DYNAMICS OF COVERT SIGNALING: MODELING THE EMERGENCE AND EXTINCTION OF IDENTITY SIGNALS

ZACKARY OKUN DUNIVIN[1,*] AND PAUL E. SMALDINO[2,3]

August 28, 2024

Abstract: Covert identity signals permit the communication of group membership to ingroup members while avoiding potentially costly detection by members of other groups. If individuals are incentivized to detect others' group memberships, however, covert signals may not remain covert for very long. We propose a theoretical extension to the literature on covert signaling in which conventionalized identity signals can become destabilized when learned by outgroup individuals, to be replaced by the emergence of new signaling conventions. We formalize this idea with both analytical and agent-based modeling of ingroup and outgroup individuals who learn about signals of group membership. Depending on the risk and associated cost of detection by the outgroup, the model yields three dynamic classes: saturation, where all identity signals become stable conventions and never go extinct; cycling, in which new signals emerge to replace old ones as they are learned by the outgroup; and suppression, in which informative identity signals never emerge. Our analysis has implications for understanding identity signaling, the emergence of conventions, coded speech, and the ebb and flow of fashion cycles.

Keywords: covert signaling; social identity; cycles; cultural evolution; agent-based model

[1]Indiana University
[2]University of California, Merced
[3]Santa Fe Institute
E-mail address: zdunivin@iu.edu.
[0]Acknowledgments omitted for review.
[0]This work was supported by ARO Grant W911NF-20-1-0220. An earlier version of this article including associated data and code were posted as a preprint: Dunivin, Z. O., & Smaldino, P. E. (2023, February 17). Dynamics of covert signaling: Modeling the emergence and extinction of identity signals. PsyArXiv https://doi.org/10.31234/osf.io/3tz2r .

1. Introduction

When interacting with other people they don't know very well, people often broadcast signals to establish the sort of person they are or aren't. These identity signals can be linguistic, sartorial, or even behavioral: they serve to inform audiences about what the sort of person the signaler is or is not (Berger and Heath, 2008; Smaldino, 2019). The identities being signaled may concern the signaler's role in society, their membership in a particular group, or even their individual behavioral characteristics (Burke and Stets, 2009). For example, a person may communicate their identity as a mother (role), a democratic socialist (group membership), or someone with a morbid sense of humor (individual behavior). Identities serve functions far beyond establishing the goals and psychological well-being of those that hold them, though these are surely important. Identity signals are used instrumentally to facilitate social assortment (Smaldino, 2019, 2022). They provide audiences with information with which to evaluate the likelihood of success of cooperative interactions (including economic or romantic interactions) and about potential dangers of direct engagement. For example, numerous studies have shown that altruistic behaviors are preferentially directed toward individuals perceived as ingroup members (Chen and Li, 2009; Henrich and Muthukrishna, 2021). The signaling function of how we present ourselves in public has long been appreciated by social scientists (Goffman, 1978; Barth, 1969; Donath, 1999; Berger and Heath, 2008; Wimmer, 2008).

At their core, identity signals serve a key social function by enabling individuals to rapidly characterize others as similar or dissimilar. In the politically polarized United States, signals have emerged to identify individuals as belonging to the political right or left (Urbatsch, 2014; Sloman et al., 2021; Powell et al., 2023). Some of these signals are directly connected to the signaler's political views, such as bumper stickers declaring support for particular candidates or polarizing social issues (e.g., "Abortion Is Healthcare"

or "Gun Control Means Using Both Hands"). Using one of these signals requires minimal common knowledge between the signaler and receiver, because the signal contains explicit information about its intended meaning. Other signals, however, are seemingly arbitrary and become conventionalized as identity signals only over time. Some of these may reflect the phenomenon of "lifestyle clustering" so that stereotypes about "latte-drinking liberals" and "bird-hunting conservatives" can emerge via the tendencies of liberals to live in cosmopolitan urban centers more likely to have upscale coffee shops and of conservatives to live more in rural areas where hunting is more accessible (McPherson, 2004; DellaPosta et al., 2015). The correlation between identity and lifestyle need not be strong, however, for signals to become conventionalized. Signaling conventions can emerge entirely through the amplification of small differences when assortment with similar individuals is beneficial. Numerous behavioral experiments and formal models have shown how incentives for coordination can facilitate strong correlations between observable signals and unobservable characteristics when no such associations existed in the initial population (Boyd and Richerson, 1987; Nettle and Dunbar, 1997; McElreath et al., 2003; Castro and Toro, 2007; Efferson et al., 2008; Puglisi et al., 2008; Cohen and Haun, 2013; Centola and Baronchelli, 2015; Bell and Paegle, 2021; Guilbeault et al., 2021).

If successfully signaling to similar others is incentivized and, as is usually assumed, signaling to dissimilar others is not penalized, then signaling conventions can not only emerge, but stabilize. In other words, a prediction that follows from the logic of most models of convention is that, once established, a convention will remain conventional. However, this prediction is not aligned with many cases of signaling conventions, in which signal trends rise and then wane in popularity (Berger and Heath, 2008). Fashions may simply go out of style. However, another reason is that signaling one's identity to dissimilar others may in fact be costly. If revealing one's otherwise-hidden identity to members of an outgroup entails costs—such as those faced by members of certain ethnic

minorities and religious groups, political dissidents, and LGBTQ+ individuals—then overt identification may not be worth the risk. In such cases, covert identity signals may arise.

Covert signals are accurately received by their intended audience but obscured when received by others (Smaldino et al., 2018; Smaldino and Turner, 2022). They allow individuals who share social traits to recognize one another while simultaneously allowing signalers to avoid being recognized as dissimilar by those not "in the know." Political dog whistling is perhaps the most widely known example of covert signaling, in which speakers will make references that are interpreted as innocuous by most listeners but signal more controversial commitments to insiders (Henderson and McCready, 2017). For example, former US President George W. Bush regularly decried the 1857 Supreme Court decision that denied the freed slave Dred Scott's right to file suit, tapping into the connections that conservative Evangelicals at the time made between that decision and the 1973 Roe vs. Wade decision that until recently upheld the right to legal abortion (Kirkpatrick, 2004). While some dog whistles are identified after the fact (as in the case just mentioned), their prevalence indicates that many if not most go largely undetected, even if little research has investigated the efficacy of dog whistles in remaining covert. More quotidian examples of covert signaling abound concerning the ways people implement fashion, humor, and other semiotic tools to subtly indicate identity (Berger and Ward, 2010; Flamson and Bryant, 2013; Fischer, 2015). These signals are probably less easily detected than more overt signals, but trade clarity for the benefits of encryption or plausible deniability (Lee and Pinker, 2010). Covert signaling may be particularly important to members of persecuted minorities, such as LGBTQ+ individuals or political dissidents, who have strong incentives to assort with one another but also to avoid detection by nonmembers.

Modeling work has formalized this idea (Smaldino et al., 2018; Smaldino and Turner, 2022), indicating that covert signals should be favored over overt identity signals when

being revealed as dissimilar is costly and when individuals cannot count on being able to partner only with those they prefer. These conditions are more likely to be met in more diverse societies and among those with minority-group status. A recent empirical study provides explicit support for the theory in the context of political identity signaling online, showing that Twitter users with more heterogeneous follower networks tweeted more covertly and that participants in a behavioral experiment strategically selected more covert signals when their audience consisted of more outgroup members (van der Does et al., 2022).

Covert signals work because they are known to insiders but not to outsiders. At minimum, they must be substantially less reliable as signals of identity when received by outsiders. However, the information content of a signal is not fixed. As noted above, signaling conventions can emerge dynamically as people learn to associate particular signals with particular identities. When the incentives of the signalers and the receivers are sufficiently aligned, these conventions can become stabilized and even institutionalized. When audiences are antagonistic, deceptive signals are incentivized (Crawford and Sobel, 1982). Covert signals occupy an in-between space (Smaldino and Turner, 2022), conveying honest messages to ingroup members (whose interests are aligned with those of the signaler) while deceiving outgroup members (whose interests are not aligned). Outgroup audience members may have incentives to avoid being deceived, however, and to correctly identify outgroup individuals despite their efforts to signal covertly. The police may wish to arrest dissidents, and an employer may wish to avoid hiring someone with divergent or unconventional views. The effectiveness of covert signals may therefore be more ephemeral when compared with overt signals. Once a particular signal has outlived its usefulness, new signals can arise to take its place.

The presence of covert signals provides the conditions for dynamic cycles of signaling conventions. Previous models of covert signaling focused on competition between strategies of covert vs. overt signaling, and did not explicitly consider the dynamic usage of

specific signals, though prior work has speculated about the possibility of such cycles (Smaldino et al., 2018; Smaldino and Turner, 2022; van der Does et al., 2022). Here, we explore this idea more extensively.

Our proposal works as follows. In any sufficiently large population in which prior knowledge of interaction partners is not guaranteed, arbitrary signals can become reliable markers of identity when individuals learn to associate particular signals with particular identities and choose their own signals accordingly. If accidentally revealing one's identity to an outgroup individual is either very unlikely or does not carry particularly high costs, a signal can become conventionalized as a stable marker of identity despite being known to both ingroup and outgroup audiences, i.e., it becomes an overt signal. If, however, becoming known to outgroup audiences carries sufficient risk, a signal may lose its value once it is in regular use by an ingroup, as the outgroup comes to associate the signal with the ingroup. In this case, the usage frequency of a dominant covert signal may decrease, and a new covert signal may rise to dominance. This dynamic should repeat indefinitely as long as other inter- and intra-group relationships remain relatively constant. In extreme cases where detection by the outgroup is both sufficiently likely and sufficiently costly, it is possible that no reliable covert signals will ever emerge, and group members will have to rely on other means to assort. In the subsequent sections of this paper, we demonstrate the plausibility of the proposal, and examine conditions for the emergence of stable signals, cycles, and the total absence of reliable signals.

Our proposal for the emergence of signaling cycles follows a logic similar to mechanisms generating cyclical group dynamics in other systems. In the social sciences, perhaps the best known is Simmel's (1904) theory of fashion. He proposed that the function of many fashion trends, which clearly serve little utilitarian purpose, is to distinguish members of the elite from members of the lower social classes. Members of the lower classes, however, are incentivized to appear more upper class, and therefore strive to copy the fashions of the elite. This in turn incentivizes the elite classes to continually innovate

new fashion trends so as to stay above the rabble. Although this idea cannot explain all fashion cycles, later work has expanded upon Simmel's theory, finding some empirical support (Miller et al., 1993; Krawczyk et al., 2014) and using formal modeling to explore specific conditions for the emergence and character of fashion cycles (Pedone and Conte, 2001; Acerbi et al., 2012; Di Giovinazzo and Naimzada, 2015). Our proposal is similar to Simmel's, but differs by virtue of the differential effect signals have on ingroup and outgroup audiences and in the use of a particular signal by only one group. We focus on scenarios in which members of a disadvantaged or persecuted group are trying to identify each other while avoiding detection by a hostile outgroup—this latter group is trying to punish rather than imitate them.

