m_1$ and $\mathcal{M}(\theta_b) = \frac{1}{4}m_0 \oplus \frac{3}{4}m_1$.

The utility $\overline{u} = 1/2$ can also be obtained by inducing posteriors $\nu_0 \leq 1/4$ and $\nu_1 = 3/4$, convincing the receiver to play $\mathcal{P}(\nu_0) = A$ and $\mathcal{P}(\nu_1) = \frac{1}{2}(B \oplus \emptyset)$ respectively. This corresponds to the sender playing the mixed strategy $\mathcal{M}(\theta_a) = \frac{2(1-\nu_0)}{3-4\nu_0}m_0 \oplus \frac{1-2\nu_0}{3-4\nu_0}m_1$ and $\mathcal{M}(\theta_b) = \frac{2\nu_0}{3-4\nu_0}m_0 \oplus$

$\frac{3-6\nu_0}{3-4\nu_0}m_1$. The case $\nu_0 = 0$ deserves particular attention: it is in some sense the 'simplest' equilibrium, corresponding to partial revelation, and furthermore both receiver- and sender-optimal.

To obtain persuasive equilibria the sender must change their strategy dependent on the state θ which is irrelevant to their preference.

Note that there are a continuum of equilibria in this example. This is true for candidate equilibria more generally, as either u^* is never a singleton on $\mathcal{B}_\sigma$, in which case eq. 4b is an interval, or u^* is a singleton at some posterior belief $\nu_0 \in \mathcal{B}_\sigma$, in which case eq. 4a permits ν_0 to vary over an open set while inducing the same action. One consequence of our work will be to refine this continuum set to a finite set of persuasive equilibria.

We pause to note the role of our finite-ness assumptions. Using a finite action space $\mathcal{A}$ ensures that u^* is well-behaved, obeying certain technical regularity conditions. Our techniques and results generally extend to equilibria of non-pathological models with infinite actions — those that are well-approximated by finite-action models. In contrast, our assumption of finite states is fundamental to our robustness analysis in later sections. While we may be able to approximate infinite-state settings with finite states, our notions of robustness may lose validity.

2 Fragility

Our first question is whether candidate equilibria are robust to slight privacy in the sender's preference. This is important as public knowledge of another's preference is generally implausible.

For this section it suffices to consider state-independent perturbations to the senders utility: $U : \Delta\mathcal{A} \times M \times \Omega \to \mathbb{R}$. Consider the perturbed utility functions

$$u^{\varepsilon U}(\pi|\omega) := u^0(\pi) + \varepsilon V(\pi|\omega).$$

We understand a candidate equilibrium as approximating an environment with idiosyncrasy when the following condition (developed as 'purification' by Harsanyi 1973[6]) is satisfied:

Definition 3 (Harsanyi robustness)**.** An equilibrium $\sigma_0 \in \Sigma(u^0, u_R)$ is **Harsanyi-robust** if for all bounded perturbations V, there exist equilibria $\sigma_\varepsilon \in \Sigma(u^{\varepsilon V}, u_R)$ of the game with sender preference $u^{\varepsilon V}$ such that $\sigma_\varepsilon \to \sigma_0$ in distribution as $\varepsilon \to 0$.

Harsanyi showed that this notion of robustness is a property of Nash equilibria in generic finite games. While these idiosyncrasies change an individual player's preference, other players, ignorant to the realization of these idiosyncrasies, will observe little difference in probabilistic behaviour. This provides an interpretation of mixed strategy equilibria as manifestations of unobserved idiosyncrasy.

[6]Harsanyi's condition is actually stronger, considering perturbations to *all* players' utilities.

For an equilibrium to be Harsanyi-robust, every strategy the sender randomizes over should be a strict best response to some local variation in the receiver's actions (constrained to their equilibrium best responses). This relies on the strategy space not featuring 'invariances' which may prevent best responses from being unique. For example:

1. In cheap talk the sender's utility is invariant to relabelling of messages.
2. With transparent preferences the sender's utility is invariant to 'informativeness'. Fixing an arbitrary receiver strategy $\mathcal{P}$, the sender's utility is equal under the following strategies:
   (a) Informative strategy: send each message m with a state-dependent probability $\mathcal{M}_m(\theta)$.
   (b) Uninformative strategy: send each message m with the state-independent probability $\sum_\theta \mathcal{M}_m(\theta)\mu(\theta)$.

The fragility associated with the first redundancy is easily resolved — either by selecting equilibria without redundant messages, or equivalently by forcing the model to remain in the realm of cheap talk with message-independent idiosyncrasies.

In contrast, the second type of redundancy is endemic to persuasive candidate equilibria:

Theorem 1 (Adaptation of Diehl and Kuzmics 2021[7])**.** *No persuasive equilibria of communication games where the sender and receiver have independent preferences is Harsanyi-robust:*

1. *No persuasive equilibrium exists if the sender's utility can be decomposed*

$$u^{\varepsilon U}(a, m|\omega) = \tilde{u}(a, m|\omega) + \tilde{U}(a|\omega), \tag{5}$$

   with $\tilde{U}$ distributed according to a density on $\mathbb{R}^{\mathcal{A}}$ when conditioned on $\tilde{u} \in \mathbb{R}^{\mathcal{A}\times M}$.

2. *For fixed receiver utility u_R, persuasion is (topologically) generically impossible over the set of preferences $\Delta\mathbb{R}^{\mathcal{A}\times M}$ under the weak-$*$ topology.*

Equation 5 represents preferences where the utility of an action is not completely determined by the message that induces it — in cheap talk this says persuasion is impossible whenever the sender's preference is distributed according to a state-independent density, no matter how concentrated. Furthermore, even if persuasion is possible for some rare utility $\tilde{u}$, this persuasion can be destroyed by adding even a small idiosyncratic perturbation that is independent of $\tilde{u}$.

The intuition behind this result is that a sender who first observes their idiosyncrasy ω will, w.p. 1, have a strict preference over equilibrium messages, and thus will choose a message independent of their observation of θ.

[7]The referenced result establishes the fragility of persuasive equilibrium in the $\Theta = \mathcal{A} = \mathbb{R}^n$ cheap talk model of Chakraborty and Harbaugh 2010. We adapt these techniques to finite pure action models where their Condition (S) may not hold.

These results rely only on the state-independence of $u^{\varepsilon U}$ and the finiteness of $\mathcal{A}$, and can be extended to settings with infinite state spaces and idiosyncratic receivers (ie. where the receiver's preference u_R depends on an idiosyncrasy ω_R distributed independently of the sender's idiosyncrasy ω). In this paper we are primarily interested in finite message equilbiria:

Proof of (1) when M is finite. Suppose there exists a persuasive equilibrium $\sigma \equiv (\mathcal{M}, \nu, \mathcal{P})$, and let p^* be an action that the receiver chooses with positive probability. Let $M^* := \mathcal{P}^{-1}(p^*)$ be the set of messages that induce this action, $\Pi^* := M^* \times \{p^*\}$ be the corresponding set of message-actions, and $M^c := M \setminus M^*$ and $\Pi^c := \{(m, \mathcal{P}(m)); m \in M^c\}$ be the remaining equilibrium message-actions.

The equilibrium probability that a message in M^* is sent in state θ is then bounded between the probabilities that a message in M^* is a strict and weak best response:

$$\mathbb{P}\left[\max_{\pi \in \Pi^*} u^{\varepsilon U}(\pi) > \max_{\pi' \in \Pi^c} u^{\varepsilon U}(\pi')\right] \leq \mathbb{P}\left[\mathcal{M} \in M^* | \theta\right] \leq$$
$$\mathbb{P}\left[\max_{\pi \in \Pi^*} u^{\varepsilon U}(\pi) > \max_{\pi' \in \Pi^c} u^{\varepsilon U}(\pi')\right] + \mathbb{P}\left[\max_{\pi \in \Pi^*} u^{\varepsilon U}(\pi) = \max_{\pi' \in \Pi^c} u^{\varepsilon U}(\pi')\right],$$

where we suppress ω in our notation. The last term is dominated by the probability that the sender is indifferent between any pair $(\pi, \pi') \in \Pi^* \times \Pi^c$:

$$\sum_{(m,m') \in M^* \times M^c} \mathbb{E}\left[\mathbb{P}\left[\tilde{U}(p^*) - \tilde{U}(\mathcal{P}(m')) = \tilde{u}(p^*, m) - \tilde{u}(\mathcal{P}(m'), m') \middle| \tilde{u}\right]\right] = 0.$$

This is zero, because when conditioned on $\tilde{u}$, the difference of $\tilde{U}$ terms is distributed with a density and the difference of $\tilde{u}$ terms is a constant. Thus every equilibrium action is chosen with a positive state-independent probability, and hence must be optimal under the prior belief.[8] □

The result of this section is that persuasion is impossible with state-independent sender preferences if there is *any* fuzziness about what these preferences are. Since perfect knowledge of other agents is impossible in reality, this suggests that to understand communication, we must understand the state-dependence inherent in a setting, for example state-dependent biases in the sender's preference.

3 Communication Graphs

While persuasion is generally impossible when the senders preference is state independent, robust persuasion is possible when preferences are slightly state-dependent. Some candidate equilibria

[8]This proof does not extend to cases where a continuum of messages are sent. In this case, each sender ω may be indifferent between messages π_ω, π'_ω varying with ω, and randomizes between these messages in an informative way. Such messages are each sent with probability zero, but cumulatively are sent with positive probability. This general case is treated in Appendix A.

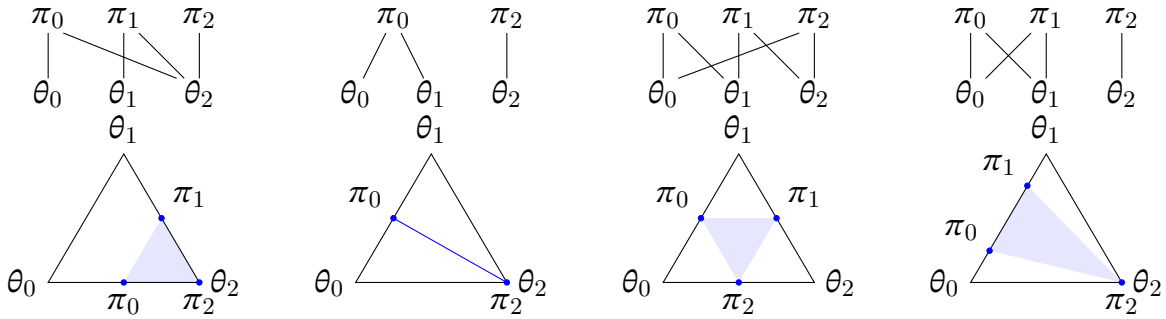


Figure 2: Example communication graphs (top) and corresponding induced beliefs (bottom) in a three-state game. The nodes of the graphs correspond to states $\{\theta_i\}$ and messages $\{\pi_i\}$ inducing the beliefs illustrated below. The line/shaded regions within the simplices correspond to the priors that render the given posteriors plausible — ie. the convex hull of these posteriors. Only (a,b) are acyclic; only (a,c) are connected.

can be approximated by introducing state-dependence, but other's cannot be robustly approximated for any state-dependence. To characterize which equilibria can be approximated, we develop a novel graph-theoretic representation of equilibria.

Formally, given a strategy profile $(\mathcal{M}, \mathcal{P})$, define $\sigma(\theta) := \{(p, m); m \in \text{supp}_{\mathbb{P}_\omega}(\mathcal{M}(\theta)), p \in \mathcal{P}(m)\}$ to be the equilibrium paths induced by the event that the state θ is realized — ie. message-action pairs that occur when the state is θ.

Definition 4. The **communication graph** $G(\sigma)$ associated with an equilibrium σ is the undirected bipartite graph whose nodes are $\Theta \sqcup \sigma(\Theta)$ and whose edges are $\big\{\{\theta, \pi\}; \theta \in \Theta, \pi \in \sigma(\theta)\big\}$.

A communication graph (and the equilibrium it represents) is

- **connected** if it contains a path between any two nodes,
- **acyclic** if it contains at most one path between any two nodes,
- a **tree** if it is both acyclic and connected,
- a **forest** if it is acyclic but not connected.

Communication graphs thus connect states with the messages sent by a sender in that state. Some example communication graphs, and corresponding beliefs, are illustrated in Figure 2.

Note that communication graphs are a coarse representation of strategies, as they do not specify probability with which a message is sent, merely what states the posterior belief designates as

plausible (ie. within its support). Nevertheless it captures sufficient detail for our robustness analysis.[9]

Acyclic equilibria can loosely be thought as a class of 'simple' equilibria, not requiring a complex network of sender indifferences.

Non-connected equilibria are *separating*. For such equilibria there is a partition $\Theta = \Theta_1 \sqcup \Theta_2$ that is communicated with certainty, ie. for all posteriors $\nu \in \mathcal{B}_\sigma$, either $\text{supp}(\nu) \subseteq \Theta_1$ or $\text{supp}(\nu) \subseteq \Theta_2$. In connected equilibria there must be a message that is sent from states in both elements of the partition.

To get a feel for the constraints that are generically imposed on candidate equilibria communication graphs, consider the subset of pure strategy equilibria (ie. $|\sigma(\theta)| = 1$ for all states).[10] These equilibria are acyclic, and, when persuasive they are also non-connected — thus forests. The need for randomization imposed by Assumption (S) means that pure sender strategies are only able to persuade the receiver if the receiver has a very specific preference. In particular:

Observation. *Persuasive, pure-strategy candidate equilibria exist only if there is a partition of states* $\Theta = \Theta_0 \sqcup \cdots \sqcup \Theta_n$ *such that the conditional beliefs*

$$\mu_{\Theta_i} := \mu[\cdot|\Theta_i]$$

induce mixed actions for $i \geq 1$.

For example, if π_2 is a pure action in Figure 2b, then this graph can only describe a persuasive equilibrium if the receiver is indifferent between two actions at the precise belief $\mu_{\{\theta_0,\theta_1\}}$ corresponding to π_0. This requires very specific receiver utilities: if this indifference occurs, a slight perturbation to the receiver's preference u_R will invalidate the indifference at this specific belief, making such an equilibrium impossible.

Fundamentally, this is because pure strategies are only capable of generating a finite number of posterior belief sets $\mathcal{B}_\sigma \subset \Delta\Theta$ which are unlikely to coincide with the (usually) measure-0 set of posteriors that induce randomization. Using dimensionality arguments, we extend this analysis to the class of acyclic equilibria:

Lemma 1. *For generic receiver preferences* $u_R \in \mathbb{R}^{\mathcal{A}\times\Theta}$ *every connected component of an acyclic candidate equilibrium's communication graph*

(i) *includes at least one message which has a unique pure action* a^* *as its best response,*

(ii) *if the component has a unique such message, then*

[9]Moreover, we will see that these additional details are uniquely identified from the equilibrium's communication graph for the acyclic candidate equilibria we will study.

[10]Green and Stokey 2007 provide a notable analysis of pure equilibria in finite-state/-action models.

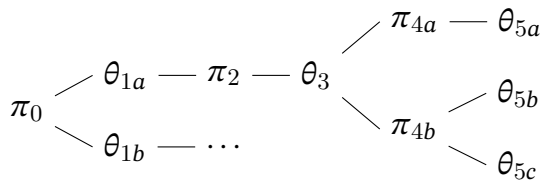


Figure 3: An example tree component of a communication graph. Nodes are indexed alphanumerically, with the numeric component indicating the node's depth, ie. the distance from the 'root' α_0.

(a) *every other message on the connected component has precisely two pure best responses,*

(b) *the connected component corresponds to a unique sender strategy (when restricted to states on this component).*

These generic conditions are the same that are necessary to ensure the equilibrium is robust to perturbations to the *receiver's* preference. The restriction to acyclic equilibria is motivated in the next section, where we find that it is a necessary property for $\mathcal{O}$-robustness. In Appendix A, we show these properties are also obtained for generic prior μ, under mild assumptions on u_R.

Sketch of proof. We outline the proof by contradiction for the specific tree in Figure 3 (where π_0 is an arbitrary action), under the assumption that every action on the tree is mixed — and fixing the actions the receiver is randomizing over for different messages.

We assume that the receiver has a unique best response to every pure state (true for generic u_R). Consequently every terminal node $(\theta_{5a}; \theta_{5b}, \theta_{5c})$ must be a state that sends its parent message $(\pi_{4a}; \pi_{4b})$ with probability 1. The posteriors associated with π_{4a}, π_{4b} are thus constrained to a 1-dimensional subspace of $\Delta\Theta$, parametrized by the probability their parent state (θ_3) sends their message. These subspaces generically intersect the hyperplane of beliefs where the receiver is indifferent between two given actions at most once, and never intersect beliefs where the receiver is indifferent between 3 actions. Thus there is a unique probability with which the sender can send these messages from θ_3, and correspondingly a unique probability with which the sender in state θ_3 sends the message π_2. Holding this probabiliy fixed, there is likewise a unique probability that the sender in state θ_{1a} sends π_2 to induce randomization, and a unique probability θ_{1a} sends π_0. Repeating this process for the branch of the tree below θ_{1b}, we also obtain a unique probability with which θ_{1b} sends π_0. This results in a unique posterior $\nu_0 \in \Delta\{\theta_{1a}, \theta_{1b}\}$ associated with π_0 that is completely determined by the rest of the tree. Since mixed actions are generically a best response to a measure-0 set of posteriors, the best response to ν_0 is generically a pure action, contradicting our assumption.

If π_0 is the unique pure action then this argument demonstrates (ii) as well.

By repeating this argument for all (finitely many) tree geometries and ways to assign pairs of randomization actions to messages, the proof is concluded.