Cyclical dynamics can arise in any coupled two-component system in which the first component activates growth in both itself and the other component, and the second component inhibits growth in both itself and the other component (Figure 1). Predator-prey systems can famously exhibit these cycles, typified by the Hudson's Bay Company data on Canadian lynx and snowshoe hare populations and formalized by the classic Lotka-Volterra model (Smaldino, 2023). In this model, the prey population grows in the absence of predators and also stimulates growth in the predator population, while the predator population shrinks in the absence of prey and also inhibits growth in the prey population. When the two populations are present simultaneously, coupled cycles of growth and decline emerge, though some parameter combinations can also lead to the collapse of one or both populations. A range of other systems exhibit cycles for similar reasons, including endoparasites (Otto and Day, 2011), neural firing rates (Stiefel and Ermentrout, 2016), and even the rise and fall of empires (Turchin, 2003).

Below, we present a formal investigations of the verbal theory presented above, using an agent-based model. In Appendix A, we also present a simple analytical model adapted from a "matching alleles" host-pathogen model (Parker, 1994; Sardanyés and Solé, 2008), in which a host evolves genetic resistance to multiple pathogens. In this

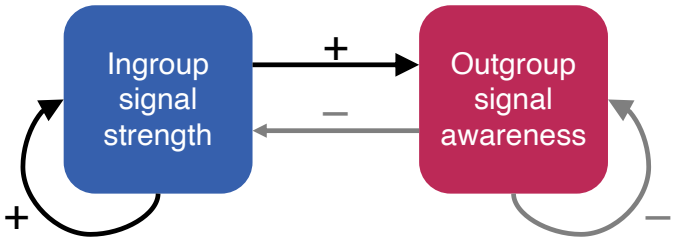


FIGURE 1. Cyclical dynamics can arise in any coupled two-component system in which the first component activates growth in both itself and the other component, and the second component inhibits growth in both itself and the other component. Here we illustrate how cycles can emerge in the usage of an ingroup signal through exposure to a hostile outgroup.

model, a population of ingroup signals changes in prevalence over time in response to growing outgroup awareness of those signals. Ingroup signal strengths and outgroup awareness of those signals are treated as a system of $2K$ coupled differential equations, where K is the maximum number of signals that can be simultaneously active in the population. This model is able to capture the broad strokes of our verbal theory, and serves to illustrate how many of the dynamics we propose may arise. However, the analytical model is also limited in its ability to capture important features of human cognition and communication due to the nature of the mathematical formalism. For example, individual decisions are not represented. Thus, the main body of this article focuses on the agent-based model, in which agents use reinforcement learning to make decisions concerning the costs and benefits associated with the use of various signals. We view this model as more realistic both in terms of its social structure and in terms of the psychological mechanisms of communication and learning implemented. The findings from the agent-based model provide support for all aspects of the verbal theory.

2. The Agent-based model

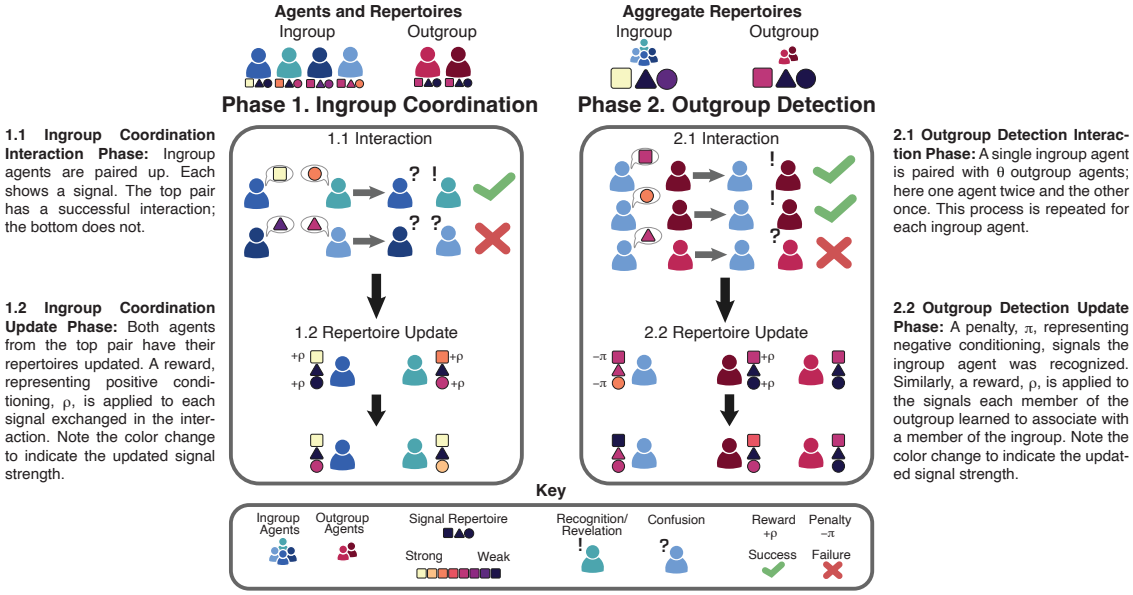
We consider a population of agents that interact with others agents both from their own ingroup and from an outgroup. As in the mathematical model, we focus on one particular group and its members' relationships with both ingroup and outgroup partners. Each interaction is an opportunity for identity signaling in which they can potentially identify a fellow ingroup member. Signals are arbitrary, and therefore the association between a particular signal and its meaning confirming the sender as an ingroup member must be learned. When signals become established as reliable indicators of ingroup status, they help members find each other and receive the benefits of cooperation and coordination. When an agent receives a signal it believes communicates ingroup status, it may take a risk and overtly declares its identity status to its partner in the hopes of initiating a successful coordination. This sort of declaration is necessary to confirm similarity and receive benefits. The use of arbitrary signals is both methodologically convenient and theoretically motivated. From a practical standpoint, the model represents the simplest possible system, and so arbitrary signals are used in principle, making them the standard for both signaling and evolutionary (genetic/phenotypic) models. Work in semiotic theory also shows that many symbols are fairly arbitrary and gain meaning only in reference to their communicative association with aspects of reality and with other symbols (Tylén et al., 2013). In this sense the model's arbitrary signals are realistic, as meaning emerges through association of symbols with group identity. That said, many signals operate through contextual and referential indicators, and these aspects are not captured by our model.

Outgroup individuals also learn to associate certain signals with group membership. If an outgroup member receives a signal from an ingroup member and chooses to identify that agent as a member of the ingroup, the outgroup agent learns that the focal agent's signal is associated with ingroup membership. If a subsequent ingroup member uses this signal with that outgroup member, the outgroup member is more likely to recognize

its ingroup status and identify the agent as a member of the ingroup. Depending on whether this outcome is costly, the ingroup agent may become less likely to use the same signal in the future. Note that for convenience, we consistently refer to the group that is motivated to signal covertly as the ingroup and to the other group as the outgroup, though of course in reality individuals will typically conceptualize members of their own group as "ingroup" and members of other groups as "outgroup."

The model dynamics proceed in discrete time steps, each of which consists of five phases: an ingroup coordination phase, the first signal repertoire update phase, an outgroup detection phase, the second signal repertoire update phase and a signal extinction/emergence phase. A simplified schematic is given by Figure 2, which omits the extinction/emergence phase. We code distinct phases for ingroup and outgroup signaling largely to simplify the code and model description. In a real system the processes in all these phases would occur concurrently. Simulations run with the two phases interspersed (not reported here) confirmed our prediction that this modeling decision did not qualitatively alter the model dynamics or outcomes. Due to the complexity of the model, we describe the initialization conditions as we introduce each parameter. A full list of model parameters and their default values is shown in Table 1. Python code for the model dynamics and analysis is available at `https://osf.io/4vcug/?view_only=14c6e87c32f146a6910a24e7dd079191`.

2.1. **Ingroup coordination interaction phase.** We consider a population of N_{IN} agents, all members of a group, which we will refer to as the ingroup. Each ingroup agent i is characterized by a *repertoire* of K identity signals, defined as a vector of signal weights $\mathbf{S}_i^{\text{IN}} = \{s_{i1}^{\text{IN}}, s_{i2}^{\text{IN}}, \ldots, s_{iK}^{\text{IN}}\}$. Each signal weight s_{ik}^{IN} represents the informational value of signal k to a particular agent i. Each signal in the repertoire of each ingroup agent i is bounded in [0, 1] so that the theoretical maximum of $\sum_k^K s_{ik}^{\text{IN}} = K$. Upon initialization of the model, every ingroup agent is instantiated with an identical signaling repertoire, such that $\forall(i,k), s_{ik}^{\text{IN}} = 0.1$.



1.1 Ingroup Coordination Interaction Phase: Ingroup agents are paired up. Each shows a signal. The top pair has a successful interaction; the bottom does not.

1.2 Ingroup Coordination Update Phase: Both agents from the top pair have their repertoires updated. A reward, representing positive conditioning, ρ, is applied to each signal exchanged in the interaction. Note the color change to indicate the updated signal strength.

2.1 Outgroup Detection Interaction Phase: A single ingroup agent is paired with θ outgroup agents; here one agent twice and the other once. This process is repeated for each ingroup agent.

2.2 Outgroup Detection Update Phase: A penalty, π, representing negative conditioning, signals the ingroup agent was recognized. Similarly, a reward, ρ, is applied to the signals each member of the outgroup learned to associate with a member of the ingroup. Note the color change to indicate the updated signal strength.

FIGURE 2. A schematic of the agent-based model representing a single time step. Signaling repertoires are denoted by the array of colored shapes. Each agent has its own signaling repertoire, distinguished by the different colors for each shape, indicating the informational value of each signal for that particular agent at that time. Stronger signals are lighter, weaker are darker. Aggregate signaling repertoires, the primary outcome of the model, are also pictured, representing informational values averaged across all agents. Not pictured is the extinction/emergence phase where signals that have risen and then fallen to particular thresholds in the ingroup population are reset to their initial values in the ingroup and outgroup. This represents the emergence of a novel signal at the location in the signal repertoires that was occupied by a newly extinct signal.

At each time step, each ingroup agent is paired with a partner, randomly chosen from among the remaining ingroup agents such that each agent has exactly one ingroup partner. Each agent chooses a signal to send to its partner from a normalized vector, $\mathbf{S'}_i^{\text{IN}} = \frac{s_{ik}^{\text{IN}}}{\sum_k^K s_{ik}^{\text{IN}}}$, such that a signal k is chosen by agent i with a probability equal to its relative informational value, $s_{ik}^{\prime\text{IN}}$. Based on their respective partners' signals, both agents independently decide whether to overtly declare their identity. This decision is made with a probability equal to the absolute informational value, s_{ik}^{IN} that a signal's receiver had previously assigned to their partner's signal. If neither agent declares their identity, they are presumed not to recognize one another as ingroup members and so nothing

happens. If either partner does declare their identity, then each agent learns to associate both the sent and received signals with the ingroup identity, and therefore increases the informational value of those signals by ρ_{IN}, such that $s_k^{\text{IN}} \leftarrow \rho_{\text{IN}} + s_k^{\text{IN}}$. Although this phase consists entirely of interactions between ingroup members, the entire set of phases are intended to represent a mixture of relatively concurrent interactions, and so agents have no *a priori* reason to assume their interaction partners are fellow ingroup members. We make the minimal assumption that signal weight determines both the probability of the agent using the signal and the informativeness of the signal to an in-group observer.

2.2. **Ingroup coordination repertoire update phase.** Following the coordination phase, the signaling repertoire of each ingroup agent is updated to account for new information acquired. First, each agent's signal repertoire is updated to account for any rewards accumulated during the coordination phase, as described above. Then, each signal is truncated such that any value great than 1 is set to 1 and any value less than 0 is set to 0.

Without truncation, values greater than one would tend towards positive infinity, and values less than 0 would tend towards negative infinity. Indefinitely increasing or decreasing signal values could result in runaway feedback loops, which would make model evaluation more challenging. However, there are theoretical reasons for constraining the informational value of the signals. From the perspective of signal interpretation, identity signals represent the probability of identity, and are thus bounded in $[0, 1]$, so that signals that are never used have a value of 0, and signals that reliably indicate ingroup membership have a value of 1. From the perspective of signal generation, relative signal salience should match interpretive probabilities. If two signals are equivalent predictors of identity, and the agent may select only one, there is no reason to prefer one to the other.

2.3. **Outgroup detection interaction phase.** After the ingroup agents interact amongst themselves and update their signal repertoires, they enter another round of signaling. Here the ingroup agents are partnered with members of an outgroup. The outgroup consists of N_{OUT} agents, each of which also tracks the information value of each signal that may be used by the ingroup. Each outgroup agent j is therefore characterized by a signal repertoire of length K, $\mathbf{S}_j^{\text{OUT}}$. Instead of using these signals amongst themselves, outgroup agents use their information to identify members of the ingroup. The strength of a particular signal for a given outgroup agent is the probability that they will correctly identify another agent using that as a member of the ingroup. We assume the outgroup is initially naïve about all signals, and so we initialize each signal in an outgroup agent's repertoire to a strength of 0.05.

At each time step each ingroup agent interacts with up to θ random outgroup agents, where $\theta \geq 0$. If θ is not an integer, the ingroup agent calculates a probability $p = \theta - \text{floor}(\theta)$, and interacts with $\text{ceil}(\theta)$ with probability p and $\text{floor}(\theta)$ with probability $1 - p$. For example, if $\theta = 3.2$, the agent will interact with four outgroup agents with probability 0.2, and three outgroup agents otherwise. We can therefore explore cases where the ingroup interacts mostly among themselves ($\theta < 1$), and mostly with the outgroup ($\theta > 1$), as might be the case for a minority group.

During each interaction, ingroup agents select their signal as before. Each ingroup agent i chooses a signal k from its repertoire with a probability equal to the signal's normalized informational value, s'^{IN}_{ik}. The outgroup agent j then recognizes the selected signal with a probability equal to the value of the same signal in its own repertoire, s^{OUT}_{jk}. If it identifies the agent as a member of the ingroup, the outgroup agent strengthens the signal's association with the ingroup, and adds ρ_{OUT} to its representation of the signal's information value. Additionally, the outgroup agent may behave in a manner to sanction the ingroup member. The member of the ingroup receives a penalty π that is subtracted from the signal's informational value, which may disincentivize the ingroup

agent from using the same signal again in future interactions. This is analogous to real-world sanctions for revealing a marginalized identity, ranging from a cold shoulder, verbal abuse, or an employer reprimand to physical violence, political imprisonment, etc.

2.4. **Outgroup detection repertoire update phase.** During this stage, signal weights are updated as a result of interactions between the ingroup and the outgroup. Both the ingroup and outgroup update their signal repertoires according to the same process, which is identical to the process for the ingroup coordination update phase. First, each agent's signal repertoire is updated to account for any rewards accumulated during the coordination phase, as described above. Then, for each agent, each signal is truncated such that any value great than 1 is set to 1 and any value less than 0 is set to 0.

2.5. **Signal extinction/emergence phase.** In real cases of identity signaling, the number of novel signals that a population may develop is effectively infinite. However, the constraints of our model demand we simplify to a small number of arbitrary signals, K. We simulate the emergence of genuinely novel signals by imposing extinction on a signal when its average informational value among ingroup agents decreases below a threshold, which we set to 0.05. When the extinction threshold is reached, the signal is reset to its initial value of 0.1 among the ingroup. Importantly, the outgroup signal knowledge is also reset to its initial value, 0.05, as the arbitrary ordinal position in the signal array now represents a new signal, of which the outgroup has minimal knowledge. This results in differing cycles of signal renewal in the ingroup and outgroup. The ingroup signals decline gradually as a result of punishment by the outgroup, whereas outgroup signals have sharp vertical drops, as seen in Figure 3.

Importantly, prior to extinction, a signal's average informational value among the ingroup must first exceed a "prevalence" threshold, which we set at 0.15, indicating it is widely acknowledged as an identity marker among the ingroup. Only after surpassing this level can a subsequent drop below the extinction threshold initiate a reset. Setting

the prevalence threshold too low results in signal extinction before becoming recognized as an identity marker, usually as a result of random noise (drift).

This mechanism of signal renewal approximates recurrent novelty in real identity signaling. Because outgroup knowledge is reset when an ingroup signal goes extinct, that signal should subsequently be interpreted as a completely new signal despite being indexed by the same number. So, even though K is small and finite, the model effectively generates new signals continually. In fact, even $K = 1$ is capable of cycling in our model. However we set $K = 3$ in our analysis in order to make cycling easier to interpret for the reader. Figure B1 in Appendix B demonstrates cycling signaling dynamics with between one and 16 signals.

While signal renewal is best understood as novel signal emergence, one could also interpret a new signal as the return of a signal which has been out of use long enough to lose salience with the outgroup to return as an effective identity signal among the ingroup. It may be that this signal never truly left the vocabulary of the ingroup, and was only removed from contexts where the audience identity is uncertain; thus, formerly covert signals may be relegated to "safe spaces", and reappear after a long period of dormancy, especially after generational turnover. While our design doesn't model such mechanisms directly, the process is effectively approximated by the extinction mechanism.

2.6. **Outcome measures.** We ran each simulation for $T = 1000$ time steps, performing 50 runs for each combination of parameters. This number of runs is justified both because there was very little variation in classification between runs using any particular combination of parameters, and because 1000 time steps was more than enough time to observe the model dynamics settle into stable, long-term behavior patterns (see Figure 3). In some cases (very infrequently) observed but not explored, cycles may be too wide to appear in 1000 time steps. This occurs when agent learning is low but balanced with ingroup punishment. We are content knowing that such cases exist and how one might

TABLE 1. Parameters for the signaling model.

Parameter	Symbol	Default Value	Range[a]
Ingroup size	N_{IN}	50	—
Outgroup size	N_{OUT}	500	—
Number of signals	K	3	—
Length of simulation runs	T	1000	—
Ingroup reward	ρ_{IN}	0.5	[0, 1]
Outgroup reward	ρ_{OUT}	0.2	[0, 1]
Ingroup penalty	π	0.3	[0, 1]
Out/Ingroup interaction ratio	θ	1	[0, 3]
Ingroup initial signal value		0.1	—
Outgroup initial signal value		0.05	—
Ingroup signal prevalence threshold		0.15	—
Ingroup signal extinction threshold		0.05	—

[a] Parameter value changes between model runs in this study.

locate them model in parameter space. These cases are not only marginal, but do not exhibit fundamentally different dynamics than those explored here; only the rate of the dynamics differs.

Our primary outcome measure of interest is the informational value of each signal (i.e., the signaling repertoires) for each agent over time. Specifically, we are interested in the conditions under which the all signals become effective and sustained identity markers, those under which cyclical dynamics would emerge, and those in which no signal ever attains (or retains) widespread use. Ultimately, it is which of these three categorical outcomes (saturation, cycling, or suppression) that we are really interested in. To categorize our model outcomes into one of these three classes, we relied on visual inspection of the model outcomes, i.e., time series of aggregate ingroup and outgroup signal informational values.

We also trained a random forest model on 11 time series features to automatically identify the three classes of dynamics. We manually classified 100 randomly selected parameter combinations (10 runs per combination) from across the ranges defined in

Table 1. Holding out 20% of parameter settings, cross-validation of random forest models achieved average accuracy of 0.96. This indicates that the criteria used for model outcome classification can be reliably automated. Appendix C describes the time series features used for training and reports Shapley values, a tool for interpretable machine learning (Lipovetsky and Conklin, 2001; Wallard, 2015), for the random forest classifier.

3. Results

3.1. **Characterizing model outcomes.** Visual inspection of the model output led us to consistently identify each model run in terms of one of three classes of behavior: *saturation, cycling*, or *suppression*. Figure 3 shows example runs in which each of these three classes of dynamics arose. Each column shows three example runs, with each example involving two graphs: the signal strengths for the ingroup (top) and outgroup (bottom). Each colored line represents the average weight a particular signal among either the ingroup or outgroup agents.

In instances of saturation, all signals become effective identity markers and remain so, despite occasional penalties from being identified by outgroup agents. This differs from the analytical model in Appendix A, which prevents multiple signals from becomingly strongly associated with the identity concurrently. In this regard the agent-based model is considerably more realistic, as groups rarely rely on a signal signal to communicate group membership.

When identification by the outgroup becomes costlier, the dominant signal may be sufficiently disincentivized once the outgroup learns to associate it with the ingroup, and so ingroup agents begin to use alternative signals. Through reinforcement learning, they converge on a new conventional signal with which to identify other ingroup agents. Once conventionalized, however, the new dominant signal becomes a new target for outgroup learning. When the outgroup learns the new ingroup signal, ingroup agents are once again forced to abandon it. A new signal emerges as dominant, the cycle begins anew.

Note that in our model, when the ingroup abandons a signal, the outgroup knowledge of that signal is also reset, effectively generating a novel signal about which the ingroup and outgroup have minimal knowledge. Thus, despite the finite and small number of signals in the model, the model should be interpreted as capable of generating infinitely novel signals across a run, and also of recycling an old signal which has lost its salience in both populations, as often happens in fashion cycles. In other words, the parameter K represents not the total the number of possible signals, but rather the maximum number of signals used *simultaneously* at any given time.

The final class occurs in cases where the outgroup learns quickly and when being identified as a member of the ingroup is severely punished. Here, no signal ever dominates or becomes informative as a group marker. Rather we see a pattern of suppression. In this circumstance all signals are equally likely and none carries identity information.

We also defined three features of the model runs that correspond with theoretically relevant features of covert signaling dynamics. The first, P(Extinction), is the probability that a signal that has become prevalent in the ingroup population goes extinct and is replaced by the emergence of new signal. This corresponds to the complete "life cycle" of a covert signal. We also measure μ(Signal Peak), the mean height of the highest value a dominant signal reaches before it is replaced by a new dominant signal. The final metric is P(Ingroup Success), the probability that an ingroup signaling interaction results in successful coordination. Each of these features is included in training the random forest classifier employed in some of our analyses.