A key property of cheap talk equilibria is that messages that result in the same action can be merged without changing sender incentives. This allows us to assume that there is a unique message resulting in the action a^*: $|\mathcal{P}^{-1}(a^*)| = 1$. This property implies the conditions of (ii), and combined with (i) ensures that acyclic equilibria are generically trees.[11]

Definition 5. A candidate equilibrium is **injective** if on-path pure actions are induced by a unique message, ie. $|\mathcal{P}^{-1}(a)| = 1$ for any $a \in \mathcal{A} \cap \mathcal{P}(\text{supp}(\mathcal{M}))$.

By assumption (S) there is at most one on-path pure action in an equilibrium. An additional property of injective equilibria is that we can assume off-path messages are not a temptation — since off-path messages do not result in a, we can assume they result in a strictly worse pure action.

Injectivity can be assumed whenever the perturbation is message-independent (ie. cheap talk and money burning) in which case merging messages is WLOG. However, in signalling models it is important to keep messages distinct; candidate equilibrium injectivity can still be assured if u^0 is injective on $\mathcal{A} \times M$. In Appendix B.4 we extend our analysis to non-injective candidate equilibria.

In the injective case we obtain the following corollary to Lemma 1:

Theorem 2 (The generic case)**.** *For generic receiver preferences $u_R \in \mathbb{R}^{\mathcal{A}\times\Theta}$, the set of injective acyclic candidate equilibria (as a subset of $\Sigma(u^0, u_R)$) satisfies the following properties:*

- ***(a)*** *all such candidate equilibria are connected (ie. trees),*
- ***(b)*** *the receiver's strategy uses precisely one pure action, with their other actions being binary (ie. p with $|supp(p)| = 2$),*
- ***(c)*** *there are finitely many such equilibria (up to information equivalence[12]), each uniquely identified by its communication graph.*

This says that tree equilibria are generically the simplest injective candidate equilibria: (a) says that it is impossible to obtain communication with simpler indifference constraints (ie. forests communication graphs, which have fewer edges), (b) says that these equilibria involve the minimum amount of receiver randomization, (c) says that these equilibria are precisely described by the communication graph, without requiring additional data.

In future sections, we will frequently constrain ourselves to the generic domain where these properties hold.

[11] This turns out to be the only property we need to directly apply our results to signalling equilibria:

[12] Two equilibria are informationally equivalent if they induce the same distribution of posterior beliefs in the receiver.

4 $\mathcal{O}$-Robustness

In previous sections we have argued that (1) state-dependence is essential for understanding communication in games, and (2) we can limit our (initial) analysis to injective candidate equilibria that are either cyclic and/or connected.

In this section we introduce a notion of $\mathcal{O}$-robustness, capturing the idea of an equilibrium being robustly approximated for a given state-dependent perturbation, without considering idiosyncrasy. We then establish three main results. Firstly, cyclic equilibria are never $\mathcal{O}$-robust, this means that the only equilibria that can be robustly approximated are generically both acyclic and connected, hence not separating. Secondly we will characterize precisely types of state-dependence that permit $\mathcal{O}$-robustness. Our most general result then extends the notion of $\mathcal{O}$-robustness to idiosyncratic perturbations, showing that our results extend to settings where the sender's perturbation is private.

We start by studying non-idiosyncratic perturbations to the sender's preference (ie. modifications $v : \mathcal{A} \times M \times \Theta \to \mathbb{R}$, extended to mixed actions via von Neumann–Morgenstern):

$$u^{\varepsilon v}(\pi|\theta) = u^0(\pi) + \varepsilon v(\pi|\theta).$$

Such state-dependencies are often naturally occuring: for examples, senders may be lying averse, sympathetic to the receiver, or face different resource constraints in different settings, skewing the sender's preference in a state-dependent direction. When these types of systematic biases occur, Harsanyi robustness may no longer be the appropriate notion of robustness, instead we propose the following:

Definition 6. For a nonempty open set of modifications $\mathcal{O} \subseteq \mathbb{R}^{(\mathcal{A}\times M)\times\Theta}$, we say a candidate equilibrium $\sigma_0 \in \Sigma(u^0)$ is **$\mathcal{O}$-robust** if for any modification $v \in \mathcal{O}$ there exist equilibria $\sigma_\varepsilon \in \Sigma(u^{\varepsilon v})$ of the game with sender preferences $u^{\varepsilon v}$ such that $\sigma_\varepsilon \to \sigma_0$ as $\varepsilon \to 0$.

This definition parallels Harsanyi robustness[13], while weakening the condition in two distinct ways: (1) we consider only non-idiosyncratic perturbations, and (2) we allow ourselves to constrain these perturbations to $\mathcal{O}$. To avoid knife-edge scenarios, we only consider open constraint sets $\mathcal{O}$. In Section 4.1 we consider a version of $\mathcal{O}$-robustness that allows for idiosyncrasy and obtain parallel results, showing relaxation (1) is merely simplifying while (2) is essential. The importance of (2) is also demonstrated by the following proposition:

Proposition 1. *Let the sender's preference be $u^0 - \varepsilon u_R$ where $\varepsilon > 0$, u_R is the receiver's preference, and u^0 is transparent. In any equilibrium all on-path actions are best responses to the prior μ.*

[13] $\mathcal{O}$-robustness is also closely related to the notion of 'essential equilibria' developed by Wu and Jiang 1962, stipulating that the equilibrium correspondence Σ is lower hemicontinuous at σ_0 in the space of non-idiosyncratic preferences.

This shows that meaningful communication is impossible if parties' interests are even slightly opposed[14], and persuasive candidate equilibria are never $\mathcal{O}$-robust for any neighbourhood $\mathcal{O} \ni -u_R$.

To establish intuition for how a modification affects an equilibrium, consider the following first order approximation of the sender's utility as the receiver shifts their action after the message π_i by an amount Δp_i:

$$u^{\varepsilon v}(\pi_i + \Delta p_i|\theta) = u^0(\pi_i) + \varepsilon v(\pi_i|\theta) + (u^0) \cdot \Delta p_i + O(\varepsilon \Delta p_i) \tag{6}$$

where $(u^0) \in \mathbb{R}^{\mathcal{A}\times M}$ is a vector of pure outcome utilities and $\Delta p_i \in \mathbb{R}^{\mathcal{A}\times M}$ captures a change in receiver's action after the message π_i[15] — what is important is that the effect Δp_i has on utility is independent of the state θ (to the first order).

For a candidate equilibrium $\sigma_0 \in \Sigma(u^0)$, the zero-th order term $u^0(\pi_i)$ is constant across $\pi_i \in \sigma_0(\Theta)$; the preference over equilibrium messages is determined by higher order terms. Suppose (π_1, θ_2, π_2) is a path on the communication graph $G(\sigma_0)$. To maintain the sender's indifference between π_1, π_2 in state θ_2, we then must have

$$v(\pi_1|\theta_2) - v(\pi_2|\theta_2) = (u^0) \cdot \left(\tfrac{\Delta p_2 - \Delta p_1}{\varepsilon}\right) + O(\Delta p).$$

If $(\pi_1, \theta_2, \ldots, \theta_N, \pi_N)$ is a path on $G(\sigma_0)$, we sum the corresponding constraints to get

$$\sum_{i=2}^{N} v(\pi_{i-1}|\theta_i) - v(\pi_i|\theta_i) = (u^0) \cdot \left(\tfrac{\Delta p_N - \Delta p_1}{\varepsilon}\right) + O(\Delta p). \tag{7}$$

But the righthand side also appears in the difference of utilities between the messages π_1 and π_N. Consequently, if $\pi_1 \in \sigma_0(\theta_1)$, we must have

$$v(\pi_1|\theta_1) - v(\pi_N|\theta_1) \geq (u^0) \cdot \left(\tfrac{\Delta p_N - \Delta p_1}{\varepsilon}\right) + O(\Delta p) = \sum_{i=2}^{N} v(\pi_{i-1}|\theta_i) - v(\pi_i|\theta_i) + O(\Delta p). \tag{8}$$

This inequality must be strict to hold over a neighbourhood of v, as $\mathcal{O}$-robustness requires.

Simplifying this inequality inspires the following definition:

Definition 7 (G-Graph Monotonicity[16])**.** For a communication graph G, we say v is ***G*-graph**

[14]It is important that the baseline sender preference is transparent. In Example 3 we study a game where adding slight antipathy to a state-*dependent* preference actually *increases* communication possibilities.

[15]Formally, if $\pi_i = (p_i, m_i)$ then $(\Delta p_i)_{a,m'}$ is 0 when $m' \neq m_i$, and otherwise is the difference in weight the receiver puts on the pure action a after the message m_i relative to p_i.

[16]This has a strong connection to the notion of 'cyclic monotonicity' developed by Rochet 1987 for mechanism design with linear transfers. This can be understood through the first-order analogy

$$\begin{pmatrix} \sigma_0(\theta) \\ (u^0) \cdot \Delta p_i \end{pmatrix} \Longleftrightarrow \begin{pmatrix} \text{decision rule} \\ \text{transfer mechanism} \end{pmatrix}.$$

See Appendix B.3 for a discussion of this relationship, along with a path-minimizing interpretation of monotonicity.

monotone (or $v \in \mathscr{M}_G(G)$) if for any path $(\theta_1, \pi_1, \ldots, \theta_N, \pi_N =: \pi_0)$ on G with $N \geq 2$ we have

$$\sum_{i=1}^{N} v(\pi_i|\theta_i) - v(\pi_{i-1}|\theta_i) > 0. \tag{9}$$

A simple result characterizes existence:

Proposition 2. *$\mathscr{M}_G(G) = \emptyset$ iff G contains a cycle.*

The 'if' case is deduced by observing that eq. 9 for a path around a cycle (ie. $\pi_N \in \sigma(\theta_1)$) changes sign if we flip the orientation of the path. Conversely, if G is acyclic, we may construct a modification in $\mathscr{M}_G(G)$ by making the $v(\pi_N|\theta_1)$ term sufficiently negative on each path — due to acyclicity this term appears *only* in the inequality associated with the unique path from θ_1 to π_N.

An immediate consequence of this proposition (by Theorem 2) is that $G(\sigma_0)$-graph monotonicity for a candidate equilibrium σ_0 generically implies $G(\sigma_0)$ is a tree.

Our main result is that graph-monotonicity precisely characterizes $\mathcal{O}$-robustness:

Theorem 3. *Within the generic case established by Theorem 2, an injective candidate equilibrium $\sigma_0 \in \Sigma(u^0, u_R)$ is $\mathcal{O}$-robust iff $\mathcal{O} \subseteq \mathscr{M}_G(G(\sigma_0))$.*

Proof. **Necessity:** Suppose $\mathcal{O} \nsubseteq \mathscr{M}_G(G(\sigma_0))$. Since $\mathcal{O}$ is open, it contains a modification $\tilde{v}$ that is not in the closure of $\mathscr{M}_G(G(\sigma_0))$ — ie. for some path $(\theta_1, \pi_1, \ldots, \theta_N, \pi_N)$ on $G(\sigma_0)$ we have

$$0 > \sum_{i=1}^{N} \tilde{v}_i(\pi_i) - \tilde{v}_i(\pi_{i-1}) \tag{10}$$

where $\pi_0 := \pi_N$ and we adopt the shorthand $\tilde{v}_i(\pi) := \tilde{v}(\pi|\theta_i)$. Since $\tilde{v}$ is continuous in π, this inequality holds over a neighbourhood N_π of π.

Now suppose our equilibrium can be approximated with actions $\hat{\pi} \in N_\pi$. For the message $\hat{\pi}_i$ to be optimal for the sender in state θ_i, we must have

$$u^0(\hat{\pi}_i) - u^0(\hat{\pi}_{i-1}) + \varepsilon\,(\tilde{v}_i(\hat{\pi}_i) - \tilde{v}_i(\hat{\pi}_{i-1})) \geq 0 \qquad \text{for } i \in \{1, \ldots, N\}.$$

Summing these equations, we obtain a contradiction of eq. 10.

Sufficiency: Injectivity allows us to assume off-path messages are strictly inferior to the sender. Thus we can constrain our analysis to on-path actions.

From Proposition 2 and Theorem 2, it suffices to check tree graphs containing a single pure action α_0. Fix the vertex associated with the pure action as the root of the $G(\sigma_0)$ (in Figure 3, this means π_0 is pure and denoted α_0).

We adopt the following notation: for a node k on a rooted tree $(G(\sigma_0), \alpha_0)$:

- $k^{\downarrow}$ is the set of k's children (ie. its neighbours that are further from the root)
- $k^{\uparrow}$ is the parent node of k (ie. its unique neighbour closer to the root)

This notation will often be pushed to subscripts.

Let θ_k be a state and fix its parent action $\hat{\pi}_{k^{\uparrow}}$, for its child actions π_j let $I_j \subseteq \text{supp}(\pi_j)$ be intervals such that

$$\min_{\pi'_j \in I_j} u^{\varepsilon v}(\pi'_j|\theta_k) \leq u^{\varepsilon v}(\hat{\pi}_{k^{\uparrow}}|\theta_k) \leq \max_{\pi'_j \in I_j} u^{\varepsilon v}(\pi'_j|\theta_k) \qquad \text{for all } j \in k^{\downarrow}. \tag{11}$$

Then by the Intermediate Value Theorem, for all $j \in k^{\downarrow}$ there exists $\hat{\pi}_j \in I_j$ such that the sender in state θ_k is indifferent between $\hat{\pi}_{k^{\uparrow}}$ and $\hat{\pi}_j$. As $\varepsilon \to 0$ and $\hat{\pi}_{k^{\uparrow}} \to \pi_{k^{\uparrow}}$, these inequalities can be satisfied with $I_j \to \{\pi_j\}$ and thus $\hat{\pi}_j \to \pi_j$. Applying this inductively down the tree gives us $\hat{\pi} \to \pi$.

This describes a potential equilibrium $\hat{\sigma}_\varepsilon \to \sigma_0$ (where in fact the sender employs the same strategy in $\hat{\sigma}_\varepsilon$ and σ_0). To show this is an equilibrium, it remains to show that these neighbouring indifferences generate preferences for neighbouring actions as $\varepsilon \to 0$. Suppose ε is sufficiently small that the actions $\hat{\pi}$ are within a neighbourhood $N_\pi \ni \pi$ where eq. 9 holds (ie. $v \in \mathscr{M}_G(G(\hat{\sigma}_\varepsilon))$), then for any path $(\theta_1, \hat{\pi}_1, \ldots, \hat{\pi}_N)$ on $G(\hat{\sigma}_\varepsilon)$ the sender in state θ_1 strictly prefers its neighbour $\hat{\pi}_1$ to $\hat{\pi}_N$:

$$\begin{aligned} u^0(\hat{\pi}_1) - u^0(\hat{\pi}_N) + \varepsilon\,(v_1(\hat{\pi}_1) - v_1(\hat{\pi}_N)) &> \sum_{i=2}^{N} u^0(\hat{\pi}_{i-1}) - u^0(\hat{\pi}_i) + \varepsilon\,(v_i(\hat{\pi}_{i-1}) - v_i(\hat{\pi}_i)) \\ &= \sum_{i=2}^{N} u^{\varepsilon v}(\hat{\pi}_{i-1}|\theta_i) - u^{\varepsilon v}(\hat{\pi}_i|\theta_i) = 0, \end{aligned} \tag{12}$$

where the last equality follows from $\hat{\pi}$ being defined to generate this indifference. □

Proposition 2 and Theorem 2 provide two immediate corollaries to this result:

Corollary 1. *Within the generic case established by Theorem 2*

1. *An injective candidate equilibrium $\sigma_0 \in \Sigma(u^0, u_R)$ is $\mathcal{O}$-robust for some non-empty set of modifications $\mathcal{O}$ iff its communication graph $G(\sigma_0)$ is acyclic.*
2. *There are finitely many candidate equilibria (up to information equivalence) that are $\mathcal{O}$-robust for some non-empty set of modifications $\mathcal{O}$.*

The second statement applies to non-injective equilibria as well, where graph-monotonicity is not sufficient for $\mathcal{O}$-robustness. We apply our techniques to this case in Appendix B.4.

4.1 Harsanyi $\mathcal{O}$-Robustness

To establish that the previous result is robust to slight idiosyncrasy, consider a sender preference with a state-dependent idiosyncratic perturbation V:

$$u^{\varepsilon V}(\pi|\theta,\omega) = u^0(\pi) + \varepsilon V(\pi|\theta,\omega).$$

It is without loss of generality to assume that $(V(\cdot|\theta))_\theta$ are mutually independent.[17] We extend the definition of $\mathcal{O}$-robustness to idiosyncratic perturbations:

Definition 8. For a nonempty open set of modifications $\mathcal{O} \subseteq \mathbb{R}^{(\mathcal{A}\times M)\times\Theta}$, we say a candidate equilibrium $\sigma_0 \in \Sigma(u^0)$ is **Harsanyi $\mathcal{O}$-robust** if, for any perturbation V with $(V(\cdot|\theta))_\theta$ mutually independent and compactly supported on $\mathcal{O}$, there exists equilibria $\sigma_\varepsilon \in \Sigma(u^{\varepsilon V})$ to the game with sender preference $u^{\varepsilon V}$ such that $\sigma_\varepsilon \to \sigma_0$ as $\varepsilon \to 0$.

Note that Harsanyi robustness is equivalent to Harsanyi $\mathbb{R}^{(\mathcal{A}\times M)\times\Theta}$-robustness. Moreover, Harsanyi $\mathcal{O}$-robustness is a stronger condition than $\mathcal{O}$-robustness. Nevertheless, our result ends up closely mirroring Theorem 3:

Theorem 4. *For an injective candidate equilibrium $\sigma_0 \in \Sigma(u^0)$, in the generic case established by Theorem 2:*

1. *σ_0 is Harsanyi $\mathcal{O}$-robust iff $\mathcal{O} \subseteq \mathscr{M}_G(G(\sigma_0))$.*
2. *The approximating equilibria in Theorem 3 are Harsanyi robust[18]: if $v \in \mathscr{M}_G(G(\sigma_0))$ and $V_n \to v$ uniformly, there exist $N, \bar{\varepsilon} > 0$, equilibria $\sigma_{\varepsilon V_n} \in \Sigma(u^{\varepsilon V_n})$ and $\sigma_{\varepsilon v} \in \Sigma(u^{\varepsilon v})$ such that*

$$\begin{aligned} &(a) \qquad \sigma_{\varepsilon V_n} \to \sigma_{\varepsilon v} \qquad \text{as } n \to \infty, \text{ for } \varepsilon < \bar{\varepsilon}, \text{ and} \\ &(b) \qquad \sigma_{\varepsilon V_n}, \sigma_{\varepsilon v} \to \sigma_0 \qquad \text{as } \varepsilon \to 0, \text{ for } n > N. \end{aligned}$$

The first statement in the theorem says that a slight idiosyncratic perturbation within $\mathscr{M}_G(G(\sigma_0))$ will allow us to approximate the candidate equilibrium σ_0.