Readers of the analytical model in Appendix A may notice parallels between these measures and the results given in Figure A2. μ(Signal Peak) is calculated and reported for both the agential and analytical models. P(Extinction) is analogous to P(Transition) in the analytical model; P(Transition) is also measurable here, but P(Extinction) is a more theoretical meaningful metric, as it corresponds to signal abandonment and novel signal emergence, rather than resurgence of an existing signal. P(Ingroup Success) has

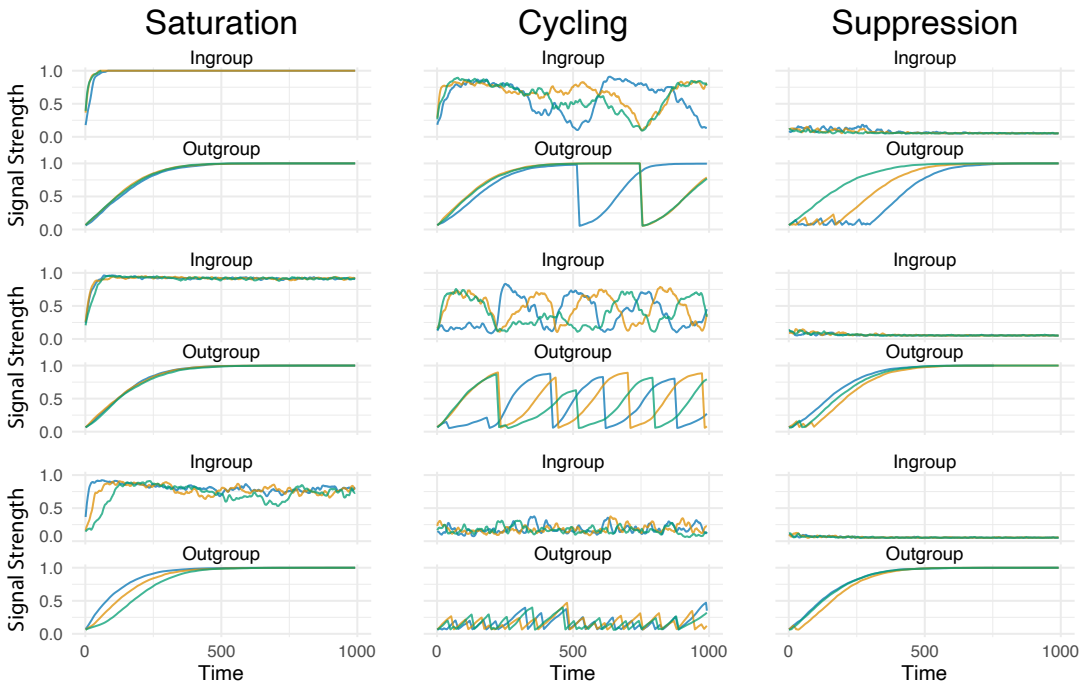


FIGURE 3. The three classes of model dynamics. Examples were selected by nonrandomly sampling runs of a parameter sweep of ingroup punishment (π). Time series occur over 1000 time steps, and data are smoothed using a moving average with a 10-time step window. The remaining parameters used in these runs are given by the Default Parameters column of Table 1.

no analogue in the analytical model, as signaling behavior is effectively abstracted away and changes in frequency are based entirely on "signal population" sizes.

3.2. **Characterizing the parameter space.** With our analytical methods established, we can say more about the relationships among the model parameters in producing each the three categorical model outcomes. The model produces three distinct classes of behavior: saturation, where all identity signals become stable conventions and never go extinct; cycling, in which new signals emerge to replace old ones as they are learned by the outgroup; and suppression, in which informative identity signals never emerge. Cycling exists in the regions of parameter space between saturation and suppression. We can draw an analogy to phases of matter. Saturation is gas. Then, as the (social)

pressure increases, we observe transitions to cycling (liquid) followed by suppression (solid), so that the state of matter changes in response to the expected costs of outgroup interaction. Holding all other parameters constant, as we sweep a particular parameter, the dynamics are driven towards or away from the neighboring class of dynamics. For example, parameter settings that produce saturation will be driven towards cycling as the expected cost of interacting with the outgroup is increased, and past cycling into suppression if it is increased further. Similarly, parameter settings that produce cycling will be driven into saturation as the ingroup reward is increased, and into suppression as the reward is decreased. The tendency for each parameter to drive the system toward saturation or suppression as it is increased is given by the "Dynamic Tendency" column of Table 2. Generally, saturation occurs under conditions that favor ingroup coordination or disfavor outgroup detection, and suppression occurs under conditions that favor outgroup detection or disfavor ingroup coordination. Put another way, suppression occurs under conditions where we expect a high cost of outgroup interaction relative to the benefit of ingroup interaction. Cycling occurs in the ranges where the expected payoffs from outgroup and ingroup interactions are more balanced. Figure 4 shows that for the default parameter values (given by the rightmost column of Table 1), most parameter settings in the sweeps result in either suppression or cycling.

TABLE 2. Four key model parameters and the effect of increasing them. Increasing each parameter drives the model toward either saturation or suppression (given by the "Dynamic Tendency" column) when all other parameters held constant. Between these two equilibria lies cycling dynamics.

Parameter		Dynamic Tendency
ρ_{IN}	Ingroup reward	Saturation
ρ_{OUT}	Outgroup reward	Suppression
π	Ingroup punishment	Suppression
θ	Out/Ingroup interaction ratio	Suppression

While the model has many parameters, we focus initially on four, given by Table 2. For two of these, predictions exist based on prior models of covert signaling (Smaldino et al., 2018; Smaldino and Turner, 2022). These models predict that, compared with a strategy of overt signaling (in which a stable dominant signal is plausible), covert signals (which are costly when detected by outgroup individuals) will be favored when the risk of detection by the outgroup is large and the cost of such detection is high. In the current model, the ingroup penalty, π, represents the cost of detection, while the ratio of outgroup-to-ingroup interactions, θ, represents the overall risk of detection by the outgroup.

Figure 4 shows the outcome variables P(Transition), μ(Signal Peak), P(Ingroup Success), and P(Outgroup Success) as we sweep across our parameters of interest (Table 2). Colored bands indicate which class of dynamics the model is exhibiting at each parameter setting. Figure 4 illustrates the tendency for each parameter to lead to one or the other pole of dynamics (saturation or suppression). Increasing ingroup reward decreases the frequency of cycling, though under these parameter settings it does not reach full saturation. Increasing the other parameters pushes the dynamics toward suppression. More specifically, ingroup punishment, which represents negative reinforcement, formalizes the outgroup's ability to sanction the ingroup when recognized. Without any such sanctioning, the initially-dominant signal can persist; at least some ingroup punishment is required to produce cycling. However, if the ingroup punishment becomes too great, cycling dynamics will give way to suppression. This is also the case for the ratio of outgroup to ingroup interactions, which represents the relative ability of the ingroup to assort preferentially among themselves (even if they don't always know it), and avoid too many potential encounters with the outgroup. Decreasing the relative number of outgroup interactions produces saturation: the ingroup is able to coordinate and is rarely sanctioned. As the rate of interaction with the outgroup increases, we see cycling dynamics: the outgroup punishment forces the ingroup to periodically abandon

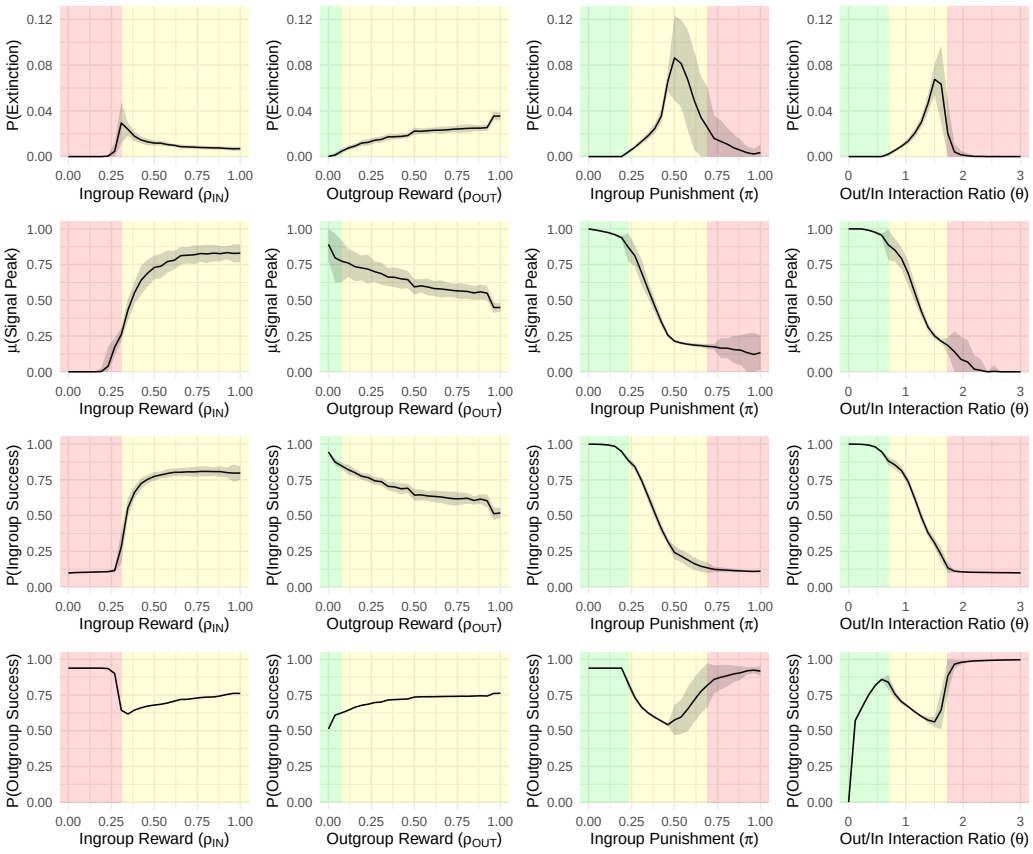


FIGURE 4. Sweeps of theoretically significant parameters holding all others constant. Each plot shows the relationship between a parameter and a feature of model dynamics. 95% confidence intervals are demarcated by the shaded gray band on either side of the line. Dynamic classes are shown by the shaded rectangles: green, saturation; yellow, cycling; red, suppression. Each column corresponds to 1 of the 4 parameters of interest. The x-axes show the parameters by which they differ: ingroup reward (ρ_{IN}), outgroup reward (ρ_{OUT}), ingroup punishment (π), and outgroup/ingroup interaction ratio (θ). Rows correspond to four outcome features. Top: P(Extinction), the probability that a signal goes extinct and restarts, corresponding to the period of the cycles. Upper Middle: μ(Signal Peak), the average knowledge of a signal among the ingroup, corresponding to the amplitude of the cycles. Lower Middle: P(Ingroup Success), the proportion of ingroup interactions which resulted in one or both of the ingroup agents recognizing their partner's signal and revealing their own identity. Bottom: P(Outgroup Success), the proportion of ingroup-outgroup interactions which resulted in the outgroup member identifying the ingroup member.

their dominant signal and learn to coordinate around a new signal. Note P(Outgroup Success) mirrors P(Extinction), illustrating how cycling is an adaptive response to moderately costly detection by the outgroup. When outgroup interaction grows too frequent, however, the ingroup cannot effectively coordinate at all as the sanctions wipe out any progress toward converging on a signal.

The simulations represented in Figure 4 were initialized with all ingroup and outgroup signals set to a low value, 0.1 and 0.05 respectively. Given that cycling emerges under conditions intermediate between saturation and suppression, one might wonder whether it is a merely transitory state and not actually stable, tending eventually toward saturation or suppression. To show that cycling is *indeed* a stable state, we ran two additional sets of parameter sweeps from initial conditions corresponding to saturation- and suppression-like states. Figure D1 demonstrates the dynamics from initial saturation-like conditions, in which all ingroup signals were initialized to 1 instead of 0.1. This figure shows nearly identical patterns to Figure 4, indicating that within parameter settings that lead to cycling, cycling is a stable attractor from both neutral and saturation-like regimes. Similarly, to show that cycling is a stable attractor from a suppression-like state we executed a set of model runs that initialized the outgroup signals to 1 and the ingroup signals to 0.1. These results are depicted in Figure D2 and are likewise similar to the results in Figure 4, though they differ slightly as some parameter combinations that typically result in cycling dynamics remained suppressive. The discrepancy is due to an artifact of our representing only a small finite number of signals for computational tractability. Figure D3 shows a sample run in which suppression-like initial conditions do resolve to cycling dynamics. These results indicate cycling appears to be a stable attractor from a suppression-like state, and are discussed in greater detail in Appendix D. Notably, the suppression-like state used for the initial conditions in these model runs is highly artificial. When the model is run with parameter settings that produce cycling from conditions that are not close to suppression (i.e., at least one signal is not

well-known to the outgroup), the model is extremely unlikely to reach a suppression-like state by random drift. Moreover, this scenario—in which no initial signals are unknown to the outgroup, as opposed to arriving at suppression from some other state—is outside of the range of conditions intended to be captured by the model. In reality, we imagine that members of a group could always put forth a novel signal for consideration as a group marker.

An alternate perspective is given by Figure 5, which examines the effects of interacting parameters. Again we sweep the four parameters of greatest interest, this time at four regular intervals in $[0.25, 1]$, giving 256 (4^4) parameter combinations (10 runs each). Each cell of figure 5 indicates the random forest classifier's average predicted probability that it is observing each class. The predicted probability should be thought of not as the probability of observing each class, as the dynamics are consistent across runs at a particular parameter setting, but rather the classifier's "confidence" that it is observing a particular class of dynamics. Figure 5 demonstrates that parameters interact additively. There is a generally continuous gradient from saturation (green), through cycling (yellow), to suppression (red). As three of the parameters—outgroup reward, ingroup punishment, and interaction rate—tend toward suppression as they increase, the gradient appears to flow toward suppression in an upward and rightward linear vector. Within each 4×4 subgrid, truly continuous gradients flow upward and *leftward* as ingroup reward tends toward saturation as it increases.

4. Discussion

Humans living in diverse societies require conventionalized signals to help them recognize and assort with similar partners, as well as to avoid or even (in some cases) aggress against dissimilar others. When group boundaries are clearly defined and delineate the structure of social interactions, the identity signals that emerge can persist for as long as

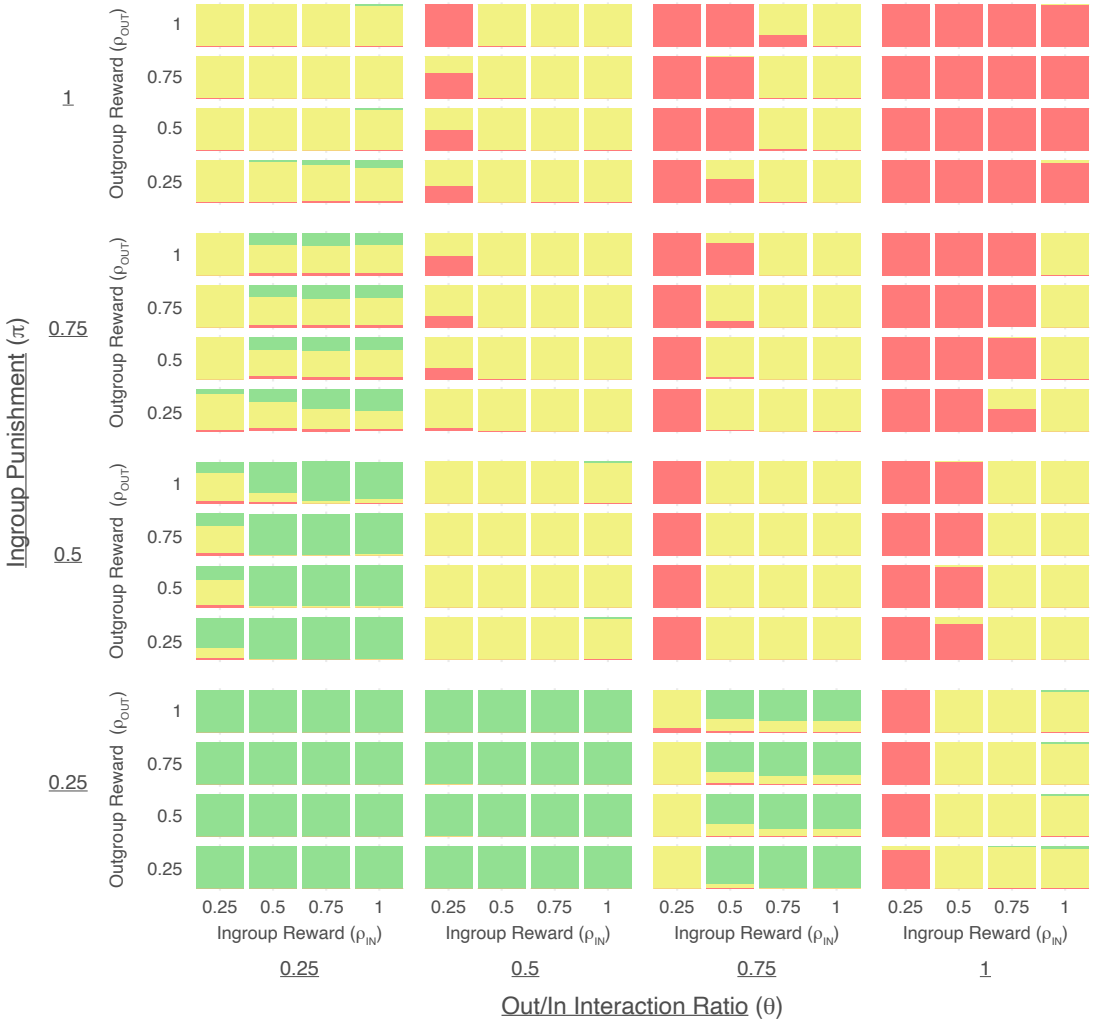


FIGURE 5. Sampling parameter space across theoretically significant parameters at regular intervals, holding all other parameters at the values given in Table 1. In total 256 (4^4) parameter combinations are visualized. Each colored bar shows random forest model confidence at a given parameter setting in observing each of the three dynamic classes: green, saturation; yellow, cycling; red, suppression. Inner x- and y-axes indicate the value of ingroup reward and outgroup reward respectively. Outer x- and y-axes indicate the value of out/in interaction ratio and ingroup punishment respectively.

the groups do. When groups mix, however—as they typically do in most places today—dividing lines between groups can cause problems. Such a scenario can select for covert signaling strategies, in which individuals are more likely to use encrypted or obfuscated

identity signals that are less likely to correctly interpreted by outsiders (Smaldino et al., 2018; Smaldino and Turner, 2022). Previous work on covert identity signaling has been largely agnostic about the specific signals individuals might be used, or for how long a particular signal might effectively remain covert if outgroup individuals are incentivized to recognize members of other groups. Our model demonstrates how identification by an outgroup, if sufficiently likely and sufficiently costly, can cause covert identity signals to have natural life cycles in which they first emerge as ingroup signaling conventions but are then discarded as they become known to outsiders, replaced by new signals.

The dynamic we describe is reminiscent of the Red Queen hypothesis from evolutionary biology (Van Valen, 1973). There, a species at risk of predation, parasitism, or competition from other species must constantly evolve new adaptations just to maintain constant fitness levels, in response to the continuous evolution of adaptations to compete or predate on the part of the opposing species. Our proposal is that, similarly, groups with incentives to remain even partially obscured to outsiders will need to constantly create new identity signals in response to the incentives faced by outsiders to recognize signals identifying other groups. In our model, particular signals appear to cycle in and out of fashion, but this need not be the case—indeed, our model does not require this interpretation. Instead, we should imagine each signal to merely be an index or pointer, for which the signal being represented is replaced by a novel one whenever the informational value of a particular index becomes sufficiently small.

We further show that if identification by the outgroup is *too* likely and *too* costly, conventionalized identity signals may *never* emerge for public display. In this scenario, members of the ingroup are quite bad at identifying one another in public, because the risk of being identified by unfriendly outsiders is too great. This is perhaps akin to the situation faced by political dissidents under a brutal totalitarian regime, or even by xenophobic extremists in a well-functioning cosmopolitan society. More mundanely, it may also represent scenarios in which the benefit of coordinating on a particular

idiosyncratic identity may be outweighed by the social costs of outing oneself as different. Our model assumes that agents must signal, and we find that under substantial threat they simply use signals that are minimally informative. In the real world, one could propose that people simply avoid signaling their identity, but in practice this is hard to do. It is nearly impossible to avoid transmitting any identity information during a social interaction (Moffett, 2019). The best one can do in these cases may be to avoid transmitting information that can *reliably* be linked to a persecuted group identity, such as by using ambiguous or generic speech, or even outright deception. This is in line with our model results.

Our analysis identifies the importance of learning rates to the dynamics of identity signal use, and shows that these rates contribute additively, along with expected costs of interactions with outgroup, to whether signal dynamics exhibit stable saturation, cycling, or suppression (see Figures 4 and 5). Sufficient reward from successfully using signals to identify ingroup is required to conventionalize particular signals as group markers. However, the rate at which the outgroup acquires knowledge of the ingroup's identity signals must be sufficiently slow as to allow time for that conventionalization to usefully take hold. Empirical determination of these learning rates, which must be determined relative to rates of outgroup interaction and punishment, is likely to be difficult. Complicating matters is the fact that in our model, learning rates could be set independently from the incentives of either successful ingroup coordination or harmful outgroup punishment, whereas in reality, these factors are likely to be entangled. We view it likely that learning rates will often be optimized to meet the strategic needs of group members, conditional upon other constraints of behavior, cognition, and physiology.