The proof has a similar structure to the proof of Theorem 3. The novel difficulty is obtaining the equilibrium actions $\hat{\pi}$, as it is no longer possible to choose actions that make the sender indifferent as in eq. 11. We replace this step with a multi-dimensional analog of the Intermediate Value Theorem described in Appendix A.

[17] By redefining the idiosycrasy space as the product probability space $\tilde{\Omega} := \Omega^\Theta$, we can define a preference $\tilde{V}(\cdot|\theta, \tilde{\omega} = (\omega_{\theta_1}, \ldots, \omega_{\theta_N})) := V(\cdot|\theta, \omega_\theta)$ that satisfies this independence and is identically distributed to $V(\cdot|\theta)$. In this case $\tilde{\omega}$ can no longer be interpreted as corresponding to an individual whose preference may be correlated across states. In Appendix A we show how Harsanyi $\mathcal{O}$-robustness translates to correlated perturbations.

[18] Other candidate equilibria may also have Harsanyi-robust approximating equilibria. Such cases, and their complications, are discussed in Appendix B.2 in the context of an example.

5 Applications

In this section, we present three applications of our results to specific modifications. Two of the modifications we consider represent moral considerations of a sender (empathy and lying aversion), and one represents a weak technical constraint to the message space, resembling a disclosure game with the ability to cheaply fabricate information. We also show how money burning can significantly benefit a sender with slightly state-dependent preferences, despite not benefitting a sender with transparent preferences.

Moral modifications seem intuitively likely to favour communication. Indeed they generally make full revelation strategies more attractive. However, candidate equilibria differ from revelation in generically requiring some obfuscation of the state (as seen in Theorem 2), and these obfuscations can conflict with senders' moral concerns.

For a modification $v \in \mathbb{R}^{(\mathcal{A}\times M)\times\Theta}$, we say the candidate equilibrium σ_0 is v**-robust** if σ_0 is $\mathcal{O}$-robust for some neighbourhood $\mathcal{O} \ni v$.

5.1 Empathetic Senders

When faced with a decision that has little effect on their material utility, even the most egoistic sender may consider the effect their actions will have on others as a deciding factor. We model this as empathy (perturbation $V \equiv u_R$), effectively aligning the interests of the two parties.

A common rule-of-thumb in communication games is that preference alignment increases the possibilities of communication: a sender with a stronger interest in the receiver will seek to ensure the receiver makes more accurate decisions, and thus better inform them.

The previous sections illustrate that this intuition is validated when the geometry of the communication structure (represented by $G(\sigma_0)$) is compatible with the monotonicity of the receiver's preference. We illustrate in the case where this geometry is linear. In particular, let $\succ$ be a linear order on the state space Θ:

Definition 9. We say the receiver utility u_R is $\succ$**-single crossing** if

$$\theta \mapsto u_R(a|\theta) - u_R(a'|\theta)$$

is $\succ$-strictly monotone for any distinct $a, a' \in \mathcal{A}$.

A candidate equilibrium $\sigma_0 \in \Sigma(u^0)$ is an $\succ$**-interval** equilibrium if the sender's strategy is $\succ$-ordered: for $\pi, \pi' \in \sigma_0(\Theta)$ the states that send one message are weakly higher than the states that send the other, ie.

$$\sigma_0^{-1}(\pi) \preceq \sigma_0^{-1}(\pi') \qquad \text{or} \qquad \sigma_0^{-1}(\pi) \succeq \sigma_0^{-1}(\pi').$$

The single crossing condition induces a linear order $\succ_{\mathcal{A}}$ on the action space $\mathcal{A}$ such that $\succ_{\mathcal{A}}$-higher actions have a comparative advantage in $\succ$-high states. Interval equilibria describe situations

where there are $\succ$-thresholds between messages: only lower states send one message, and only higher states send the other. Such equilibria are necessarily acyclic.

Combining these structures ensures that the equilibrium will be u_R-robust:

Proposition 3. *Let u^0 be a transparent sender preference, and u_R be a $\succ$-single crossing receiver preference in the generic class established by Theorem 2. Any $\succ$-interval candidate equilibrium $\sigma_0 \in \Sigma(u^0, u_R)$ is then u_R-robust.*

If further (1) u_R is such that the receiver is only ever indifferent between $\succ_{\mathcal{A}}$-neighbouring actions, and (2) u^0 is a $\succ_{\mathcal{A}}$-single-dipped preference, then the sender-preferred candidate equilibrium is the unique (non-babbling) candidate equilibrium that is u_R-robust.

We term the latter set of models as having *quantitative* actions. These are settings where as their posterior changes, the receiver's best-response shifts incrementally: they are never indifferent between two non-neighbouring actions. Single-dipped preferences describe senders that benefit from the receiver taking $\succ_{\mathcal{A}}$-extreme actions, which correspond to $\succ$-extreme beliefs. This is a natural setting for communication, as the sender benefits most from splitting the receiver's prior into one low and one high posterior.

1-dimensional quadratic preference models are a classic example of quantitative actions, as illustrated in the following example studied by Lipnowski and Ravid 2020:

Example 2 (Advice on Investing in an Asset)**.** An investor consults a broker about what share of their wealth to invest in an asset. The action space is thus a proportion of their wealth $\mathcal{A} = [0, 1]$,[19] with the investor holding an initial position $a_0 \in [0, 1]$. The broker is aware of some information $\theta \in \Theta$ (drawn from the prior μ with finite support), indicating the investor's optimal position is actually $a^*(\theta) \in [0, 1]$, and receives a fee proportional to the investor's trade volume paid by the investor. The investor (receiver) and broker (sender) material preferences are thus

$$u_R(a|\theta) = -\tfrac{1}{2}(a - a^*(\theta))^2 - \kappa|a - a_0| \qquad u^0(a) = |a - a_0|$$

We assume that the investor's initial position is optimal under their prior knowledge, $a_0 = \mathbb{E}_\mu [a^*(\theta)]$. Note that the investor's preference satisfies single crossing when the states are ordered by their optimal positions $a^*(\theta)$. Furthermore the investor's best response shifts continuously along this order as their posterior varies, matching the description of quantitative actions.[20]

Candidate equilibria in this setting consist of two recommendations: buy an additional amount δ of the assset, or sell that same amount. There are a continuum of candidate equilibria corresponding to different broker profits δ.

[19] This action space is continuous. Because the utilities (in particular u^*) are well behaved our results still apply. The skeptical reader may assume a finite approximation of this space.

[20] The notion of quantitative actions can be formally extended to continuous action models through the more general condition: for any two beliefs $\nu_1 \succ_{\text{FOSD}} \nu_0$ the image of the receiver's best-response over any $\succ_{\text{FOSD}}$-path $\gamma \subset \Delta\Theta$ between ν_0, ν_1 does not depend on the path. For example, in the $\kappa = 0$ case, this image is always $[\mathbb{E}_{\nu_0} [a^*(\theta)], \mathbb{E}_{\nu_1} [a^*(\theta)]]$.

A tree candidate equilibrium in this setting has only one state θ^* randomizing between the two messages, with every other state sending only one message. For this candidate equilibrium to have an interval structure, every state above θ^* must recommend buying, while every state below θ^* recommends selling. There is a unique candidate equilibrium with this structure — the only non-babbling candidate equilibrium that is u_R-robust— which is also the candidate equilibrium that maximizes the broker's profit δ.

With two states Proposition 3 shows that every tree candidate equilibrium is u_R-robust. However there are simple three state models where empathy is not $G(\sigma_0)$-graph monotone for any persuasive candidate equilibrium $\sigma_0 \in \Sigma(u^0)$. This changes as we move to settings with '*vertically-differentiated*' actions, for example a salesperson selling nested bundles of products:

Example 3 (Selling Nested Products)**.** An insurance salesperson advises a consumer about which policy to buy. There are two policies, an expensive full insurance policy F that covers a wide range of outcomes, and a cheaper partial insurance policy P that covers a few specific risks. The consumer possesses the outside option $\emptyset$, representing some minimal insurance plan. The optimal policy for the consumer depends on their type, the salesperson knows the optimal policy (due to specialized knowledge of the policies, and a proprietary algorithm on the consumer's demographics), but the consumer has a prior μ.

Specifically, the consumer may be type θ_n in which case the outside option of minimal coverage is optimal; if the consumer's type is θ_p then their specific risks are covered by both policies, and the partial policy is optimal due to its cheaper cost; if the consumer's type is θ_f their risks are not covered by the partial policy, and the full insurance is optimal.

The salesperson is evaluated by the number of policies they sell, with an additional bonus for each full coverage policy sold.

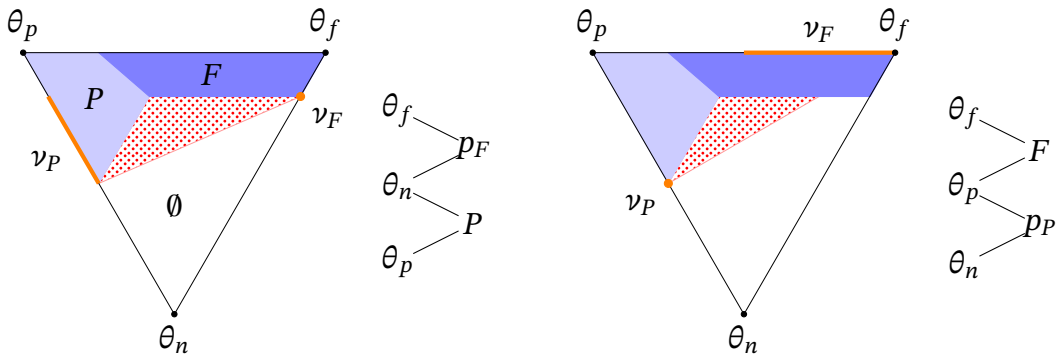
Specifically, the consumer (receiver) and salesperson (sender) to have the following material preferences, normalizing the value of the outside option to 5:

$\theta \backslash a$	$\emptyset$	P	F
$u_R(\cdot\|\theta_n)$	5	3	0
$u_R(\cdot\|\theta_p)$	5	7	6
$u_R(\cdot\|\theta_f)$	5	3	6

	$\emptyset$	P	F
u^0	0	3	4

These preferences lead to the indirect utility illustrated in Figure 4a.

With cheap talk, persuasive candidate equilibria involve the receiver choosing the partial policy P after one message, and randomizing between the high coverage policy and the outside option after the other message $p_F = \frac{3}{4}F \oplus \frac{1}{4}\emptyset$. This can persuade a receiver who would otherwise choose the outside option if the receiver's prior is in the hashed region of the figure. The unique acyclic candidate equilibrium for such priors involves the sender pooling some θ_n-consumers with θ_p-consumers to convince them to purchase the specific policy, and to pool the remaining θ_n-consumers with θ_f-consumers to rationalize the randomization action.



(a) (b)

Figure 4: The sender's indirect utility over the receiver's belief simplex in Example 3, with darker hues indicating higher utility. (a) demonstrates the unique acyclic cheap talk equilibrium, which is not u_R-robust. (b) demonstrates the money-burning equilibrium that is u_R-robust. These correspond to the illustrated communication graphs, and induce posteriors ν in the orange boundary regions. These candidate equilibria benefit the sender when the prior μ is in the cross-hatched region.

While this candidate equilibrium superficially resembles the acyclic equilibrium of Example 1, if we look deeper we see that this candidate equilibrium is somewhat counterintuitive from the consumer's perspective. According to this candidate equilibrium, θ_n-consumers often end up with their worst insurance policy F. Given these equilibrium actions, it seems more natural to pool the consumers that benefit from the full insurance policy together to choose policy p_F, ie. θ_p-with θ_f-consumers as in Figure 4b. However this pooling can never rationalize the mixed action p_F, indeed the action arising from this pooling will weakly dominate the action chosen from the complementary θ_p-θ_n pooling, violating incentive constraints.

Applying Theorem 3, we find that this counter-intuitive nature is precisely why this candidate equilibrium is not u_R-robust, in particular graph-monotonicity fails for the path $(\theta_p, P, \theta_n, p_F)$:

Proposition 4 (Uncooperative empathy)**.** *For any prior μ, no candidate equilibrium σ_0 of Example 3 is u_R-robust. If the prior is in the cross-hatched region there are persuasive candidate equilibria $\sigma_0 \in \Sigma(u^0)$ that are $\mathcal{O}$-robust for an open set $\mathcal{O}$ of modifications that excludes empathy.*

Note that this is despite the candidate equilibrium being a pareto improvement.

Empathy can even be harmful to communication in this setting. Suppose the prior belief is in the cross-hatched region of Figure 4a, and consider an acyclic persuasive candidate equilibrium σ_0 with a modification $v \in \mathscr{M}_G(G(\sigma_0))$ that permits persuasion. Adding a sufficient amount of empathy will shift the modification outside of the set $\mathscr{M}_G(G(\sigma_0))$, removing the possibility of persuasion. In this way, increasing preference alignment decreases the receiver's information. Conversely, making the sender more antipathic undoes this alignment and increases communication possibilities.

If we allow the sender to burn money there exist candidate equilibria that are u_R robust, as we discuss below.

The numbers for this example are chosen for simplicity and plausibility. We enumerate the general properties that permit this result:

(1) policy P is never a best response on $\Delta\{\theta_n, \theta_f\}$.

(2) policy F provides value over $\emptyset$ in state θ_f and θ_p due to its broad coverage.

(3) policy P provides value over policy F in state θ_p and θ_n due to its cheaper cost.

(4) $u^0(P)$ is 'close' to $u^0(F)$.

Note that while u_R in our example does not obey single-crossing (for any ordering of states) it is possible to obtain these properties with a receiver utility that satisfies single-crossing relative to the order $\theta_n \prec \theta_p \prec \theta_f$. (1) is thus the essential property distinguishing qualitative actions from models with quantitative actions where Proposition 3 applies.

Property (1) ensures that a persuasive candidate equilibrium exists, as the receiver can be made indifferent between $\emptyset$ and F. (2-3) ensure that u_R will fail to be graph monotone when the mixed action p_F is close to F, as implied by (4).

5.2 Money-Burning

Money-burning is well-known technique in communication games to relax incentive constraints — previously studied in cheap talk settings by Austen-Smith and Banks 2000; Kartik 2007. A sender may profit significantly from the receiver taking an action a, but this profit prevents the sender from credibly recommending a. If the sender publicly burns a portion of their future profits, the incentive to mislead the receiver to take this action is reduced, potentially restoring credibility to the recommendation.

One question is whether this process benefits the sender, or if it requires the sender to burn all of their profit, providing no benefit over cheap talk equilibria. The latter occurs when preferences are transparent: while money burning greatly expands the set of candidate equilibria in such settings, none of these equilibria deliver a higher utility to the sender than cheap talk.[21] This is surprising, as many interpretations of money-burning (e.g. expensive advertising campaigns, wining and dining potential clients, lobbyists expending effort to build relationships with lawmakers) focus on settings where preferences are fairly transparent.

Moving to *nearly* transparent preferences, money burning can be advantageous for the sender through two distinct mechanisms.

[21] See Appendix B.1 for details.

Firstly, only enough money needs to be burnt to dissuade defection. Since states that send a message have a relative bias towards that message, they retain some excess profit. However, since this relative bias is a product of state-dependence, the increase in utility will only be proportional to the degree of state-dependence ε.

A larger benefit occurs when cheap talk candidate equilibria require garbling posteriors in ways that are incompatible with the state-dependent sender preference, while money-burning equilibria persist by permitting different communication graphs. We see this in Example 3 where empathy is incompatible with cheap talk candidate equilibria:

Example 3 (continued). If the sender is able to accompany a message with the burning money in discrete increments, they are able to create a new type of candidate equilibrium that involve the receiver choosing the action F after the sender accompanies the message with burning a value $M_b \in]1,3[$ of wealth, and choosing the mixed action $p_P := \frac{3-M_b}{2}P \oplus \frac{M_b-1}{2}\emptyset$ after observing a message where the sender does not burn any wealth.

Such candidate equilibria exist for the same range of prior beliefs as the cheap talk equilibrium exists (ie. priors in the cross-hatched region of Figure 4a), but for prior beliefs in the cross-hatched region of Figure 4b it enables communication graphs with a new geometry. In particular, this allows an equilibrium where θ_p-consumers are pooled with each of the two other types, matching our earlier argued intuition. This candidate equilibrium is u_R-robust.

This allows the sender to obtain a material utility of $3 - M_B \in]0,2[$. When the increments of money are small, the sender-preferred equilibrium gives them a utility of almost 2, where without money-burning they would be unable to persuade the receiver and be left with a utility of 0. The consumer also obtains a positive *ex ante* benefit of $\mu[\theta_f] + \mu[\theta_p] - \mu[\theta_n]$ relative to babbling[22].

Note that this equilibrium only exists for a strict subset of prior beliefs for which persuasive candidate equilibria exist. For priors outside of this cross-hatched region persuasive candidate equilibria fail to be u_R-robust.

The advantage of money burning in this setting relies on the properties (2-4) discussed above, and a stronger version of (1):

(1*) $\emptyset$ provides value over policy P in states θ_n and θ_f.