The theory of covert signaling (Smaldino et al., 2018; Smaldino and Turner, 2022; van der Does et al., 2022) indicates that speech or other communicative acts that deliberately or at least instrumentally (and so potentially unconsciously) obscure identity

information from non-insiders should be more common when ingroup members are relatively uncommon and the cost of being identified as such by outgroup individuals is nontrivial. Our results here indicate that individuals must do more than simply attune themselves to signals that are intrinsically overt or covert. Rather, the results imply several consequences for individuals living in diverse populations in which covert signals are the best strategic choice for identity signaling. Individuals must of course possess, consciously or not, some understanding of the informational value of both sent and received identity signals (Skyrms, 2010; Bergstrom and Rosvall, 2011), so that they correctly present themselves to others and correctly identify coalitional commitments and behavioral tendencies in others. Individuals must of course also possess appropriate strategic caution to favor covert or encrypted signals in environments in which both friendly and unfriendly audience members are present (Loury, 1994; Smaldino et al., 2018; Smaldino and Turner, 2022). But individuals for whom covert signaling is strategic must also learn and continuously re-learn (1) the likely informational content of potential signals so as to accurately present themselves, (2) the information content of particular signals used by others so as to accurately identify them, and (3) the likely risks of being discovered by outgroup listeners condition on using particular signals. The key point is that each of these estimates (1–3) may be continuously changing. Understanding the cognitive mechanisms, and their associated accuracy, behind these estimates is an important task for research on the psychology of social cognition and intergroup interactions.

Previous work described overt and covert signals as if they were distinct classes of symbolic communication, with distinct lexicons (Smaldino et al., 2018; Smaldino and Turner, 2022). The present study illustrates how this need not be the case. Rather, the property of a signal as covert or overt results from a dynamic process that emerges from the changing state of collective knowledge. An initially arbitrary signal may become a covert signal of identity within an ingroup through repeated association while remaining obscure to outsiders, only to later become an overt signal once sufficiently common use

allows those outsiders to determine its meaning as an identity signal. At this point, the signal may either remain an oft-used identity signal if the costs of identification by outsiders is negligible, or be abandoned if covertness is required.

This work presented here is agnostic about some of the psychological process that determine or moderate the model parameters. For example, the reward for successfully using identity signals for assortment may stem from emotional responses, monetary gains, or the maintenance of partnerships or coalitions. Any of these could be used by learning mechanisms to reinforce some signaling behaviors and to suppress others. We view this agnosticism as a strength of the model, as it is similar to the phenotypic gambit common in behavior ecology (Grafen, 1991), which focuses on how behavioral strategies create adaptive value without worrying overly about the mechanisms that generate those behaviors (temporarily at least). This approach is no less fruitful when applied to human behaviors that may evolve culturally (Smith and Winterhalder, 1992), even if behaviors must eventually be understood in terms of their generating mechanisms (Heyes, 2016). Indeed, the present work can be viewed in the context of other attempts to understand dynamic patterns of social and cultural change using formal modeling approaches (e.g., Turchin, 2003, 2011), along with more recent efforts to better integrate cultural evolution with cognitive science (e.g., Heyes, 2018).

In general, although our model does embrace some of the important complexity present in the real world, it is still a drastically simplified representation of human behavior. This kind of simplification is necessary for the development of robust formal theories of social behavior, particularly in domains where the number of formal models is still relatively small (Smaldino, 2017, 2023). Nevertheless, it is important to acknowledge how some of our model assumptions constrain our ability to generalize. We assume that the benefits to successful assortment manifest simply as positive reinforcement for the use of a particular signal, while the costs of being identified by the outgroup manifest simply as negative reinforcement for the use of that signal, with both of these identically

implemented across all agents. We view this is as a plausible dynamic that is consistent with well-mixed models of weak selection commonly used to model both genetic and cultural evolution (Mullon and Lehmann, 2014; Rodrigues and Kokko, 2016). In doing so, we ignore issues like individual differences in signaling strategies and power dynamics, as well as the potential for the population structure to evolve if punished individuals are removed from the population. Exploring these and other limitations are important avenues for future modeling work.

Our work highlights the need for more empirical work to investigate the dynamics of identity signals, particularly those that may be covert and used only within particular groups. Studying such signals is difficult, because covert signals are not easily identified by non-group members (van der Does et al., 2022). This difficulty is further complicated if the use of a particular signal as an identity marker and the level of covertness attributable to that signal both change over time. Our model yields qualitative but testable hypotheses about the nature of these signaling dynamics. We expect to see clear relationships between the expected costs of detection by outgroup individuals and the character and lifespan of the identity signals used by ingroup individuals to coordinate. These predictions should aid future empirical investigation. Given the importance of considering identity for understanding social processes, such investigations are warranted.

References

Acerbi, A., Ghirlanda, S., and Enquist, M. (2012). The logic of fashion cycles. *PLOS ONE*, 7(3):e32541.

Barth, F. (1969). Introduction. In Barth, F., editor, *Ethnic Groups and Boundaries*, pages 9–38. Little, Brown, New York.

Bell, A. V. and Paegle, A. (2021). Ethnic markers and how to find them. *Human Nature*, 32(2):470–481.

Berger, J. and Heath, C. (2008). Who drives divergence? identity signaling, outgroup dissimilarity, and the abandonment of cultural tastes. *Journal of Personality and Social Psychology*, 95(3):593.

Berger, J. and Ward, M. (2010). Subtle signals of inconspicuous consumption. *Journal of Consumer Research*, 37(4):555–569.

Bergstrom, C. T. and Rosvall, M. (2011). The transmission sense of information. *Biology & Philosophy*, 26:159–176.

Boyd, R. and Richerson, P. J. (1987). The evolution of ethnic markers. *Cultural Anthropology*, 2(1):65–79.

Burke, P. J. and Stets, J. E. (2009). *Identity theory*. Oxford University Press.

Castro, L. and Toro, M. A. (2007). Mutual benefit cooperation and ethnic cultural diversity. *Theoretical Population Biology*, 71(3):392–399.

Centola, D. and Baronchelli, A. (2015). The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences*, 112(7):1989–1994.

Chen, Y. and Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, 99(1):431–457.

Cohen, E. and Haun, D. (2013). The development of tag-based cooperation via a socially acquired trait. *Evolution and Human Behavior*, 34(3):230–235.

Crawford, V. P. and Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50(6):1431–1451.

DellaPosta, D., Shi, Y., and Macy, M. (2015). Why do liberals drink lattes? *American Journal of Sociology*, 120(5):1473–1511.

Di Giovinazzo, V. and Naimzada, A. (2015). A model of fashion: Endogenous preferences in social interaction. *Economic Modelling*, 47:12–17.

Donath, J. S. (1999). Identity and deception in the virtual community. In Kollock, P. and Smith, M., editors, *Communities in cyberspace*, pages 29–59. Routledge.

Dunivin, Z. O. and Smaldino, P. E. (2023). Dynamics of covert signaling: Modeling the emergence and extinction of identity signals. *PsyArXiv. https://doi.org/10.31234/osf.io/3tz2r*.

Efferson, C., Lalive, R., and Fehr, E. (2008). The coevolution of cultural groups and ingroup favoritism. *Science*, 321(5897):1844–1849.

Fischer, H. (2015). *Gay Semiotics: A Photographic Study of Visual Coding Among Homosexual Men*. Cherry and Martin.

Flamson, T. J. and Bryant, G. A. (2013). Signals of humor: Encryption and laughter in social interaction. In Dynel, M., editor, *Developments in Linguistic Humour Theory*, volume 1, pages 49–73. John Benjamins Publishing, Amsterdam.

Goffman, E. (1978). *The presentation of self in everyday life*. Harmondsworth London.

Grafen, A. (1991). Modelling in behavioural ecology. In Krebs, J. and Davies, N., editors, *Behavioural ecology, 3rd edition*, pages 5–31. Blackwell Scientific Publications.

Guilbeault, D., Baronchelli, A., and Centola, D. (2021). Experimental evidence for scale-induced category convergence across populations. *Nature Communications*, 12(1):1–7.

Henderson, R. and McCready, E. (2017). How dogwhistles work. In *JSAI International Symposium on Artificial Intelligence*, pages 231–240. Springer.

Henrich, J. and Muthukrishna, M. (2021). The origins and psychology of human cooperation. *Annual Review of Psychology*, 72:207–240.

Heyes, C. (2016). Blackboxing: social learning strategies and cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1693):20150369.

Heyes, C. (2018). Enquire within: Cultural evolution and cognitive science. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1743):20170051.

Kirkpatrick, D. D. (2004). Speaking in the tongue of evangelicals. *New York Times*, 17 October 2004.

Krawczyk, M. J., Dydejczyk, A., and Kułakowski, K. (2014). The Simmel effect and babies' names. *Physica A: Statistical Mechanics and Its Applications*, 395:384–391.

Lee, J. J. and Pinker, S. (2010). Rationales for indirect speech: The theory of the strategic speaker. *Psychological Review*, 117(3):785.

Lipovetsky, S. and Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330.

Loury, G. C. (1994). Self-censorship in public discourse: A theory of "political correctness" and related phenomena. *Rationality and Society*, 6(4):428–461.

McElreath, R., Boyd, R., and Richerson, P. J. (2003). Shared norms and the evolution of ethnic markers. *Current Anthropology*, 44(1):122–130.

McPherson, M. (2004). A Blau space primer: prolegomenon to an ecology of affiliation. *Industrial and Corporate Change*, 13(1):263–280.

Miller, C. M., McIntyre, S. H., and Mantrala, M. K. (1993). Toward formalizing fashion theory. *Journal of Marketing Research*, 30(2):142–157.

Moffett, M. W. (2019). *The human swarm: How our societies arise, thrive, and fall.* Basic Books.

Mullon, C. and Lehmann, L. (2014). The robustness of the weak selection approximation for the evolution of altruism against strong selection. *Journal of Evolutionary Biology*, 27(10):2272–2282.

Nettle, D. and Dunbar, R. (1997). Social markers and the evolution of reciprocal exchange. *Current Anthropology*, 38:93–99.

Otto, S. P. and Day, T. (2011). *A biologist's guide to mathematical modeling in ecology and evolution.* Princeton University Press.

Parker, M. A. (1994). Pathogens and sex in plants. *Evolutionary Ecology*, 8:560–584.

Pedone, R. and Conte, R. (2001). Dynamics of status symbols and social complexity. *Social Science Computer Review*, 19(3):249–262.

Powell, M., Kim, A. D., and Smaldino, P. E. (2023). Hashtags as signals of political identity: #BlackLivesMatter and #AllLivesMatter. *PLOS ONE*, 18(6):e0286524.

Puglisi, A., Baronchelli, A., and Loreto, V. (2008). Cultural route to the emergence of linguistic categories. *Proceedings of the National Academy of Sciences*, 105(23):7936–7940.

Rodrigues, A. M. and Kokko, H. (2016). Models of social evolution: can we do better to predict 'who helps whom to achieve what'? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1687):20150088.

Sardanyés, J. and Solé, R. V. (2008). Matching allele dynamics and coevolution in a minimal predator–prey replicator model. *Physics Letters A*, 372(4):341–350.

Simmel, G. (1904). Fashion. *International Quarterly*, 10:130—155.