This property ensures that the cross-hatched region of Figure 4b is non-empty. Along with (2-3), this precludes u_R from satisfying a single-crossing condition.

5.3 Weak Signalling Models

Weak signalling models involve modifications that are message independent. We consider two examples: lying aversion, which we will find is only capable of preserving specific geometries

[22]The cross-hatched region of priors where this equilibrium is valid constrains $\mu[\theta_p] - \mu[\theta_n] \geq 0$.

(a) (b)

Figure 5: The two types of communication structures stabilized by the lying-averse modification v_{LA} defined in eq. 13.

of communication; and weakly verifiable disclosure which preserves general acyclic candidate equilibria. In this case the assumption of injective equilibria has some bite — we use results from Appendix B.4 to extend our results to non-injective equilibria.

5.3.1 Lying Aversion

Consider a cheap talk sender concerned with honesty, opting to break their indifferences in favour of the truth. In this case, the message space is identified with the state space through a bijection $\psi : M \leftrightarrow \Theta$ where a message m is the claim 'The state is $\psi(m)$.' The sender views the moral cost of lying through the following modification

$$v_{\text{LA}}(m|\theta) := \begin{cases} 0 & \psi(m) = \theta \\ \ell_\theta & \psi(m) \neq \theta, \end{cases} \tag{13}$$

where $\ell_\theta < 0$ for all θ. This supposes all lies from a state θ are equally costly.[23]

Theorem 2 shows that candidate equilibria will generically involve lying. The challenge with lying averse senders is ensuring that they are only incentivized to send the 'right' lies. This significantly constrains communication:

Proposition 5. *For generic receiver utilities u_R, an injective[24] candidate equilibrium σ_0 iff $G(\sigma_0)$ is one of the graphs illustrated in Figure 5.*

When a communication graph features long paths it becomes impossible to incentivize a specific lie from one state without tempting other states to make the same lie. As a result, there can only be one lying message (making the other message a 'single-disclosure') as in Figure 5a, or there can only be one state that lies (garbling the revelation of other messages) as in Figure 5b.

[23] This model of lying is similar to Kartik 2009, which applies a more general version to Crawford-Sobel cheap talk.

[24] In Appendix B.4 we show that non-injective candidate equilibria can be v_{LA}-robust for additional communication graph geometries. These geometries are composed of multiple connected components, one resembling Figure 5a or b, and all other components have the form of a single state θ_k truthfull disclosing their state with the message $\psi^{-1}(\theta_k)$. These equilibria exist only if each of these states have the same receiver best response $a^*(\theta_k)$.

5.3.2 Weakly Verifiable Disclosure

Our last application is a modification to the communication technology. As with lying aversion this turns our communication game into a signalling game, where the signal cost is only weakly state-dependent.

Consider a *weakly verifiable disclosure* game[25] where the sender discloses 'pieces of evidence' $e \in E$ to the receiver, each of which rules out a corresponding state $\psi(e)$ — however the verification process is extremely unreliable, allowing the sender to fabricate information at vanishing cost/likelihood of being caught. The message space is thus $M = 2^E$ where $\psi : E \leftrightarrow \Theta$ is a bijection.

The corresponding modification is then the sum of the cost of presenting the evidence plus the cost of fabricating any false evidence. For a message $m \subseteq E$ providing evidence the state is not in $\psi(m)$, the modification takes the form

$$v_{\text{WD}}(m|\theta) := \begin{cases} \tau_m & \theta \notin \psi(m) \\ \ell_{m,\theta} & \theta \in \psi(m), \end{cases} \tag{14}$$

where $\tau_m > \ell_{m,\theta}$ for all $\theta \in m$ and messages m.

This model allows flexibility in the message content, making this maximally stabling:

Proposition 6. *Let $\sigma_0 \in \Sigma(u^0)$ be an acyclic candidate equilibrium of a cheap talk model (identified up to relabelling of messages). If there is evidence corresponding to every state $\psi(E) = \Theta$, then $v_{WD} \in \mathscr{M}_G(G(\sigma_0))$.*

The key fact in this argument is that acyclicity implies equilibrium beliefs have distinct supports, and thus every belief in such an equilibrium may be associated with a distinct truthful message.

6 Discussion

In this paper we propose that transparent preference models should be interpreted as approximations of models with slightly state-dependent preferences. This interpretation is only valid for a limited subset of candidate equilibria, depending on the type of state-dependence in the preference, through a relationship that we characterize.

In abstract terms, we study the continuities and (lower hemi-)discontinuities of the equilibrium correspondence at transparent sender preferences. This discontinuity in equilibria also allows for many 'discontinuities' in the properties of candidate equilibria: for example, money-burning is not useful if the sender has transparent preferences, but can be significantly useful in some settings with slight state-dependence. Similarly there will be discontinuities in other properties of candidate

[25] We reach verifiable disclosure games as $\varepsilon \to \infty$. Note that we consider a model allowing vague disclosure. Bertomeu and Cianciaruso 2018 provide a broad analysis of such games.

equilibria — eg. long/mediated cheap talk provide no additional value to a sender with transparent preferences, but can be shown to provide significant value to a sender in some models with slightly state-dependent preferences. Adapting our techniques to these communication technologies may provide insight as to whether there are specific state-dependencies that preserve these properties, and if there are realistic classes of models where these technologies can provide significant value.[26]

While we focus on finite states, it seems natural that our model should be able to approximate interval equilibria in one-dimensional communication games ($\Theta, \mathcal{A} \subseteq \mathbb{R}$) as we increase the number of states/actions. It is unclear whether our robustness techniques can be extended to multidimensional continuous state games ($\Theta \subseteq \mathbb{R}^n$). Our analysis does not straight-forwardly apply — connectedness tends to imply cyclicity with such state-topologies — however monotoncity may still be relevant, especially when using finite approximations of state space.

Our analysis focuses on the limiting case of nearly transparent preferences $\varepsilon \to 0$. In a work-in-progress, we show that these results can be extended to 'ordinally transparent' preferences (specifically, sender preferences that order $\mathcal{A} \times M$ according to a transparent ranking) that satisfy a graph version of 'increasing differences'.

To outline the scope of our approach, recall that in Bayesian Perfect Equilibria of communication games the receiver's decision is determined through the following causal process:

1. The sender observes the state and chooses a message,
2. The receiver observes the message and forms a belief,
3. The receiver chooses an action that is a best response to their belief.

We focus on perturbations that introduce small state dependence in to the sender's preference over the first and third steps. Since both correspond to adding a state-dependent component to the sender's utility, they can be studied within the same framework.

The reader might wonder what would be the effect of adding state-dependence to the second step. One way to interpret this is as a *psychological* modification (in the sense of Geanakoplos, Pearce, and Stacchetti 1989) where the sender's utility is dependent on the receiver's belief. This might be induced through a variation on lying aversion where the sender feels guilt for inducing a misleading belief (see e.g. Khalmetski and Sliwka 2019), or an 'effort' associated with shifting the receiver's beliefs. Since this too is a modification of message-induced subgame utility, it can be treated within the above framework.

A distinct way to introduce state-dependence in the second step is if messages induce beliefs in a state dependent manner. This naturally occurs if the receiver is partially informed from an independent source. For example, the receiver observes an informative signal that is (conditional on the state) independent from the sender's message. Since this signal is informative, it will affect the receiver's 'prior' belief. But this informativeness also means that the likelihood of various

[26] I thank Johannes Hörner in particular for bringing these questions to my attention.

signal realizations — which induce different priors — varies across states, thereby introducing state-dependence in the sender's incentives. Arieli, Gradwohl, and Smorodinsky 2023 study this mechanism in binary-state environments. Ideas developed in this paper may help to analyze this mechanism more generally.

While we limit our attention to one-shot games, state-dependence can also arise from weak reputation costs in a repeated game setting. Weak reputation costs can be understood as arising in a population model with random matching. This work then informs the types of punishments that can sustain communication.

References

Arieli, Itai, Ronen Gradwohl, and Rann Smorodinsky (2023). *Informationally Robust Cheap-Talk*. arXiv: 2302.00281.

Austen-Smith, David and Jeffrey S. Banks (2000). "Cheap Talk and Burned Money". In: *Journal of Economic Theory* 91.1, pp. 1–16. ISSN: 0022-0531. DOI: https://doi.org/10.1006/jeth.1999.2591. URL: https://www.sciencedirect.com/science/article/pii/S0022053199925917.

Bertomeu, Jeremy and Davide Cianciaruso (2018). "Verifiable disclosure". In: *Economic Theory* 65.4, pp. 1011–1044. DOI: 10.1007/s00199-017-1048-x. URL: https://doi.org/10.1007/s00199-017-1048-x.

Chakraborty, Archishman and Rick Harbaugh (Dec. 2010). "Persuasion by Cheap Talk". In: *American Economic Review* 100.5, pp. 2361–82. URL: https://www.aeaweb.org/articles?id=10.1257/aer.100.5.2361.

Crawford, Vincent P. and Joel Sobel (1982). "Strategic Information Transmission". In: *Econometrica* 50.6, pp. 1431–1451. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1913390 (visited on 05/23/2022).

Diehl, Christoph and Christoph Kuzmics (2021). "The (non-)robustness of influential cheap talk equilibria when the sender's preferences are state independent". In: *International Journal of Game Theory* 50.4, pp. 911–925. DOI: 10.1007/s00182-021-00774-0. URL: https://doi.org/10.1007/s00182-021-00774-0.

Geanakoplos, John, David Pearce, and Ennio Stacchetti (1989). "Psychological games and sequential rationality". In: *Games and Economic Behavior* 1.1, pp. 60–79. ISSN: 0899-8256. DOI: https://doi.org/10.1016/0899-8256(89)90005-5. URL: https://www.sciencedirect.com/science/article/pii/0899825689900055.

Green, Jerry R. and Nancy L. Stokey (July 2007). "A two-person game of information transmission". In: *Journal of Economic Theory* 135.1, pp. 90–104. URL: https://ideas.repec.org/a/eee/jetheo/v135y2007i1p90-104.html.

Harsanyi, John C. (1973). "Games with randomly disturbed payoffs: A new rationale for mixed-strategy equilibrium points". In: *International Journal of Game Theory* 2.1, pp. 1–23. URL: https://doi.org/10.1007/BF01737554.

Kamenica, Emir and Matthew Gentzkow (Oct. 2011). "Bayesian Persuasion". In: *American Economic Review* 101.6, pp. 2590–2615. DOI: 10.1257/aer.101.6.2590. URL: https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2590.

Kartik, Navin (2007). "A note on cheap talk and burned money". In: *Journal of Economic Theory* 136 (1). ISSN: 00220531. DOI: 10.1016/j.jet.2006.07.001.

— (Oct. 2009). "Strategic Communication with Lying Costs". In: *The Review of Economic Studies* 76.4, pp. 1359–1395. ISSN: 0034-6527. DOI: 10.1111/j.1467-937X.2009.00559.x. eprint: https://academic.oup.com/restud/article-pdf/76/4/1359/18358398/76-4-1359.pdf. URL: https://doi.org/10.1111/j.1467-937X.2009.00559.x.

Khalmetski, Kiryl and Dirk Sliwka (Nov. 2019). "Disguising Lies—Image Concerns and Partial Lying in Cheating Games". In: *American Economic Journal: Microeconomics* 11.4, pp. 79–110. DOI: 10.1257/mic.20170193. URL: https://www.aeaweb.org/articles?id=10.1257/mic.20170193.

Lipnowski, Elliot and Doron Ravid (2020). "Cheap Talk With Transparent Motives". In: *Econometrica* 88.4, pp. 1631–1660. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA15674.

Rochet, Jean-Charles (1987). "A necessary and sufficient condition for rationalizability in a quasi-linear context". In: *Journal of Mathematical Economics* 16.2, pp. 191–200. ISSN: 0304-4068. DOI: https://doi.org/10.1016/0304-4068(87)90007-3. URL: https://www.sciencedirect.com/science/article/pii/0304406887900073.

Steg, Jan-Henrik et al. (2023). *Robust equilibria in cheap-talk games with fairly transparent motives*. arXiv: 2309.04193 [econ.TH].

Vohra, Rakesh V. (2011). *Mechanism Design: A Linear Programming Approach*. Econometric Society Monographs. Cambridge University Press.

Wu Wen-Tsün and Jiang Jia-He (1962). "Essential Equilibrium Points of N-Person Non-Cooperative Games". In: *Scientia Sinica* 11, pp. 1307–1322.

A Additional Proofs

Fragility Proofs

We present an alternate interpretation of Theorem 1 not contained in the body of the paper. In this case $\tilde{U}$ can be interpreted as an estimation of $u^{\varepsilon U}$ based on some individual characteristics $\omega_x(\omega)$ and omitting m. Statement (1) then says that as long as (i) the characteristics ω_x are continuously distributed, (ii) $\omega_x \mapsto \tilde{U}(\cdot|\omega_x)$ has full span, and (iii) the residual $\tilde{u}$ is distributed according to a

density with support independent of $\omega_x(\omega)$ (e.g. full support), then persuasion is impossible.

Proof of Theorem 1. **Proof of (1):** Let σ be an equilibrium with strategy profile $(\mathcal{M}, \mathcal{P})$ and a^* be a pure action played with positive probability. That is, if $M^* := \mathcal{P}^{-1}\{p; a^* \in \text{supp}(p)\}$ is the set of messages leading to a^* being played with positive probability, then the sender sends messages in M^* with positive probability, ie. $\mathbb{P}\left[\mathcal{M}(\theta, \omega) \in M^*\right] > 0$.

Let $\Pi^* := \{(m, \mathcal{P}(m)); m \in M^*\}$ be the set of equilibrium paths leading to action a^*, and $\Pi^c := \{(m, \mathcal{P}(m)); m \notin M^*\}$ be the remaining equilibrium paths. We obtain the following bounds:

$$\mathbb{P}\left[\max_{\pi\in\Pi^*} u(\pi) > \max_{\pi'\in\Pi^c} u(\pi')\right] \leq \mathbb{P}\left[\mathcal{M}(\theta, \omega) \in M^* | \theta\right] \leq \mathbb{P}\left[\max_{\pi\in\Pi^*} u(\pi) \geq \max_{\pi'\in\Pi^c} u(\pi')\right].$$

However

$$\mathbb{P}\left[\max_{\pi\in\Pi^*} u(\pi) = \max_{\pi'\in\Pi^c} u(\pi')\right] = \mathbb{E}\left[\mathbb{P}\left[\max_{\pi\in\Pi^*} u(\pi) = \max_{\pi'\in\Pi^c} u(\pi') \middle| \tilde{u}, (\tilde{U}(a))_{a\neq a^*}\right]\right].$$

Since $u(\pi')$ is determined by $(\tilde{u}, (\tilde{U}(a))_{a\neq a^*})$, the conditional probability amounts to the probability that $\max_{\pi\in\Pi^*} u(\pi)$ is equal to a constant. But $\max_{\pi\in\Pi^*} u(\pi)$ is strictly increasing in $\tilde{U}(a^*)$, so there is a unique value of $\tilde{U}(a^*)$ that will yield this equality. Since the utility admits a density, the conditional probability of $\tilde{U}(a^*)$ being precisely this value is 0, allowing us to conclude

$$\mathbb{P}\left[\mathcal{M}(\theta, \omega) \in M^* | \theta\right] = \mathbb{P}\left[\max_{\pi\in\Pi^*} u(\pi) > \max_{\pi'\in\Pi^c} u(\pi')\right]$$

is state-independent. As $\mathbb{P}\left[\mathcal{M}(\theta, \omega) \in M^*\right] > 0$, this means the average posterior that induces a randomization including a^* is equal to the prior μ. Thus the prior is a convex combination of beliefs to which a^* is a best response, indicating a^* is a best response to the prior. Thus pure actions are either sent with probability zero, or are a best response to the prior. Since there are finite pure actions, the total probability of one of the former actions being played is zero, making the equilibrium unpersuasive.

Proof of (2): Meagreness is a corollary of (1), under the following two observations:

1. The set of preferences described by eq. 5 is dense.

2. The set of preferences that permit persuasive equilibria is the countable union of closed sets.

The first property is trivial. To observe the second, consider the set of persuasive equilibria where an action a^* that is not a best response to the prior is chosen w.p. at least $\frac{1}{n}$. This is a closed set of equilibria, by upper-hemicontinuity the corresponding set of preferences is also closed. We know that the complement is then open, and by the first property includes a dense set, thus these equilibria only occur over a nowhere dense set of preferences.

By taking the union of these preferences over n, we find that persuasion is only possible within a meagre set (ie. a countable union of nowhere dense sets). □

Note that the topological properties invoked in the last proof are desirable for any topology on random preferences: the first is a consequence of preferences converging as a perturbation vanishes, the second is implied by best responses being upper hemicontinuous. Thus persuasion is generically impossible for any 'reasonable' topology, not just the weak-$*$ topology.

Generic Structure of Communication Graphs

We state our alternate result for fixed u_R and generic prior beliefs. We make the following weak (generic) assumption:

Assumption (R). *[Receiver Distinguishes between Best Responses.] If $a, a' \in \mathcal{A}$ are distinct receiver-best responses to a belief $\nu \in \Delta\Theta$ then there exists a state $\theta \in supp(\nu)$ such that $u_R(a|\theta) \neq u_R(a'|\theta)$.*

An immediate consequence of this assumption is that pure beliefs have unique best responses. More generally, this says that when restricted to any subset $\Theta_1 \subseteq \Theta$, if two actions are equivalent to the reciever (ie. they yield identical utilities for all states in Θ_1), they must be irrelevant (ie. they are never best responses for beliefs in Θ_1). As a result the set of beliefs in $\Delta\Theta_1$ where the receiver is indifferent between two best responses has codimension 1.

There is also a slightly stronger version of this assumption that we will also use:

Assumption (R$*$). *[Receiver Distinguishes within 3-Best Responses.] Let $a, a', a'' \in \mathcal{A}$ be (possibly non-distinct) receiver-best responses to a belief $\nu \in \Delta\Theta$. If $p, p' \in \Delta\{a, a', a''\}$ and $u_R(p|\theta) = u_R(p'|\theta)$ for all $\theta \in supp(\nu)$, then $p = p'$.*

This can be equivalently stated in terms of the preference over pure actions as follows: if a is equivalent to a convex combination of a', a'' (when restricted to any $\Theta_1 \subseteq \Theta$) then a is irrelevant (over $\Delta\Theta_1$).

This implies that for any $\Theta_1 \subseteq \Theta$, (i) the set of beliefs in $\Delta\Theta_1$ where the receiver is indifferent between two best responses has codimension 1, and (ii) the set of beliefs in $\Delta\Theta_1$ where the receiver is indifferent between three best responses has codimension 2.

These assumptions are necessary conditions for an equilibrium involving such mixed actions to be Harsanyi-robust to perturbations in the *receiver's* utility. They are generically satisfied on the function space $\mathbb{R}^{\mathcal{A}\times\Theta}$.

Theorem 2'. *If Assumptions (S) and (R$*$) are satisfied and the sender has transparent preferences, then properties (a-c) of Theorem 2 hold over the set of injective acyclic candidate equilibria for generic priors $\mu \in \Delta\Theta$.*

This is a corollary to the following version of Lemma 1, under the observation that Assumption (S) and injectivity implies that candidate equilibria contain at most one pure action.

Lemma 1'. *Under Assumptions (S,R), for generic $\mu \in \Delta\Theta$ every connected component of an acyclic candidate equilibrium's communication graph*

(i) *includes at least one message whose best response is a unique pure action a^*,*

(ii) *in the case where this message is unique and (R$*$) holds, then*

- ***(a)*** *every other message on the connected component has best responses that are a binary lottery,*
- ***(b)*** *this connected component corresponds to a unique sender strategy (when restricted to states on this component).*

To prove this lemma (and Lemma 1), we first establish some notation:

For a rooted tree G_0, for a node j, we denote the set of j's children by $j^{\downarrow}$ and its parent by $j^{\uparrow}$. This will often be pushed to subscripts — eg. a message π_j will have children from the set of states $\Theta_{j^{\downarrow}}$, whereas a state θ_j will have children $\Pi_{j^{\downarrow}}$. The set of j's neighbours will be denoted $\mathcal{N}(j) := \{j^{\uparrow}\} \cup j^{\downarrow}$.

We also require notation for the sets of beliefs where the receiver is indifferent between multiple actions. For a set of pure actions $A \subseteq \mathcal{A}$, define

$$\mathcal{I}(A) := \left\{ \nu \in \Delta\Theta; A \subseteq \arg\max_{a \in \mathcal{A}} \mathbb{E}_{\theta \sim \nu} \left[u_R(a, \theta) \right] \right\}$$

to be the set of beliefs such that the receiver is willing to randomize over A (ie. choosing an action $p \in \text{int}(\Delta A)$).

We let

$$\mathcal{I}_2 := \{\mathcal{I}(A); |A| = 2\} \qquad \mathcal{I}_{3+} := \{\mathcal{I}(A); |A| \geq 3\}$$

refer to collections of these sets where the receiver is willing to randomize between 2 actions, and 3 or more actions respectively. Assumption (R) implies every set in $\mathcal{I}_2$ has codimension 1, while (R$*$) implies every set in $\mathcal{I}_{3+}$ has codimension at least 2.

Proof of Lemma 1 and 1'. We first prove Lemma 1', before showing how this proof can be translated to show Lemma 1.

WLOG we limit ourselves to priors $\mu \in \text{int}(\Delta\Theta)$. Suppose there is an acyclic connected component $G_0 \subseteq G$ of the communication graph that does not include a pure action. We turn G_0 into a rooted tree with arbitrary message root π_0. We make an inductive argument up the tree to the root:

1. Leaf Case: We begin with the vertices furthest from the root. Since there are no pure actions in G_0, all the leaves are states (otherwise they are actions corresponding to a known state — thus

pure under Assumption (R)). The deepest vertex is thus a state connected to a single message, say π_j, which it recommends w.p. 1 in equilibrium. Since this is a deepest action, all its other children $\theta_k \in \Theta_{j\downarrow}$ also recommend it w.p. 1.

The only degree of freedom that G admits in the belief ν_j is the probability that it is recommended from its parent state $\theta_{j\uparrow}$. Thus for a fixed prior, the set of possible ν_j corresponding to this graph has dimension 1.

Letting the probability $\mu_{j\downarrow}$ of child states occuring vary, the set of possible posteriors ν_j under G is a subset of

$$\mathcal{P}_j := \left\{ \tfrac{1}{\lambda+|\mu_{j\downarrow}|} \left(\lambda \delta_{\theta_{j\uparrow}} + \mu_{j\downarrow} \right) ; \, \lambda \in \left]0,1\right[, \mu_{j\downarrow} \in \operatorname{int}\left(\Delta\Theta_{j\downarrow}\right) \right\} \subseteq \operatorname{int}\left(\Delta\Theta_{\mathcal{N}(j)}\right),$$

where the denomonator is normalization. Note that this set is an open subset of $\Delta\Theta_{\mathcal{N}(j)}$, hence has codimension 0. Thus it is trivially transversal to any best response set.

Applying the transversality theorem, for a.a. $\mu_{j\downarrow} \in \Delta\Theta_{j\downarrow}$, the set

$$\mathcal{P}_j(\mu_{j\downarrow}) := \left\{ \tfrac{1}{\lambda+|\mu_{j\downarrow}|} \left(\lambda \delta_{\theta_{j\uparrow}} + \mu_{j\downarrow} \right) ; \, \lambda \in \left]0,1\right[\right\}$$

is transversal to any $\mathcal{I}(A)$. Thus, within this generic set, $\mathcal{P}_j(\mu_{j\downarrow})$ intersects $I \in \mathcal{I}_2$ at finitely many points (by convexity, at most one), and, under assumption (R$*$), never intersects the codimension 2 sets $I \in \mathcal{I}_{3+}$.

Fixing such a $\mu_{j\downarrow}$, there are finite probabilities that the sender can send the message π_j from state $\theta_{j\uparrow}$ and induce a mixed action, and by convexity a single probability for a given mixed action.

2. Induction step: We now let π_j be a non-root message further up the tree G_0, and $\theta_{j\uparrow}$ be its parent state. Our inductive assumption is that there are finite probabilities $\lambda_{j\downarrow}$ that its child states $\theta_k \in \Theta_{j\downarrow}$ recommend their own child actions $\pi_{k\downarrow}$ while rationalizing these actions. We then claim that fixing this $\lambda_{j\downarrow}$, there are finite probabilities λ that its parent state can send the message π_j to make the receiver indifferent between two actions.

Indeed, fixing the posteriors generated in descendent actions, we find that the range of posteriors that can be generated by the message π_j is a subset of

$$\mathcal{P}_j := \left\{ \tfrac{1}{\lambda+|\mu_{j\downarrow}-\lambda_{j\downarrow}|} \left(\lambda \delta_{\theta_{j\uparrow}} + (\mu_{j\downarrow} - \lambda_{j\downarrow}) \right) ; \, \lambda \in \left]0,1\right[\left[, \mu_{j\downarrow} > \lambda_{j\downarrow}, |\mu_{j\downarrow}| < 1 \right\} \subseteq \operatorname{int}\left(\Delta\Theta_{\mathcal{N}(j)}\right).$$

where $\lambda_{j\downarrow}$ is fixed. This set has codimension 0 in $\Delta\Theta_{\mathcal{N}(j)}$.

Applying the transversality theorem once more, the set of feasible posteriors

$$\mathcal{P}_j(\mu_{j\downarrow}) := \left\{ \tfrac{1}{\lambda+|\mu_{j\downarrow}-\lambda_{j\downarrow}|} \left(\lambda \delta_{\theta_{j\uparrow}} + (\mu_{j\downarrow} - \lambda_{j\downarrow}) \right) ; \, \lambda \in \left]0,1\right[\right\}$$

is thus transversal to any $\mathcal{I}(A)$ for a.a. $\mu_{j\downarrow} \in \Delta\Theta_{j\downarrow}$. As a result, $\mathcal{P}_j(\mu_{j\downarrow})$ intersects any $I \in \mathcal{I}_2$ at most once, and (under assumption (R$*$)) never intersects any $I \in \mathcal{I}_{3+}$ within this generic set.

3. Conclusion: This shows that generically (non-root) actions have support of at most two. But if we consider the rooted message, the above reasoning shows that there are finite posteriors that can be generated when all its descendants are constrained to generate mixed actions. Applying the transversality theorem (as above, but with no λ as there is no parent state) we find that for a.a. $\mu_{j\downarrow}$ the posteriors do not intersect any set $I \in \mathcal{I}_2$ of posteriors inducing a mixed action. Thus one action must be pure.

Since there are finite acyclic communication structures (up to choice of π), the set of priors that admit a forest candidate equilibrium is the finite union of measure-0 sets, and hence is measure-0.

Moreover, fixing the indifferences required by each message $(\mathcal{I}_2)^M$, there is a at most one sender strategy capable of generating these indifferences within the acyclic communication graph, giving us property (ii).

Translation to proof of Lemma 1: To show the above results also hold for fixed μ and generic u_R, we make Assumption (R$*$) (recalling that it holds for generic receiver preferences), and note that the above theorem can be reproduced by holding $\mu_{j\downarrow}$ fixed at each step, and varying $I(A) \in \mathcal{I}_2$ through the receivers utility.

Observe that the set $\{I(A; u_R)\}$ can be perturbed in any direction by adjusting u_R within this set of utilities, hence it is trivially transversal to $\mathcal{P}_j(\mu_{j\downarrow})$. Thus, for generic u_R, $I(A; u_R)$ is transversal to $\mathcal{P}_j(\mu_{j\downarrow})$. Since the latter is one dimensional (parametrized solely by the probability λ), and the former has codimension one when $|A| = 2$ (by Assumption (R$*$)), this means that the intersection consists of a finite set of weights the parent state can send the message with that will induce randomization over two actions (and no weights that will induce higher support randomizations). □

$\mathcal{O}$-robustness Proofs

Proposition 1 is an immediate corollary of the more general proposition:

Proposition 7. *Any pair of equilbirium paths $(p_0, m_0), (p_1, m_1)$ taken in an equilibium where the senders preference is $u^0 - \varepsilon u_R$ (with $\varepsilon > 0$) satisfy $u^0(p_0, m_0) = u^0(p_1, m_1)$ and $u_R(p_0|\theta) = u_R(p_1|\theta)$ for μ-a.a. θ.*

We prove this result in the more general setting where the state belongs to an arbitrary measure space, and actions belong to an arbitrary set, possibly infinite.

Proof. Let one equilibrium action p_0 be associated with the message m_0 and posterior ν_0, and another equilibrium action p_1 be associated with message m_1 and posterior ν_1. For these to be optimal messages for the sender it must be the case that

$$\begin{aligned} 0 \leq & u^{-\varepsilon u_R}(p_0, m_0|\theta_0) - u^{-\varepsilon u_R}(p_1, m_1|\theta_0) \qquad \text{for } \nu_0\text{-a.a. } \theta_0 \\ 0 \leq & u^{-\varepsilon u_R}(p_1, m_1|\theta_1) - u^{-\varepsilon u_R}(p_0, m_0|\theta_1) \qquad \text{for } \nu_1\text{-a.a. } \theta_1. \end{aligned} \tag{15}$$

Integrating these inequalities over their respective posteriors and summing, we obtain

$$0 \leq -\int \varepsilon(u_R(p_0|\theta_0) - u_R(p_1|\theta_0))\, d\nu_0(\theta_0) - \int \varepsilon(u_R(p_1|\theta_1) - u_R(p_0|\theta_1))\, d\nu_1(\theta_1) \leq 0,$$

where the last inequality is obtained from p_i being a best response to ν_i. This implies eq. 15 holds with equality ν_0-/ν_1-a.e. If there are only two messages we are done.

Otherwise repeat the above procedure, comparing p_0, p_1 to other equilibrium actions p' corresponding to the message m' and posterior ν' to find

$$u^{-\varepsilon u_R}(p_0, m_0|\theta) = u^{-\varepsilon u_R}(p', m'|\theta) = u^{-\varepsilon u_R}(p_1, m_1|\theta) \qquad \text{for } \nu'\text{-a.a. } \theta.$$

Integrating over the equilibrium distribution of posteriors then shows that $u^{-\varepsilon u_R}(p_0, m_0|\theta) = u^{-\varepsilon u_R}(p_1, m_1|\theta)$ and $u_R(p_0|\theta) = u_R(p_1|\theta)$ for μ-a.a. θ. □

Harsanyi $\mathcal{O}$-robustness Proofs

If we wish to work with perturbations that do not satisfy mutual independence, we must separate out the distribution of perturbations conditional on each state — these conditional distributions are what actually determines the likelihood a message is sent from a particular state. We define the **state-factored support** (denoted **supp(·|Θ)**) of an idiosyncratic perturbation V:

$$\text{supp}(V|\Theta) := \bigtimes_{\theta\in\Theta} \text{supp}_{\mathbb{P}_\omega}(V(\cdot|\theta)) \subseteq \mathbb{R}^{(\mathcal{A}\times M)\times\Theta}. \tag{16}$$

This is the smallest 'rectangular' set that contains $\text{supp}_{\mathbb{P}_\omega}(V)$. If $\tilde{V}$ is the reparametrization of V that satisfies mutual independence (defined in footnote 17), then $\text{supp}(\tilde{V}) = \text{supp}(V|\Theta)$. This means that the definitionn of Harsanyi $\mathcal{O}$-robustness can be extended to V that are not mutually independent by replacing $\text{supp}(V)$ with $\text{supp}(V|\Theta)$ in the definition, allowing us to retain the interpretation of ω as reflecting an individual's preference.

The following lemma is essential for the proof of Theorem 4, essentially replacing the use of intermediate value theorem in the proof of Theorem 3.

Lemma 2 (Neighbour Incentive Compatability)**.** *Let $u^{\varepsilon V} : \Delta\mathcal{A} \times M_0 \times \Omega \to \mathbb{R}$ be an idiosyncratic sender preference (fixing the state), where M_0 is finite. Fix the action π_0 associated with some message $m_0 \in M_0$, and for $j \neq 0$, let $B_j \subseteq \Delta supp(\pi_j)$ be closed, convex intervals satisfying*

$$\inf_{\pi'\in B_j} \mathbb{P}_\omega\left[u^{\varepsilon V}(\pi') > u^{\varepsilon V}(\pi_0)\right] = 0 \qquad \sup_{\pi'\in B_j} \mathbb{P}_\omega\left[u^{\varepsilon V}(\pi') > u^{\varepsilon V}(\pi_0)\right] = 1. \tag{17}$$

For any mixed sender strategy $\mathcal{M} \in \Delta M_0$ there exists a profile of actions $\hat{\pi}_+ \in \bigtimes_{j\neq 0} B_j$ that induce the sender population best response $\mathcal{M}$.

This generalizes intermediate value theorem in the sense that eq. 17 is equivalent to each pure strategy $\mathcal{M} \in M_0$ being the unique best response to some action profile π_+, and we conclude that every mixed strategy in ΔM_0 also a best response to some action profile.

Proof of Lemma 2. Fix the message-action associated with m_0 to be π_0 and denote $B_0 := \{\pi_0\}$. Given actions $\pi_+ \in \bigtimes_{j\neq 0} B_j$, the probability the sender sends the message π_j is described by the correspondence $\hat{\mathcal{M}} : B \rightrightarrows \Delta M_0$ bounded[27] by

$$\mathbb{P}_\omega\left[u^{\varepsilon V}(\pi_j) \geq \max_{j'\neq j} u^{\varepsilon V}(\pi_{j'})\right] \geq \hat{\mathcal{M}}_j(\pi) \geq \mathbb{P}_\omega\left[u^{\varepsilon V}(\pi_j) > \max_{j\neq j'} u^{\varepsilon V}(\pi_j)\right].$$

Order B_j so that higher actions are more attractive to the sender, so that

$$\underline{\pi}_j := \arg\min_{\pi'\in B_j}\left\{\mathbb{P}_\omega\left[u^{\varepsilon V}(\pi') > u^{\varepsilon V}(\pi_0)\right]\right\} \qquad \overline{\pi}_j := \arg\max_{\pi'\in B_j}\left\{\mathbb{P}_\omega\left[u^{\varepsilon V}(\pi') > u^{\varepsilon V}(\pi_0)\right]\right\}$$

are the lower/upper bound of B_j. By eq. 17, the action $\underline{\pi}_j$ makes it a best response to never send its message: $0 \in \hat{\mathcal{M}}_j(\underline{\pi}_j, \pi_{-j})$ for any $\pi_{-j} \in B_{-j} := \bigtimes_{j'\neq j} B_{j'}$; while a receiver having the strategy corresponding to $\overline{\pi}_j$ ensures that it is optimal to never send π_0 in equilibrium, ie. $0 \in \hat{\mathcal{M}}_0(\overline{\pi}_j, \pi_{-j})$.

Define the correspondence $g_j^* : B_{-j} \rightrightarrows B_j$

$$g_j^*(\pi_{-j}) := \begin{cases} \{\overline{\pi}_j\} & \hat{\mathcal{M}}_j(\overline{\pi}_j; \pi_{-j}) < \mathcal{M}_j \\ \{\pi'_j \in B_j; \hat{\mathcal{M}}_j(\pi'_j; \pi_{-j}) = \mathcal{M}_j\} & \text{otherwise.} \end{cases}$$

Note this is never empty-valued — since $\hat{\mathcal{M}}_j(\cdot, \pi'_{-j})$ is increasing, convex valued, and upper hemicontinuous it satisfies an intermediate value theorem, furthermore eq. 17 implies $\hat{\mathcal{M}}_j(\underline{\pi}_j; \pi_{-j}) < \mathcal{M}_j$ for any π_{-j}. Together these properties show that g_j^* is upper-hemicontinuous and interval-valued.