Skyrms, B. (2010). *Signals: Evolution, learning, and information*. Oxford University Press.

Sloman, S. J., Oppenheimer, D. M., and DeDeo, S. (2021). Can we detect conditioned variation in political speech? Two kinds of discussion and types of conversation. *PLOS ONE*, 16(2):e0246689.

Smaldino, P. E. (2017). Models are stupid, and we need more of them. In Vallacher, R. R., Nowak, A., and Read, S. J., editors, *Computational social psychology*, pages 311–331. Routledge.

Smaldino, P. E. (2019). Social identity and cooperation in cultural evolution. *Behavioural Processes*, 161:108–116.

Smaldino, P. E. (2022). Models of identity signaling. *Current Directions in Psychological Science*, page 09637214221075609.

Smaldino, P. E. (2023). *Modeling social behavior: Mathematical and agent-based models of social dynamics and cultural evolution*. Princeton University Press.

Smaldino, P. E., Flamson, T. J., and McElreath, R. (2018). The evolution of covert signaling. *Scientific Reports*, 8(1):1–10.

Smaldino, P. E. and Turner, M. A. (2022). Covert signaling is an adaptive communication strategy in diverse populations. *Psychological Review*, 129(4):812–829.

Smith, E. A. and Winterhalder, B. (1992). Natural selection and decision-making: Some fundamental principles. In *Evolutionary ecology and human behavior*, pages 25–60. De Gruyter.

Stiefel, K. M. and Ermentrout, G. B. (2016). Neurons as oscillators. *Journal of Neurophysiology*, 116(6):2950–2960.

Turchin, P. (2003). *Historical dynamics: Why states rise and fall*. Princeton University Press.

Turchin, P. (2011). Toward cliodynamics: An analytical, predictive science of history. *Cliodynamics*, 2:167—186.

Tylén, K., Fusaroli, R., Bundgaard, P. F., and Østergaard, S. (2013). Making sense together: A dynamical account of linguistic meaning-making. *Semiotica*, 194:39–62.

Urbatsch, R. (2014). Nominal partisanship: Names as political identity signals. *PS: Political Science & Politics*, 47(2):463–467.

van der Does, T., Galesic, M., Dunivin, Z. O., and Smaldino, P. E. (2022). Strategic identity signaling in heterogeneous networks. *Proceedings of the National Academy of Sciences*, 119(10):e2117898119.

Van Valen, L. (1973). A new evolutionary law. *Evolutionary Theory*, 1:1–30.

Wallard, H. (2015). Using explained variance allocation to analyse importance of predictors. In *16th ASMDA conference proceedings*, volume 30, pages 1043–1054.

Wimmer, A. (2008). The making and unmaking of ethnic boundaries: A multilevel process theory. *American Journal of Sociology*, 113(4):970–1022.

Appendix A. Analytical Model

Here we present an analytical model based on coupled differential equations. This model captures several core elements of the theory discussed in this paper, and also produces patterns of stability, cycling, and noise that are qualitatively similar to the agent-based model presented in the main text. Nevertheless, the complexity of the theory means that this model omits some key elements, such as explicit agent learning and the ability of multiple signals to be simultaneously expressed. In the interest of robustness and completeness, we present this model here.

We consider a signaling system in which there are a set of K possible signals, any of which can be used by an ingroup to effectively assort. The strength of a signal i among the ingroup, x_i, is its proclivity to be used by members of the ingroup, and is therefore equivalent to its frequency of use in that population. In the absence of an outgroup, popular signals will increase in strength more rapidly than unpopular ones. However, popular signals are also likely to be learned by the outgroup. We designate the strength of awareness of a signal i among the outgroup as y_i. Awareness among the outgroup decreases the utility and therefore the strength of the signal among the ingroup. However, when a signal is not in high usage, outgroup awareness will diminish to baseline levels. We can represent this dynamic as a system of coupled differential equations. The strength of ingroup signals changes as follows:

$$\dot{x}_i = rx_i \left(1 - \sum_j x_j\right) - cpy_i \frac{x_i}{1 + x_i}, \tag{1}$$

where r is the reward for successfully using the signal among the ingroup, and the parenthetical in the first term represents the fact that the strengths of all signals must sum to one; c is the contact rate between in- and outgroup members, p is the punishment of ingroup members from detection by the outgroup, and the fraction in the second term indicates diminishing marginal returns to punishment.

We similarly represent the change in outgroup signal awareness as follows:

$$\dot{y}_i = c\rho y_i(1 - y_i)\frac{x_i}{1 + x_i} - \delta y_i + \epsilon, \tag{2}$$

where ρ is the rate at which outgroup individuals learn to recognize ingroup signals, which we model as a logistic function; δ is the intrinsic decay rate of signal knowledge among the outgroup, and ϵ is the baseline awareness of a signal among the outgroup.

For simplicity, we present model explorations with the following default parameter values: $K = 3$, $r = 1$, $\delta = 0.05$, $\rho = 0.4$, $\epsilon = 0.001$. R code to fully explore the model through numerical simulation is provided at `https://osf.io/4vcug/?view_only=14c6e87c32f146a6910a24e7dd079191`. Our focus is on the two parameters most central to our verbal theory: the punishment to ingroup individuals from outgroup detection, p, and the relative rate of contact between the ingroup and the outgroup, c. These parameters are at the core of our verbal theory described above. When p is very low, detection by the outgroup should matter little, and stable signaling conventions should emerge. As p increases, dominant signals should be replaced with increasing frequency until persistent dominance is either transient or impossible. Example dynamics for both ingroup signal strength and outgroup signal awareness are shown in Figure A1. Similarly, lower contact rate (c) reduces the expected cost of punishment, so that a stable, dominant signal can persist more easily for a given level of punishment, p. The left panel of Figure A2 shows the probability of a new dominant ingroup signal (peak) arising per unit time, estimated over the course of a simulation run lasting 10,000 time steps. Contact rates of $c > 1$ imply that ingroup individuals are more likely to interact with members of the outgroup than members of their own group. The middle panel of Figure A2 extends this analysis for much larger values of p for $c = 1$, showing a critical transition in which the number of peaks stops increasing with p and instead decreases sharply. This is when the overall strength of the dominant signal starts to become indistinguishable from that of the other signals. This corresponds to a suppression of effective identity signaling among ingroup

members. This transition from cycling to suppression happens somewhat gradually as p increases, with the mean strength of the dominant signal decreasing gradually starting with much lower values of p (Figure A2, right panel).

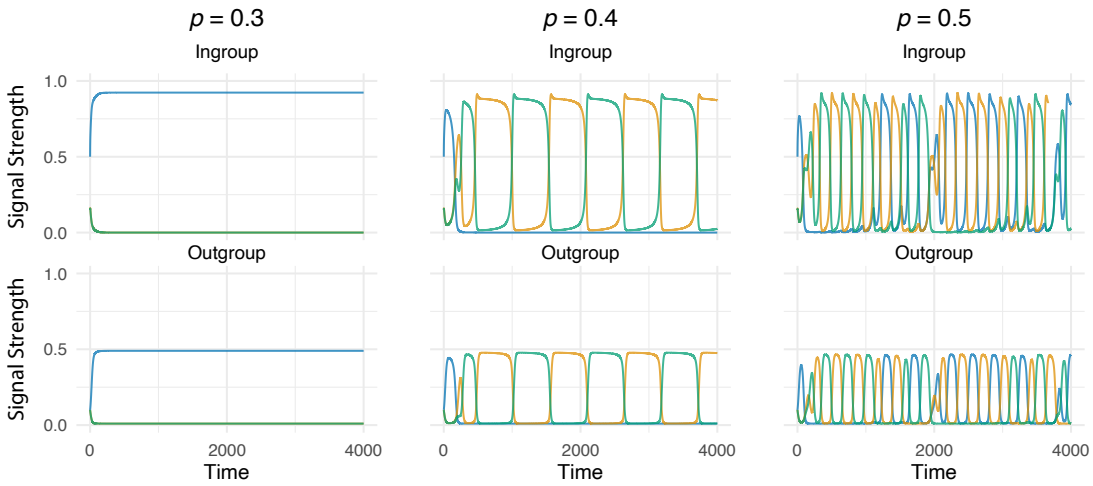


FIGURE A1. Example dynamics for ingroup signal strengths (top) and outgroup signal awareness (bottom) for different values of punishment, p. Here $c = 1$.

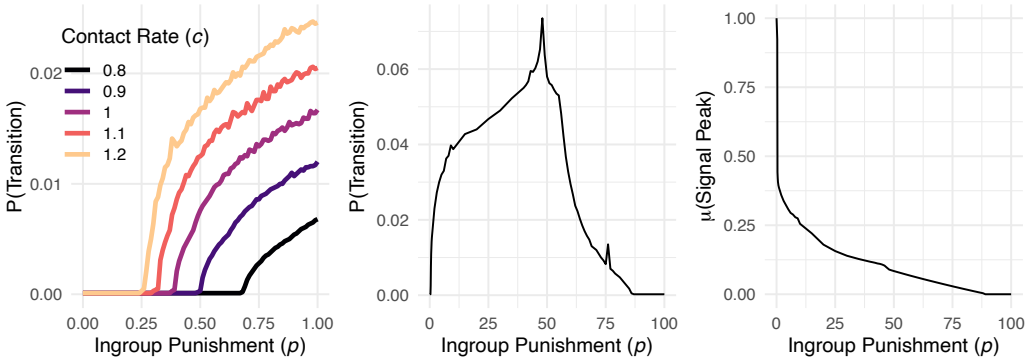


FIGURE A2. Summary results from numerical simulations of the analytical model. Left: Probability of a new dominant ingroup signal, P(Transition), as a function of p, for several values of c. For low values of p and c, there is only one stable peak. Middle: P(Transition) for much larger values of p. As p approaches a critical value just below 50, the rate of cycling stops increasing and begins to decrease. Right: Cycle amplitude diminishes as p increases. When p gets very large, all signals are suppressed. In both the middle and right panels $c = 1$.

This mathematical model captures several key features of our theory. We observe the persistence of a single dominant ingroup signal when the expected cost of outgroup interactions are low (driven by low contact rate and/or cost of punishment for detection), cycling between different signals with increasing frequency when the expected cost increases sufficiently, and the suppression of any dominant signals once the expected cost of signaling exceeds some threshold. However, this model also has several limitations. First, the model assumes that all individuals have identical knowledge of signal prevalence, and therefore identical proclivities to use those signals. This means that ingroup and outgroup knowledge *necessarily* rise and fall together. Second, the model requires that individuals' knowledge of signals must exactly track the relative frequency at which they use those signals. This means that increased knowledge of one signal automatically implies decreased knowledge of other signals. Third, the model does not allow for the emergence of novel signals. In reality, when a signal is driven out of use, human signalers should be able to replace it with a new signal that is (initially) unknown to the outgroup. Finally, the analytical model represents human learning as a process of selection at the population level. While this sort of abstraction is useful, it must still be tested against more individual-level cognitive mechanisms for learning to verify that the behavior of individual learners can in fact be accurately represented by the mean-field approach described above.