Define the self-map $g : B \rightrightarrows B$

$$g(\pi) = \bigtimes_{j\neq 0} g_j^*(\pi_{-j})$$

By Kakutani fixed point theorem, there exists a fixed point π^* of this map. Suppose this fixed point does not solve $\hat{\mathcal{M}}_j(\pi^*) = \mathcal{M}_j$, then $\pi_j^* = \overline{\pi}_j$ for at least one j. But then $\hat{\mathcal{M}}_0(\pi^*) < \mathcal{M}_0$, as earlier observed. As a result there must be a message j' (not necessarily j) sent with probability $\hat{\mathcal{M}}_{j'}(g_{j'}^*(\pi_{-j'}^*); \pi_{-j'}^*) > \mathcal{M}_{j'}$, contradicting the definition of $g_{j'}^*$. □

[27] No insight is lost by assuming u admits a density, hence $\hat{\mathcal{M}}$ is single-valued. More generally $\hat{\mathcal{M}}(\pi)$ is the intersection of the simplex ΔM_0 with the 'box' in $[0,1]^{M_0}$ whose sides are described by the given interval.

Proof of Theorem 4. We use the same rooted tree and notation as in the proof of Theorem 3.

1. The Neighbour Problem: To ensure each state on the tree randomizes over its neighbouring messages to the desired degree, first constrain each state to neighbouring messages. Suppose $N_j \ni \pi_j$ are interval neighbourhoods of π_j for all non-pure actions π_j (let $N_0 := \{\alpha_0\}$ for the pure action). We assume ε is sufficiently small that the sender's ordinal preference over pure actions is independent of idiosyncrasy.

We seek $\overline{\varepsilon} > 0$ and interval neighbourhoods B_j such that

$$\min u^{\varepsilon V}(B_j|\theta_k) < u^{\varepsilon V}(\pi_{k^\uparrow}|\theta_k) < \max u^{\varepsilon V}(B_j|\theta_k) \qquad \mathbb{P}_\omega\text{-a.a., for all } \pi_{k^\uparrow} \in B_{k^\uparrow}, \theta_k, \varepsilon < \overline{\varepsilon}. \quad (18)$$

for all $j \in k^\downarrow$. This allows us to apply Lemma 2 to find mixed actions $\hat{\pi} \in B := \bigtimes B_j$ such that each state sends its neighbouring messages with the desired probability.

Beginning with mixed actions π_j without grandchildren, define $B_j := N_j$. We denote the endpoints of $B_j =: [\underline{\pi}_j, \overline{\pi}_j]$ where actions on the right side of the interval are unanimously prefered to actions to their left. Moving up the tree, for mixed actions π_j with grandchildren, define

$$B_j := \bigcap_{k \in j^\downarrow} \bigcap_{j' \in k^\downarrow} \{\pi' \in N_j; u^{\varepsilon V}(\underline{\pi}_{j'}|\theta_k) \leq u^{\varepsilon V}(\pi'|\theta_k) \leq u^{\varepsilon V}(\overline{\pi}_{j'}|\theta_k) \text{ w.p. } 1\}.$$

to be the constraint imposed by its grandchild actions $j^{\downarrow\downarrow}$. This set may be empty for some ε, but observe that if $\pi_{j'} \in B_{j'}$ for all $j' \in j^{\downarrow\downarrow}$, then $B_j \to O_j$ as $\varepsilon \to 0$ where

$$O_j := \left\{\pi' \in N_j; u^0(\pi') \in \left]\max_{j' \in j^{\downarrow\downarrow}} \min u^0(B_{j'}), \min_{j' \in j^{\downarrow\downarrow}} \max u^0(B_{j'})\right[\right\} \ni \pi_j,$$

since $B_{j'} \ni \pi_{j'}$. Thus when ε is sufficiently small, B_j is a neighbourhood of π_j. Applying this inductively up the tree, we simultaneously have neighbourhoods around every mixed action, and can apply Lemma 2 to find actions $\hat{\pi}$ that make senders in each state θ send each neighbouring message with the prescribed probability.

1.1 Harsanyi robustness To get the limit (1) in the second part of the theorem, bound $\varepsilon < \overline{\varepsilon}$ so that the proof of Theorem 3 ensures states have strict preferences for their own limit equilbirium actions at v. Then repeat the above step with $V \to v$ in distribution instead of $\varepsilon \to 0$, and π representing the equilibrium actions of $u^0 + \varepsilon v$ rather than u^0.

2 Non-neighbouring deviations Assume that the neighbourhoods N_π collectively solve eq. 9. Once we have actions $\hat{\pi}$ that induce the desired degree of randomization when constrained to on-path messages, we need to check that agents are not tempted by off-path messages. To observe a sender in state θ_1 will w.p. 1 have a strict preference for neighbouring messages, consider a sequence $(\theta_1, \hat{\pi}_1, \ldots, \hat{\pi}_N)$. Applying intermediate value theorem, we find a modification

$\hat{v} \in \text{co}(\text{supp}(V|\Theta)) \subseteq \mathscr{M}_G(G(\sigma_0))$ that is indifferent over each message is this sequence. Consider the realized modification

$$v_\omega(\pi|\theta) := \begin{cases} \hat{v}(\pi|\theta) & \theta \neq \theta_1 \\ V(\pi|\theta_1, \omega) & \theta = \theta_1. \end{cases}$$

Since $v_\omega \in \mathscr{M}_G(G(\hat{\sigma}))$ w.p. 1 we can apply Theorem 3 to conclude that non-neighbouring messages are not attractive. □

Proofs for Applications

Proof of Proposition 3. We will show that for any path $(\theta_0, \pi_0, \ldots, \theta_N, \pi_N)$ where $(\theta_0, \ldots, \theta_N)$ is $\succ$-monotone (WLOG increasing), we have

$$u_R(\alpha_0|\theta_0) - u_R(\alpha_N|\theta_0) > u_R(\alpha_0|\theta_1) - u_R(\alpha_N|\theta_1)$$

for all $\alpha_i \in \text{supp}(\pi_i)$. Single-crossing implies that for a fixed α_0, α_N pair either

$$\begin{aligned} u_R(\alpha_0|\theta) - u_R(\alpha_N|\theta) > u_R(\alpha_0|\theta') - u_R(\alpha_N|\theta') \quad &\text{for all } \theta' \succ \theta, \text{ or} \\ u_R(\alpha_0|\theta) - u_R(\alpha_N|\theta) < u_R(\alpha_0|\theta') - u_R(\alpha_N|\theta') \quad &\text{for all } \theta' \succ \theta. \end{aligned}$$

Note that our equilibrium requires that α_N is a best response to a belief supported on states $\theta' \succeq \theta_N$, and that α_0 is a best response to a (non-degenerate) belief supported on states $\theta \preceq \theta_N$. This is incompatible with the second possibility, so the first must hold. This implies that

$$u_R(\pi_i|\theta_i) - u_R(\pi_N|\theta_i) > u_R(\pi_i|\theta_{i+1}) - u_R(\pi_N|\theta_{i+1})$$

for all paths $(\theta_0, \pi_0, \ldots, \theta_N, \pi_N)$ on $G(\sigma_0)$ and $i \in \{0, \ldots, N-1\}$. By rearranging and summing this inequality over $i = 0$ to $i = N - 1$ we obtain graph monotonicity:

$$\sum_{i=0}^{N-1} u_R(\pi_i|\theta_i) - u_R(\pi_N|\theta_i) - u_R(\pi_i|\theta_{i+1}) + u_R(\pi_N|\theta_{i+1}) = \sum_{i=0}^{N} u_R(\pi_i|\theta_i) - u_R(\pi_i|\theta_{i-1}) > 0.$$

Quantitative model, uniqueness: First note that because the receiver only randomizes between $\succ_{\mathcal{A}}$ neighbouring actions, equilibrium actions can be ordered by $\succ_{\mathcal{A}}$. Since u^0 is $\succ_{\mathcal{A}}$-single-dipped, there are at most two actions that the sender is ever indifferent between. Consider the state that randomizes between the two recommendations. For Graph Monotonicity to hold, this equilibrium must have a threshold structure: every state that recommends the higher (lower) action must be higher (lower) than this state. interval equilibria will thus be threshold equilibria where one message induces high belief $\overline{\nu}$ and the other a low belief $\underline{\nu}$.

Define Δ_T to be the set of posteriors that can be induced by a threshold equilibrium[28], this set is linearly ordered by $\succ_{\text{FOSD}}$. The set of threshold strategies can be parametrized by $t = (\theta^*, \underline{p})$ where θ^* is the threshold and $\underline{p}$ is the probability of sending the high message from the threshold state, note that if we order t lexicographically then $t \mapsto (\overline{\nu}(t), -\underline{\nu}(t))$ is monotonically increasing.

Let $\hat{\nu} \in \arg\min_{\nu \in \Delta_T} u^*(\nu)$ be the posterior where the sender's least preferred action is induced. WLOG $\hat{\nu} \succeq_{\text{FOSD}} \mu$. Let $\underline{t} := \overline{\nu}^{-1}(\hat{\nu})$. A candidate equilibrium must have $\overline{\nu} \succ_{\text{FOSD}} \hat{\nu} \succ_{\text{FOSD}} \underline{\nu}$, otherwise the utility from one posterior dominates the other. This implies that the threshold satisfies $t \geq \underline{t}$, But $t \mapsto u^*(\overline{\nu}(t)) - u^*(\underline{\nu}(t))$ is monotonically increasing for $t \geq \underline{t}$ (only constant when both posteriors induce pure actions). Thus there is a unique u_R-robust equilibrium.

Sender-preferred-ness: Consider any non-interval candidate equilibrium. Due to our assumptions on u_R and u^0, this involves two actions that are $\succ_{\mathcal{A}}$-ordered, induced by posteriors $\underline{\nu}_2, \overline{\nu}_2$. One of these actions is mixed $p \in \Delta\{\underline{a}, \overline{a}\}$. There will be a threshold posterior $\underline{\nu}_0$ that induces this mixed action as well (with complementary posterior $\overline{\nu}_0$). We assume (WLOG) that $\underline{a} \prec_{\mathcal{A}} \overline{a}$, $\underline{\nu}_2$ corresponds to this mixed action, and $\underline{\nu}_0$ is a lower threshold interval.

We claim that $\overline{\nu}_0 \succ_{\text{FOSD}} \overline{\nu}_2$. In this case $u^*(\overline{\nu}_0) \geq u^*(\overline{\nu}_2)$. If this holds with equality, then we have found an interval equilibrium that obtains the same sender utility. If this holds with a strict inequality, then by shifting the threshold lower, we increase $u^*(\underline{\nu}_0)$ and decrease $u^*(\overline{\nu}_0)$, eventually attaining equality between the two, at a higher sender-utility.

To show that $\overline{\nu}_0 \succ_{\text{FOSD}} \overline{\nu}_2$, we parametrize the sender's strategy $\mathscr{M}$ by the probability $\overline{m}(\theta)$ that it sends the high message in each state θ. For the mixed action p to be a best response to the low message, this strategy must solve

$$\begin{aligned} \sum_{\theta \in \Theta} \left[u_R(\overline{a}|\theta) - u_R(\underline{a}|\theta)\right]\left[1 - \overline{m}(\theta)\right] &= 0 \\ 0 \leq \overline{m}(\theta) &\leq 1 \end{aligned} \tag{19}$$

Let $\theta_a := \min\{\theta; u_R(\overline{a}|\theta) > u_R(\underline{a}|\theta)\}$ be the first state where the receiver prefers the action $\overline{a}$ to $\underline{a}$.

The threshold strategy $\overline{m}_0$ corresponding to ν_0 is characterized by solving eq. 19 and having the structure of being 1 above some state θ^* and 0 below it, where single-crossing implies $\theta^* \geq \theta_a$.

We consider an intermediate solution $\overline{m}_1$ of eq. 19 that for some θ^{**} satisfies

$$\begin{cases} \overline{m}_1(\theta) = \overline{m}_2(\theta) & \theta < \theta^{**} \\ \overline{m}_1(\theta) \in [\overline{m}_0(\theta), \overline{m}_2(\theta)] & \theta = \theta^{**} \\ \overline{m}_1(\theta) = \overline{m}_0(\theta) & \theta > \theta^{**} \end{cases}$$

where $\theta_a \leq \theta^{**} \leq \theta^*$ (intermediate value theorem assures such a solution exists). Intuitively, we are splitting the transformation into two steps: (1) moving from $\overline{m}_0$ to $\overline{m}_1$ focuses on changes below the threshold θ^*, (2) moving from $\overline{m}_1$ to $\overline{m}_2$ focuses on changes above the threshold θ_a.

[28]Formally $\Delta_T := \bigcup_{\theta^* \in \Theta} \text{co}(\{\mu|_{\theta \succ \theta^*}\}, \{\mu|_{\theta \succeq \theta^*}\}); \theta^* \in \Theta\} \cup \text{co}(\{\mu|_{\theta \prec \theta^*}\}, \{\mu|_{\theta \preceq \theta^*}\}); \theta^* \in \Theta\}$

The change in posteriors from the first step is

$$\overline{\nu}_0(\theta) - \overline{\nu}_1(\theta) = \frac{\overline{m}_0(\theta)}{p_0} - \frac{\overline{m}_1(\theta)}{p_1} \qquad \text{where } p_i = \sum_{\theta'} \mathscr{M}_i(\overline{m}|\theta').$$

Note that $p_0 < p_1$, making $\overline{\nu}_0(\theta) - \overline{\nu}_1(\theta)$ negative before θ^{**} and positive after. This implies $\overline{\nu}_0 \succ_{\text{FOSD}} \overline{\nu}_1$.

The second shift only affects states weakly above θ^{**} (therefore weakly above θ_a). Since both strategies solve eq. 19, $\theta \mapsto u_R(\overline{a}|\theta) - u_R(\underline{a}|\theta)$ is increasing, and $\text{sgn}(\overline{m}_2 - \overline{m}_1)$ has a threshold structure (being positive until θ^* and negative after), the high message is sent more often in the second strategy: $p_2 > p_1$. We then apply the above reasoning to deduce that $\overline{\nu}_2 \prec_{\text{FOSD}} \overline{\nu}_1$. □

Proof of Proposition 4. Acyclic candidate equilibria that attain a non-zero payoff will involve two messages: one recommending mixed action p_F, enacting policy F and $\emptyset$ with equal probability; and one recommending policy P.[29]

Observe that a receiver only chooses P if θ_p is in the support of their posterior, likewise for F and θ_f. However

$$v_E(p_F|\theta_f) - v_E(P|\theta_p) = \tfrac{11}{4} > v_E(p_F|\theta_p) - v_E(P|\theta_p) = \tfrac{-5}{4} > v_E(p_F|\theta_n) - v_E(P|\theta_n) = \tfrac{-7}{4}.$$

Suppose there is a persuasive candidate equilibrium:

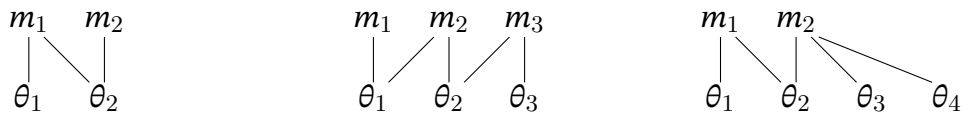
- If state θ_n recommends p_B with positive probability, then graph monotonicity implies that state θ_a only recommends p_B, contradicting our previous observation.
- If θ_n only recommends A, then p_B is induced by a message sent only by states θ_a, θ_b. However $\emptyset$ (hence p_B) is never a best response on $\Delta\{\theta_a, \theta_b\}$.

There are evidently acyclic candidate equilibria whose convex hull contains priors in the cross-hatched region, which can thus be preserved by appropriate modifications. □

The following lemma is essential to the proof of Proposition 5:

Lemma 3 (N-Shaped Subgraphs)**.** *Consider a lying averse model, and a candidate equilibrium σ_0 with communication graph $G(\sigma_0)$ containing an* N*-shaped subgraph, in that two states θ_1, θ_2 both send a message m_1 and one of them (θ_2) sends another message m_2. If $v_{LA} \in \mathscr{M}_G(G(\sigma_0))$, then $\psi(m_1) = \theta_1$ or $\psi(m_2) = \theta_2$.*

The N-shaped subgraph is illustrated in Figure 6a. We will say such an 'N' is *spanned by* its endpoints — in this case θ_1-m_2. The content of this lemma is intuitive: for lying aversion to be monotone relative to this 'N', the sender in state θ_1 must be biased towards the message they



(a) (b) (c)

Figure 6: In (a), the communication sub-graph referred to in Lemma 3. In (b), (c) are the two sub-graphs the lemma rules out.

send more than the sender in state θ_2. The only way this can happen is if they are telling the truth ($\psi(m_1) = \theta_1$) or the other message is the truth for the sender in state θ_2 ($\psi(m_2) = \theta_2$).

The two other possibilities — (1) $\psi(m_1) = \theta_2$, or (2) all messages are lies (ie. $\psi(m_1) \neq \theta_1, \theta_2$ and $\psi(m_2) \neq \theta_2$) are easily verified to fail graph monotonicity:

$$v_{\text{LA}}(m_1|\theta_2) - v_{\text{LA}}(m_2|\theta_2) \geq 0 \geq v_{\text{LA}}(m_1|\theta_1) - v_{\text{LA}}(m_2|\theta_1).$$

Proof of Proposition 5. (2)⇒(1) is easily verified by computation. It remains to show that (1)⇒(2), ie. if $v_{\text{LA}} \in \mathscr{M}_G(G(\sigma_0))$ then $G(\sigma_0)$ is one of the graphs in Figure 5.

We show the previous lemma rules out the subgraphs illustrated in Figures 6b, 6c, which constricts the communication graph's *half squares*.[30]

We use the assumption of generic u_R to assume that the graph is a tree and the receiver has a unique best response to each state. By injectivity and Assumption (S), this means at most one state can be disclosed in equilibrium. This implies that if $G^2[M]$ is not complete, G must have a subgraph of the form of Figure 6b.

The impossibility of Figure 6c shows that if three states send a message, then every state sends that message — knowing that there is at least one state that sends every message, this implies that $G^2[\Theta]$ is either a star or a complete graph.

The impossibility of these figures are obtained by repeated applications of Lemma 3:

Impossibility of Figure 6b: Applying the lemma to the 'N' spanned by θ_1-m_3 implies that $\psi(m_2) \neq \theta_2$.
With this knowledge, examining the 'N' spanned by θ_2-m_1, we find $\psi(m_1) = \theta_1$.
Revisiting θ_1-m_3, knowing that $\psi(m_2) \neq \theta_1$, we deduce $\psi(m_3) = \theta_2$.
But this makes graph monotonicity on the 'N' spanned by θ_3-m_2 impossible.

[29]A measure 0 of priors will also allow an equilibrium where mixed action p_P is recommended. These equilibria are fragile to perturbations to the *receiver*'s utility. Nevertheless the proposition extends to these equilibria.

[30]The half square $G^2[X]$ of a bipartite graph G contains all the vertices of one side X of the bipartition, and draws edges between vertices that share a neighbour in G.

Single-disclosure captures communication graphs whose half-squares $G^2[\Theta]$ and $G^2[M]$ are both complete.

Garbled-revelation describes communication graphs whose state half-square $G^2[\Theta]$ is a star and whose message half-square $G^2[M]$ is complete.

Impossibility of Figure 6c: for the two 'N' shapes spanned by θ_3-m_1 and θ_4-m_1 to hold, we must have $\psi(m_1) = \theta_2$ (since $\psi(m_2)$ cannot simultaneously be both θ_3 and θ_4).
But as before this contradicts the 'N' shape spanned by θ_1-m_2. □

Proof of Proposition 6. Let the equilibrium message inducing a belief ν correspond to the complement of its support $m = \psi^{-1}(\text{supp}(\nu)^c)$. Then for a path $(\theta_0, \pi_0, \ldots, \theta_N, \pi_N)$, we have

$$\begin{cases} v_{\text{WD}}(\pi_i|\theta_i) - v_{\text{WD}}(\pi_{i-1}|\theta_i) = 0 & i > 0 \\ v_{\text{WD}}(\pi_i|\theta_i) - v_{\text{WD}}(\pi_{i-1}|\theta_i) > 0 & i = 0. \end{cases} \quad \square$$

B Supplemental Material

B.1 Money Burning Example

We illustrate this technique in constructing equilibria of the money burning model:

Example 4. Consider the message space $M = M_1 \times M_2$ where M_1 denotes cheap talk messages, and $M_2 = \mathbb{N}_0$ denotes a discrete quantity of money burnt that accompanies the message. Suppose the sender's utility takes the form

$$u^0(a, (m_1, m_2)) = \tilde{u}(a) - \frac{m_2}{N}$$

for some N (representing currency increments). By assuming the currency space is discrete we maintain Assumption (S).

When N is large, this model allows us to attain the maximum set of candidate equilibrium:

Proposition 8. *Suppose $|M_1| \geq |\Theta|$, then for sufficiently large N, any Bayes-plausible set of posterior beliefs that induces at most one pure action corresponds to a money-burning equilibrium.*

However these equilibria cannot improve the sender's utility beyond the best cheap talk equilibria.

Note that this applies whenever persuasion is possible (ie. for some belief the receiver would prefer an action that is not a best response to the prior — for generic μ this is equivalent to the receiver not possessing a dominant action across all states). In particular, this permits communication even when the sender has monotonic preferences over the receiver's beliefs.

Proof. Let $\mathcal{B}$ be a plausible set of posterior beliefs, with $\mathcal{B}_+ \subseteq \mathcal{B}$ the set of posterior beliefs inducing a mixed action. We show how to design scheme of money-burning that satisfies sender-incentive compatibility. Let N be such that

$$\frac{2}{N} \leq \min_{\nu \in \mathcal{B}_+} \{\max \tilde{u}^*(\nu) - \min \tilde{u}^*(\nu)\}$$

where $\tilde{u}^*(\nu)$ is the range of utilities attained by a belief ν assuming no money is burnt. For convenience define $\underline{u} := \min_{\nu \in \mathcal{B}}\{\max \tilde{u}^*(\nu)\}$. Our analysis at this point diverges depending on where this minimizing belief is in $\mathcal{B}_+$.

The minimizing belief is in $\mathcal{B}_+$: If the minimizing belief induces a mixed action, then $\underline{u} - \frac{2}{N} \geq \min \tilde{u}^*(\nu)$. This establishes the bound of utility for eq. 4c.

For posteriors $\nu \in \mathcal{B}_+$ with $\min \tilde{u}(\nu) \leq \underline{u} - \frac{1}{N}$, we allow the belief to be induced without burning money. Otherwise, we require the sender to burn an amount $\frac{m_\nu}{N}$ so that

$$\underline{u} - \tfrac{2}{N} < \min \tilde{u}^*(\nu) - \tfrac{m_\nu}{N} \leq \underline{u} - \tfrac{1}{N}$$

Then $[\underline{u} - \frac{1}{N}, \underline{u}] \subseteq \tilde{u}^*(\nu) - \frac{m_\nu}{N}$ for all beliefs ν inducing a mixed action. For the message that induces the pure action a, we require burning an amount of money $\frac{m}{N}$ so that $\tilde{u}(a) - \frac{m}{N} \in [\underline{u} - \frac{1}{N}, \underline{u}]$. This gives us eq. 4b.

The minimizing belief is not in $\mathcal{B}_+$: If the minimizing belief leads to a pure action, then $\underline{u} \geq \min \tilde{u}^*(\nu)$. This establishes the bound of utility for eq. 4c.

For posteriors $\nu \in \mathcal{B}_+$, we require the sender to burn an amount $\frac{m_\nu}{N}$ so that

$$\underline{u} - \tfrac{1}{N} \leq \min \tilde{u}^*(\nu) - \tfrac{m_\nu}{N} < \underline{u}$$

ensuring eq. 4b is solved by $\underline{u}$.

Compared with cheap talk: To see that this cannot improve the sender's utility beyond the best cheap talk case, suppose a money-burning equilibrium improves on the utility generated by the prior. Utility is maximized when there is at least one message where money isn't burnt, attaining the candidate equilibrium $\overline{u}$. But then every message involving the burning of money must lead to a strictly higher utility. By garbling these messages to move their posteriors towards the prior, the intermediate value theorem ensures that at some point the posterior have a best response leading to sender utility $\overline{u}$, meaning no money need be burnt to send this message. □

This construction subsumes the cheap talk equilibria as the special case where $\max_{\nu \in \mathcal{B}} \min \tilde{u}^*(\nu) \leq \min_{\nu \in \mathcal{B}} \max \tilde{u}^*(\nu)$ and no burning of money is required.

An advantage of money-burning over cheap talk is that it allows persuasion when the sender's payoff is monotonic in the receiver's belief space. For example: a salesperson exists in two possible states: θ_a their product is good, θ_n their product is bad; and the receiver has two actions: A buy the product, or $\emptyset$ buy nothing. Persuasive cheap talk is impossible in such a model, but money-burning can be persuasive in convincing the receiver to purchase the product.

It is interesting to note that money burning with continuum quantities of money literally translates communication into mechanism design with linear transfers, where beliefs map to allocations whose utility is given by $u^*(\nu|\theta)$ with the feasibility constraint $\mathbb{E}[\nu] = \mu$.

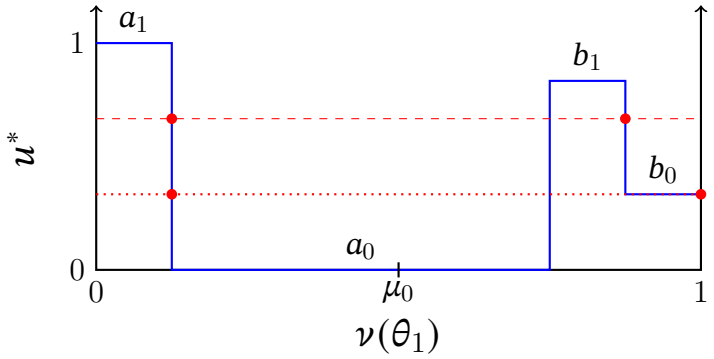


Figure 7: The indirect utility considered in Example 5.

B.2 Weak Robustness

Consider the following notion of stability:

Definition 10. We say a candidate equilibrium σ_0 is **weakly-stabilized** by a modification v if there exists Harsanyi-robust equilibria σ_ε to the game with preference $u + \varepsilon v$ such that $\sigma_\varepsilon \to \sigma_0$.

For contrast, we say a candidate equilibrium σ_0 is **strongly-stabilized** by a modification v if σ_0 is $\mathcal{O}$-robust for some $\mathcal{O} \ni v$.

If we decompose our sender utility as

$$u^0(\pi) + \varepsilon v(\pi|\theta) + U_\varepsilon(\pi|\theta, \omega)$$

then strong stability requires that there exists an N such that an equilibrium approximates σ_0 whenever $\|U_\varepsilon\|_\infty < N\varepsilon$ asymptotically. This inequality indicates the degree by which dependency must dominate idiosyncrasy for the desired communication to be an equilibrium.

Weak stability requires merely that there is some strictly positive function f such that an equilibrium approximates σ_0 whenever $\|U_\varepsilon\|_\infty < f(\varepsilon)$ asymptotically. Thus weak-stability may demand that state-dependency dominate idiosyncrasy by an arbitrary degree of magnitude.

We use the following example to illustrate the fundamental difference between candidate equilibria that can only be weakly stabilized and those that may be strongly stabilized:

Example 5 (Weak Stabilization)**.** Consider cheap talk with the indirect utility u^* illustrated in blue in Figure 7, similar to Example 2 of Steg et al. 2023. Utility will be normalized so that modifications only adjust the utility of b_1, b_0. There are three separate types of candidate equilibria we will analyze:

1. The dotted red line illustrates the unique candidate equilibrium that can be strongly stabilized, which involves a message m_b that induces action b_0, and a message m_a that induces the appropriate degree of randomization p_a between $[a_0, a_1]$.

2. There is also a range of candidate equilibria that can only be weakly stabilized —- an example is given by the dashed red line — that involve a message m_b that induces a randomization p_b strictly between $]b_0, b_1[$, and another message m_a that induces the corresponding degree of randomization p_a between $[a_0, a_1]$.

3. There are additional candidate equilibria cannot be even weakly stabilized — these are the candidate equilibria that involve the receiver randomizing between $[a_0, b_1]$ after one message, and $[a_0, a_1]$ or $[b_0, b_1]$ after another message.

In what follows, we label message-action pairs by their action for simplicity.

Candidate Equilibrium 1: The first situation is analyzed in the general case of Section 4 (as is the inability of the remaining equilibria to be strongly stabilized).

Candidate Equilibria 2: The second situation can be weakly stabilized in the following manner, illustrated in Figure 8: by shifting the utility of either b_0, b_1 — in this case b_1 — above the normalized line $[a_0, a_1]$ and moving the other below this line, we create an intersection point. This intersection represents a degree of randomization p_a, p_b between the two pairs of actions that is necessary to make the sender in each state indifferent between the messages m_a, m_b. Observe that this modification can be interpreted as making one message riskier in one state compared with another (in this case message m_b has a better best outcome and worse worst outcome in state θ_1 than in state θ_0).

Given a small degree of sender idiosyncrasy, we can then locally manipulate the two variables p_a, p_b, until we get the appropriate degree of mixing from the senders in each state (a two-dimensional problem), to produce the desired posteriors.

The reason this equilibrium is only weakly stabilized is that this degree of randomization p_a, p_b is highly sensitive to the modification. If we adjust the utility of $u^{\varepsilon v}(b_1|\theta_0)$ slightly (holding everything else constant), the intersection, and thus degree of randomization, will shift a proportionate amount. This constant of proportionality explodes as preferences approach transparency, where we are finding the intersection of nearly incident lines — unlike Candidate Equilibrium 1 (and strongly stable equilibria more generally) where the proportionality remains constant as $\varepsilon \to 0$.

Candidate Equilibria 3: The third family of candidate equilibria cannot even be weakly stabilized. This is because, in their situation, weak stabilization requires creating dual indifference between two mixed actions whose support only differ in one action. This is illustrated in Figure 9. To create dual indifference it is necessary that b_1 lies precisely on the normalized (state-independent) line $[a_0, a_1]$. That is, such communication requires state-independence, which we already know is fragile to idiosyncrasy.

Four types of complexity This example illustrates many of the complexities required and demanded by weak (but not strong) stabilization: (1) a higher degree of complexity in the equilibrium

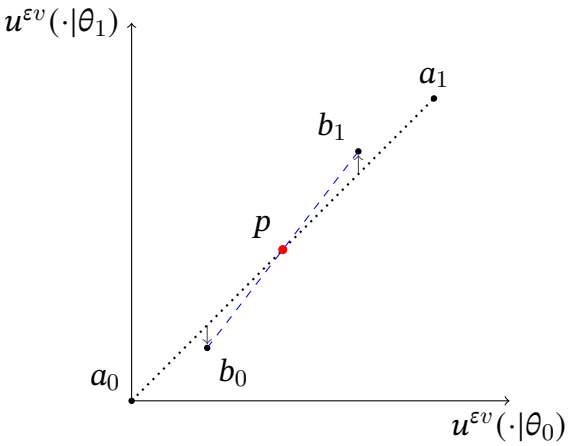


Figure 8: An illustration of a modification that can weakly stabilize the candidate equilibrium in Example 5. The axes are the utilities in the two states. The dotted line indicates the utility of mixed actions between $[a_0, a_1]$ (normalized to be state independent), while the dashed line indicates the state-dependent preference of mixed actions $[b_0, b_1]$ (the modification is illustrated by the small grey arrows). This creates a point of dual indifference p at the intersection of the two lines, determining the degree of receiver randomization after each message.

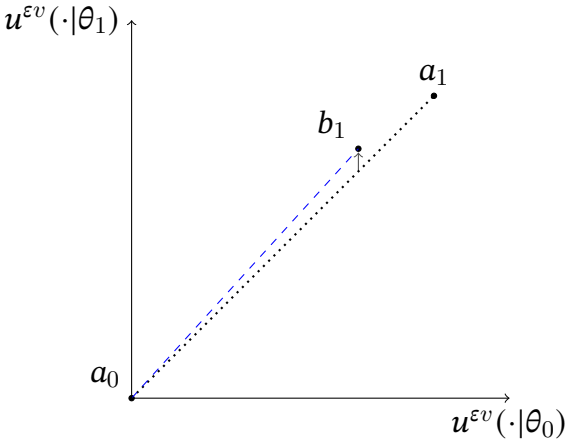


Figure 9: An illustration of how the third type of equilibria in Example 5 require state independence to create a point of dual indifference, and hence are inherently fragile. The dotted line is the state-independent utility from a_0, a_1 (determined by normalization), which $u^{\varepsilon v}(b_1|\theta)$ must lie on to create the point of dual indifference. Otherwise (as in the dashed line) they only intersect at the same action (a_0), and the action is message independent (ie. no persuasion occurs).

message structure, (2) a complex receiver environment, (3) complex modifications, and (4) (even small) uncertainty plays a role in which equilibria can be stabilized:

(1) Message Structure: The presence of cyclicity in the message structure demands more complex indifferences from the sender's perspective (and more strategic complexity) as well as being in a less informative family of equilibria from the receiver's perspective (acyclic equilibria are a more informative class of equilibria). In the above example the strongly stabilized equilibrium is strictly more Blackwell-informative than the weakly stabilized equilibrium.

(2) Receiver environment: As the third family shows, not all candidate equilibria can be weakly stabilized: weak stabilization requires the presence of additional actions, and will not differ from strong stability for equilibria where a single pure action is in the support of all equilibrium actions.

(3) Modification: Furthermore, the reason these additional actions are necessary is that the required modifications are slightly more complex: they require adjusting the 'riskiness' of a message in a state dependent way. This can be constrasted with strong stabilization where only the state-dependent *attractiveness* of equilibrium messages needs to modified. To observe the complexity of the required modifications, note that action-independent modifications can never produce this effect on riskiness.[31]

(4) Uncertainty: Lastly, this necessity of adjusting the riskiness means that pure actions are typically not present in weakly stabilized equilibria (as shown with the second family). The precise mixing required by an equilibrium is also highly sensitive to the *direction* of the sender's state dependence, unlike in strongly stabilized equilibria.

These points do not show that weak stabilization is impossible, but rather that it is distinct from strong stabilization and, in some informal sense, demands more from both the environment and the parties involved.

B.3 Relations with Mechanism Design

Our first order approximation very closely parallels the notion of mechanism design with linear transfers. In this section we will consistently use terminology from mechanism design, but maintain our notation from the body of the paper to make these parallels clear:

In mechanism design $\sigma : \Theta \rightrightarrows \Delta\mathcal{A}$ is an allocation correspondence, that maps types $\theta \in \Theta$ to allocations $\pi \in \Delta\mathcal{A}$. Each type θ has a preference $v(\cdot|\theta) : \Delta\mathcal{A} \to \mathbb{R}$.

The allocation map σ is *implementable* if there exists a transfer $T : \sigma(\Theta) \to \mathbb{R}_+$ such that

$$v(\pi|\theta) + T(\pi) \geq v(\pi'|\theta) + T(\pi') \qquad \text{for all } \pi \in \sigma(\theta), \theta \in \Theta, \pi' \in \sigma(\Theta)$$

A necessary and sufficient condition for implementability, related by Rochet 1987, closely relates to our notion of graph monotonicity:

[31] Empathy is also incapable of producing this effect in two-state environments, but may with more states.

Definition 11 (σ-Cyclic Monotonicity)**.** For a strategy profile σ, a utility v is **σ-cyclically monotone** if, for any sequence $(\theta_1, \ldots, \theta_{N-1}, \theta_N =: \theta_0)$ with $N \geq 2$, and any $\pi_i \in \sigma(\theta_i) \setminus \{\pi_{i-1}\}$ we have

$$\sum_{i=1}^{N} v(\pi_i|\theta_i) - v(\pi_{i-1}|\theta_i) \geq 0. \tag{20}$$

If the inequality is strict for all paths then v is **strictly σ-cyclic-monotone** (or $v \in \mathscr{M}(\sigma)$).

The conditions $N \geq 2$ and $\pi_i \neq \pi_{i-1}$ are to avoid trivial equations.

To parallel our definition of graph monotonicity, we interpret this as a constraint on preferences with the allocation as fixed, whereas the standard definition takes eq. 20 as a constraint on allocations with preferences fixed.

This differs from graph monotonicity in that the sequence does not have to follow chains of indifferences. As a result this is a stronger condition. However in the situations that we consider in this paper, the notions coincide:

Proposition 9. *In general $\mathscr{M}(\sigma) \subseteq \mathscr{M}_G(G(\sigma))$, with equality iff one of the following holds:*

1. *$G(\sigma)$ is cyclic, in which case the sets are empty.*
2. *$G(\sigma)$ is a tree, in which case the sets are non-empty.*

The first property immediately follows from $\mathscr{M}_G(G(\sigma))$ being empty in such situations, as shown in Proposition 2.

The second property results from the inequalities of eq. 9 forming a basis for the inequalities of eq. 20 when $G(\sigma)$ is connected.

A slight modification to our proof of Theorem 3 will show that cyclic monotonicity is a necessary condition for sender $\mathcal{O}$-robustness, as such it is a 'tighter' condition than graph monotonicity. However, in light of Theorem 2 showing that the situations where these sets differ are fragile and rare, we opt for the simpler concept, which is more relevant to our proof strategies. This is further justified by Appendix B.4, where we find that cyclic monotonicity is no longer relevant for $\mathcal{O}$-robustness of non-injective equilibria, nevertheless we can still extend our techniques to this setting.

Shortest Paths and Incentive Graphs

Suppose the allocation map σ is single-valued. In this case Vohra 2011 proposes an alternate graph-theoretic interpretation of eq. 20, as an *incentive graph*.

Definition 12A (Incentive Graphs)**.** *The **transfer incentive graph** $G_T(\sigma; v)$ associated with an allocation map σ and utility v is the directed weighted graph with nodes Θ. Between any pair of nodes θ_0, θ_1 there is a directed edge from θ_0 to θ_1 with weight $v(\sigma(\theta_1)|\theta_1) - v(\sigma(\theta_0)|\theta_1)$.*

We can see that v is σ-cyclic monotone iff every cycle on $G_T(\sigma; v)$ has weakly positive length — where length of a directed path is equal to the sum of weights of the edges on the path. This graph has the property that the linear transfer map T necessary to make the allocation map σ incentive compatible solves

$$T(\sigma(\theta_1)) - T(\sigma(\theta_0)) = d_{G_T}(\theta, \theta_0) + C$$

where $d(\theta, \theta_0)$ is the length of the shortest path between θ and θ_0.

Transfer incentive graphs cannot be defined if σ is multi-valued. Instead we modify the definition by identifying edges with actions — this requires a *multi-graph* (ie. a graph allowing multiple edges between nodes):

Definition 12B (Utility Incentive Graph)**.** *The **utility incentive graph** $G_U(\sigma; v)$ associated with an allocation map σ and utility v is the directed weighted multi-graph with nodes Θ. The set of edges from θ_0 to θ_1 is $E(\theta_0, \theta_1) := \{\pi_0; \pi_0 \in \sigma(\theta_0)\}$ where the weight associated with the edge π_0 is $v(\pi_0|\theta_0) - v(\pi_0|\theta_1)$.*

As before, the utility v is σ-cyclic monotonicity iff every cycle on $G_U(\sigma; v)$ has weakly positive length. Unlike transfer incentive graphs, the shortest path between two nodes defines the difference in utility obtained by these two states:

$$[v(\pi_1|\theta_1) + T(\pi_1)] - [v(\pi_0|\theta_0) + T(\pi_0)] = d_{G_U}(\theta_1, \theta_0)$$

Implementability can also be verified by seeing how the communication graph $G(\sigma)$ embeds in $G_U(\sigma; v)$. Formally, for a communication graph define the embedding $\text{proj}_U(G; v)$ to be the graph with nodes Θ and edges from θ_0 to θ_1 given by the allocations $\{\pi \in \sigma(\theta_0) \cap \sigma(\theta_1)\}$ with weight $v(\pi|\theta_0) - v(\pi|\theta_1)$.

This takes the communication graph and maps the allocations π into edges.[32] Note that $\text{proj}_U(G(\sigma); v) \subseteq G_U(\sigma; v)$. The embedding $\text{proj}_U(G(\sigma); v)$ may contain a cycle even if $G(\sigma)$ is acyclic, — if more than two types are mapped to the same allocation — but all cycles will have length zero, and the embedding will be a graph, with at most one edge from one vertex to another. One last key property is that edges are anti-symmetric, in that the weight on the edge from θ_0 to θ_1 is the negative of the weight on the edge from θ_1 to θ_0.

Proposition 10. *Let σ be an allocation map such that $G(\sigma)$ is connected.*

1. *The utility v is σ-cyclic monotone for the connected communication graph $G(\sigma)$ iff every path on $proj_U(G(\sigma); v)$ is a shortest path on $G_U(\sigma; v)$.*

2. *The utility v is strictly σ-graph monotone iff $G(\sigma)$ is acyclic and the paths on $proj_U(G(\sigma); v)$ are precisely the shortest paths on $G_U(\sigma; v)$.*

[32] This essentially results in the half-square graph $G^2(\sigma)[\Theta]$, with bidirectional weighted edges.

Proof. Let $(\theta_1, \ldots, \theta_N)$ be a path in the embedding with length L. Suppose there is a shorter path in $G_U(\sigma; v)$, WLOG from θ_1 to θ_N with length $L' < L$. Then the cycle from $(\theta_N, \ldots, \theta_1, \theta_N)$ along these paths has length $L' - L < 0$. Moreover, every cycle in $G_U(\sigma; v)$ can be decomposed into the sum of cycles with all but one edge in the embedded graph (similar to Proposition 9), so if all these cycles are weakly postive, then all cycles on $G_U(\sigma; v)$ will be weakly positive.

To obtain strict σ-graph monotonicity, every cycle on $G_U(\sigma; v)$ with non-positive length must correspond to a single action. This can fail if there is an equally short path that is not included, following the above logic with weak inequality, or if $G(\sigma)$ has a cycle. If $G(\sigma)$ has no cycles, then all cycles on in the embedding are connected by a single action (and exempt from the strict cyclic monotonicity definition), this shows that strict cyclic monotonicity applies for cycles contained withing the embedded graph. Cycles not entirely contained within the embedded graph can be decomposed into cycles with at most one edge outside the connected graph, showing that these conditions are equivalent. □

B.4 $\mathcal{O}$-Robustness of Non-Injective Equilibria

We use injectivity three times in our results:

1. To assume eq. 4c does not bind.
   This affects our analysis: when studying the worst candidate equilibria from the sender's perspective. In at least one state they will attain their minimum equilibrium payoff, in other states they do at most $O(\varepsilon)$ better.
   Required Adjustment: perturbations must be constrained to ensure off-path messages are unattractive.

2. For Lemma 1ii and Theorem 2 to apply. The former ensures that each connected component includes at most one pure action, the latter ensures there is a unique connected component in acyclic equilibrium.
   When this affects our analysis: it turns out that Lemma 1ii is necessary for robustness. However there may be multiple connected components, each rooted at a different message with the same pure action component. This is possible only if the receiver has the same best response to two disjoint beliefs, ie. there exist ν, ν' with $\text{supp}(\nu) \cap \text{supp}(\nu') = \emptyset$ and $a^*(\nu) = a^*(\nu')$.
   Required Adjustment: the perturbation must be adjusted to make comparisons on different trees.

3. So the receiver never chooses a pure action at a posterior where they have multiple best responses (when Lemma 1iia applies).
   Required Adjustment: the perturbation must shift the sender's utility in a direction that can be compensated by receiver randomization (which now can only move in one direction).

Of course it is also possible that multiple reasons may hold at once, which requires combining the constraints we discuss below. We leave our reasoning informal, the same proof techniques from Theorem 3 can be used to formally prove these results.

1. Off-path Temptations

In our proofs we can prescribe the receiver responds to off-path messages with the least attractive action to the sender: $\mathcal{P}(m) \in \arg\min_a\{u^0(a,m)\}$ whenever $m \notin \text{supp}(\mathcal{M})$. However, if eq. 4c binds, this action may not decrease the utility of the sender with transparent utility u^0, and thus may still tempt senders with perturbed utility $u_S^{\varepsilon v}$. This occurs for example, if an equilibrium requires a lying averse sender to lie and gives them the least preferred action — this is worse than the utility the sender can obtain by telling the truth and obtaining their least preferred action.

Working in the generic case of Lemma 1, we know that these equilibria involve a pure message-action α, which thus must maximinimize the right side of eq. 4c. Ie., $\alpha = (\underline{a}(\overline{m}), \overline{m})$ where

$$\underline{a}(m) \in \arg\min_{a\in\mathcal{A}} u^0(a,m) \qquad\qquad \overline{m} \in \arg\max_{m\in M} u^0(a(m),m).$$

If there is a unique maximizing message $\overline{m}$, then eq. 2a is obtained by verifying that each state prefers their message-action to the message-action α corresponding to $\overline{m}$. This is done in the main theorem, and requires no additional analysis.

However, if there are multiple maximizing messages then different states may be tempted by different messages, resulting in an additional constraint. Note that Assumption (S) implies that $\underline{a}(m) \equiv a^*$ is then constant over these messages. Define

$$\alpha(\theta) :=(a^*, m(\theta)) \qquad\qquad m(\theta) \in \arg\max_{m\in\underline{a}^{-1}(a^*)} v(a^*, m|\theta)$$

to be the most tempting off-path messages for a state θ.

To obtain eq. 2a, it is then necessary that for a path $(\theta_1, \pi_1, \ldots, \theta_N, \pi_N = \alpha)$ from the state θ_1 to the pure root action α the sender in state θ_1 prefers π_1 to $\alpha(\theta_1)$. Equation 7 shows that the equilibrium perturbation to p_1 must solve

$$\left\langle (u^0_\alpha), \Delta p_1 \right\rangle = \varepsilon \sum_{i=2}^{N} v(\pi_i|\theta_i) - v(\pi_{i-1}|\theta_i) + O(\varepsilon^2). \tag{21}$$

in order to preserve indifferences.

Thus for the off-path message $\alpha(\theta_1)$ to be suboptimal, it is necessary that

$$v(\pi_1|\theta_1) - v(\alpha(\theta_1)|\theta_1) + \sum_{i=2}^{N} v(\pi_i|\theta_i) - v(\pi_{i-1}|\theta_i) > 0 \tag{22}$$

This is just an alteration to the definition of graph monotonicity so that $\pi_0 := \alpha(\theta_1)$ — by the definition of $\alpha(\theta_1)$ this results in a stronger inequality. In contrast to Graph Monotonicity, this must also hold for $N = 1$ to ensure that states that recommend α still attain their lowest utility (in this case the sum is empty).

In lying aversion this means that only m_* and m^* can result in this minimal action (otherwise multiple actions send this minimal action and eq. 22 with $N = 1$ fails). The m^* case is always stabilized by lying aversion, while m_* requires that $\ell_{\theta_*} > \ell_\theta$ for all $\theta \neq \psi(m_\emptyset)$.

2. Two messages with same pure best response

Suppose there are two equilibrium posteriors for which the pure action a^* is a unique best response — corresponding to the message-actions α, α'. If θ, θ' send message α, α' respectively, then it is necessary that $v(\alpha|\theta) > v(\alpha'|\theta)$ to ensure that state θ will not deviate to send α'.

Same connected component Suppose α, α' lie on the same connected component of $G(\sigma_0)$. If they are the unique actions on this connected component then there is a state θ that randomizes between the two messages. To retain indifference it must be the case that

$$v(\alpha|\theta) = v(\alpha'|\theta)$$

which does not hold on any open set of modifications.

If there is a message π_1 on this connected component to which a^* is not the unique best response, then there are two paths on $G(\sigma_0)$ for which eq. 21 must hold. In particular, there are paths $(\pi_1, \theta_2, \ldots, \theta_N, \pi_N)$ and $(\pi_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_M, \tilde{\pi}_M)$ on $G(\sigma_0)$ with $\pi_N = \alpha$, and $\pi_M = \alpha'$. It is then necessary that

$$\sum_{i=2}^{N} v(\pi_i|\theta_i) - v(\pi_{i-1}|\theta_i) = \sum_{i=2}^{N} v(\tilde{\pi}_i|\theta_i) - v(\tilde{\pi}_{i-1}|\theta_i)$$

where $\tilde{\pi}_1 = \pi_1$. Since this cannot be satisfied on an open set of modifications, this cannot be $\mathcal{O}$-stable — this overdetermines Δp_1.

Different Connected Components In this case the negative result from Theorem 3 applies, allowing us to restrict our attention to forest equilibria. Lemma 1 implies that (generically) every connected component of the communication graphs includes a pure action (as there is a posterior with unique best response). By Assumption (S) these must all be the same pure action.

Graph monotonicity ensures that no state will deviate to send a message in the same connected component, it remains to check that states will not deviate to send messages in different connected components.

Consider a path $(\theta_1, \pi_1, \ldots, \theta_M, \pi_M)$ on $G(\sigma_0)$ from θ_1 to the pure action $\pi_M \equiv \alpha$ in its component, and a path $(\pi'_M, \theta_{M+1}, \ldots, \theta_N, \pi'_N)$ from the pure action $\pi'_M \equiv \alpha'$ in π'_N's component to π'_N. Applying eq. 21, we require

$$v(\pi_1|\theta_1) + \sum_{i=2}^{M} v(\pi_i|\theta_i) - v(\pi_{i-1}|\theta_i) > v(\pi'_N|\theta_1) + \sum_{i=M+1}^{N} v(\pi'_{i-1}|\theta_i) - v(\pi'_i|\theta_i)$$

or, defining $\pi_m := \pi'_m$ for $m > M$:

$$v(\pi_M|\theta_{M+1}) - v(\pi'_M|\theta_{M+1}) + \sum_{i=1}^{N} v(\pi_i|\theta_i) - v(\pi_{i-1}|\theta_i) > 0. \tag{23}$$

where $\theta_{M+1} := \theta_1$ if $N = M$. Note that the first two terms correspond to substituting $v(\alpha'|\theta_{M+1})$ for $v(\alpha|\theta_{M+1})$. Since these have the same action component a^*, this does not affect cheap talk models.

This is equivalent to graph monotonicity on the communication graph that merges messages that result in the same pure action into a single message action α^* with

$$v(\alpha^*|\theta_i) := v(\alpha_i|\theta_i)$$

for the unique pure α_i on the connected component containing θ_i.

This situation requires that the pure action a^* appears on distinct components of the communication graph, so a^* must be a best response to disjoint beliefs (ν, ν' with $\mathrm{supp}(\nu) \cap \mathrm{supp}(\nu') = \emptyset$).

In lying aversion forest equilibria provide states another way to recommend a pure action a^*: by disclosing the state. To be stabilized by v_{LA} each message that results in a^* must be truthful, and thus have a single state sending it.

If we have a forest equilibrium then by Proposition 5 each connected component of the communication graph must resemble either Figure 5ab (with m_*, m^* inducing the pure action). Suppose two states θ_1, θ_2 recommend a^* through truthful disclosure, and θ_2 also recommends another action π, then the path from θ_1 to π gives us the constraint

$$\ell_2 - \ell_1 > 0.$$

Note that if θ_1 also recommends another action, we obtain the reverse inequality: at most one state can send multiple messages. In this case our communication graph is a subgraph of Figure 5ab, where θ_*, θ^* must be the state that minimizes $-\ell_\theta$ from all those that recommend the pure action a^*.

3. A pure action is non-unique best response Let α_1 be this action. In this case, Lemma 1' implies that (generically) there is another pure action α_N on the same connected component that is

a unique best response to its associated posterior (by Assumption (S) α_1, α_N have the same action component, but differ in their message).

Suppose that α_1' is the receiver's other best response (unique by Lemma 1) to the posterior associated with α_1 satisfies $u^0(\alpha_1') > u^0(\alpha_1)$. Applying eq. 21, we find the necessary condition that for the path $(\pi_1, \theta_2, \ldots, \theta_N, \pi_N)$ with $\pi_N = \alpha_N$ we have

$$\sum_{i=2}^{N} v(\pi_i|\theta_i) - v(\pi_{i-1}|\theta_i) > 0$$

in addition to the other required inequalities. If $u^0(\alpha_1') < u^0(\alpha_1)$, we require the opposite inequality.

In lying aversion graph monotonicity requires a subgraph of Figure 5ab. In this case the state $\theta_0 := \theta^*$, or θ_* recommends a^* through two messages: one where it is the unique best response, and the other where it is not. If both of these messages are lies (ie. not $m_0 := \psi^{-1}(\theta_0)$), then the equilibrium is not stabilized by lying aversion, if one is a lie, it necessary that this message can be made relatively more attractive through changing the action. Thus if m_0 induces the belief where multiple actions are best responses, then it must be the case that $u^0(\alpha_1') < u^0(\alpha_1)$, otherwise it is necessary (and sufficient) that $u^0(\alpha_1') < u^0(\alpha_1)$.