Appendix B. Cycling Occurs across Many Values of K

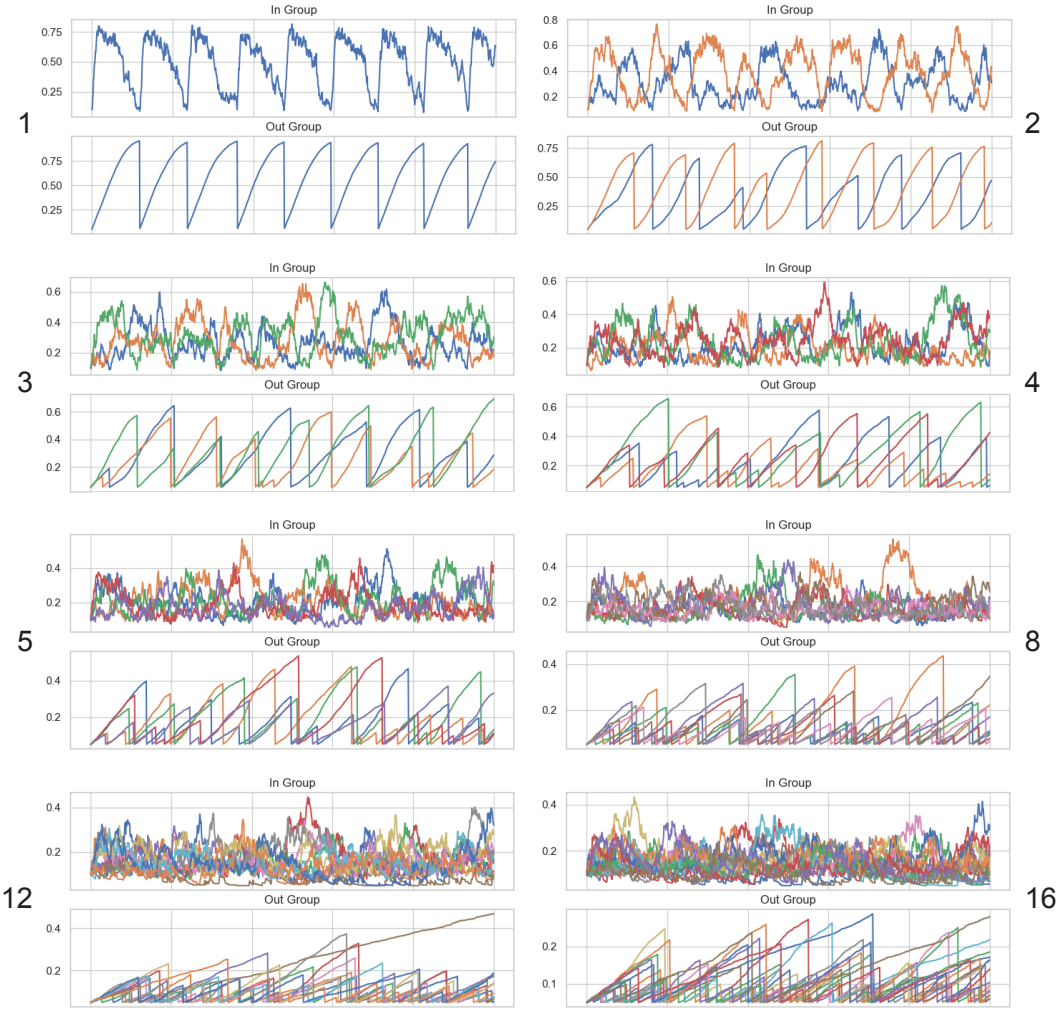


Figure B1. Signaling dynamics sweeping K. All other values set to the defaults given by Table 1

Appendix C. Random Forest Classification of Model Dynamics

We trained a random forest model on time series features. We manually classified 100 randomly selected parameter combinations (10 runs per combination) from across the ranges defined in Table 1. Holding out 20% of parameter settings, cross-validation of random forest models achieved average accuracy of 0.96. Figure C1 reports the results of Shapley decomposition, an analytical tool for estimating the contribution of features to the output of machine learning models.

The human-defined features were selected to mimic the authors' intuitive process for identifying the dynamical class exhibited by a particular run of the model. Most of these features are derived from the mean (μ) and standard deviation (σ) of each signal's informational value summed and normalized across all agents across all time steps within a model run. This is calculated separately for the ingroup and outgroup. Within a single model run, we use the following 5 features calculated for both ingroup and outgroup to train the random forest model and 1 feature calculated for only the ingroup for a total of 11 features: (1,2) the mean signal value across all signals, $\mu(\mu)$; (3,4) the mean signal standard deviation, mean(σ); (5,6) the mean signal value of a signal peak, μ(Signal Peak); (7,8) the standard deviation of all signal peak values, σ(Signal Peak); (9,10) the probability that a signal goes extinct per time step (definitionally the same for in- and outgroup) P(Extinction); and (11) the proportion of ingroup interactions which successfully resulted in rewards for the agents, % Ingroup Success.

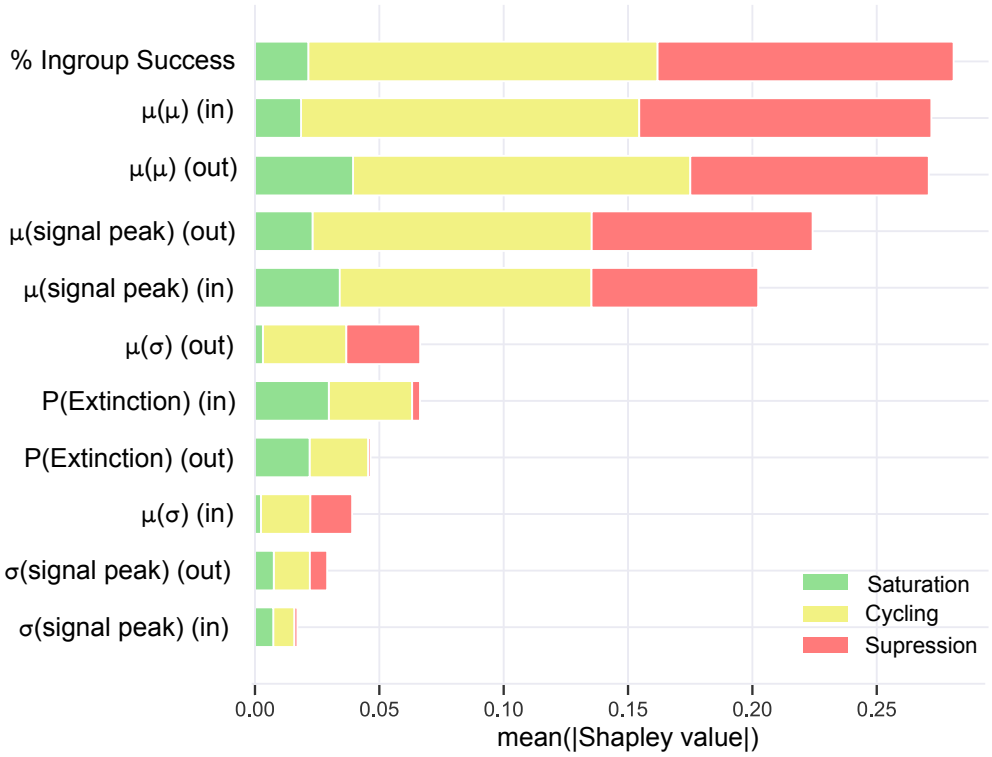


FIGURE C1. Shapley decomposition of the random forest model. Length of each bar shows the contribution of the feature to determining the probability of each class of model dynamics. A long bar in a particular color indicates that the feature is highly informative in determining whether it belongs to the class associated with that color. A short bar indicates that the feature is not very informative.

Appendix D. Parameter Sweeps from Other Initial Conditions

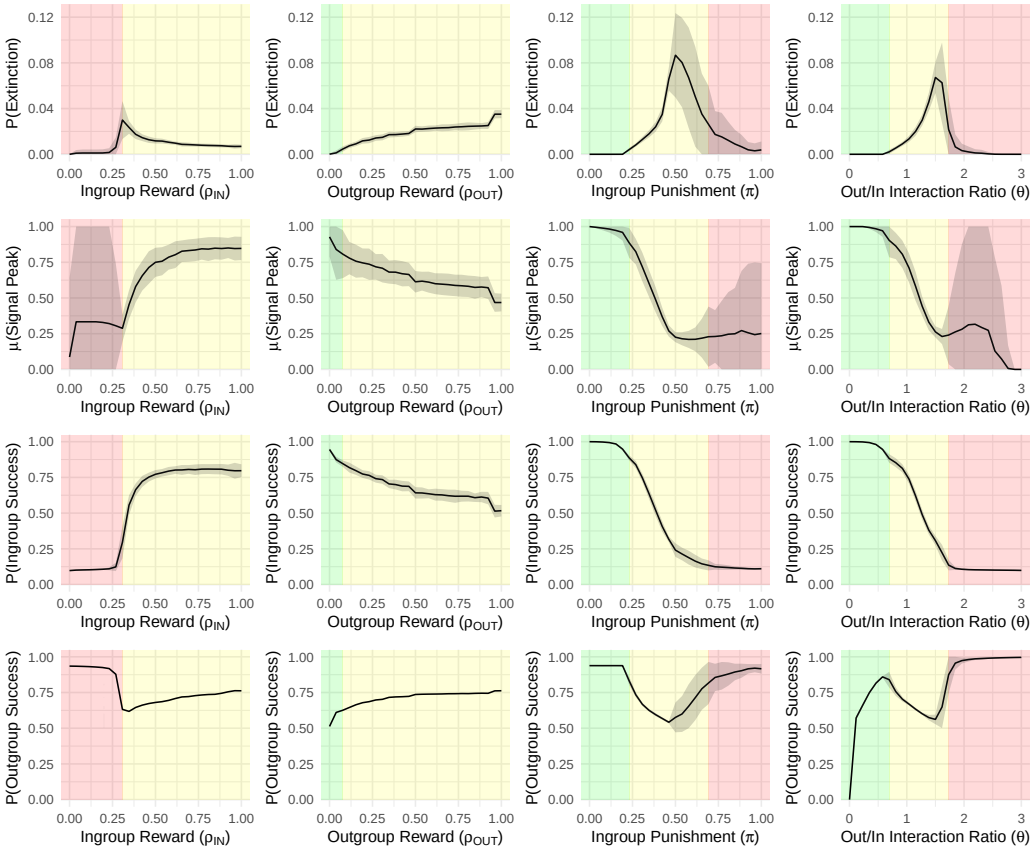


Figure D1. Sweeps of theoretically significant parameters holding all others constant as in Figure 4. Unlike Figure 4, which initializes with suppression-like conditions, wherein all ingroup signals are 0.1, these simulations initialize with saturation-like conditions, wherein all ingroup signals are 1. The patterns observed are nearly identical to Figure 4, differing only for the suppression regime for μ(Signal Peak). Here we observe a wide confidence interval indicating that sometimes 1000 time steps are sufficient to shift the signals away from 1, and sometimes they are not.

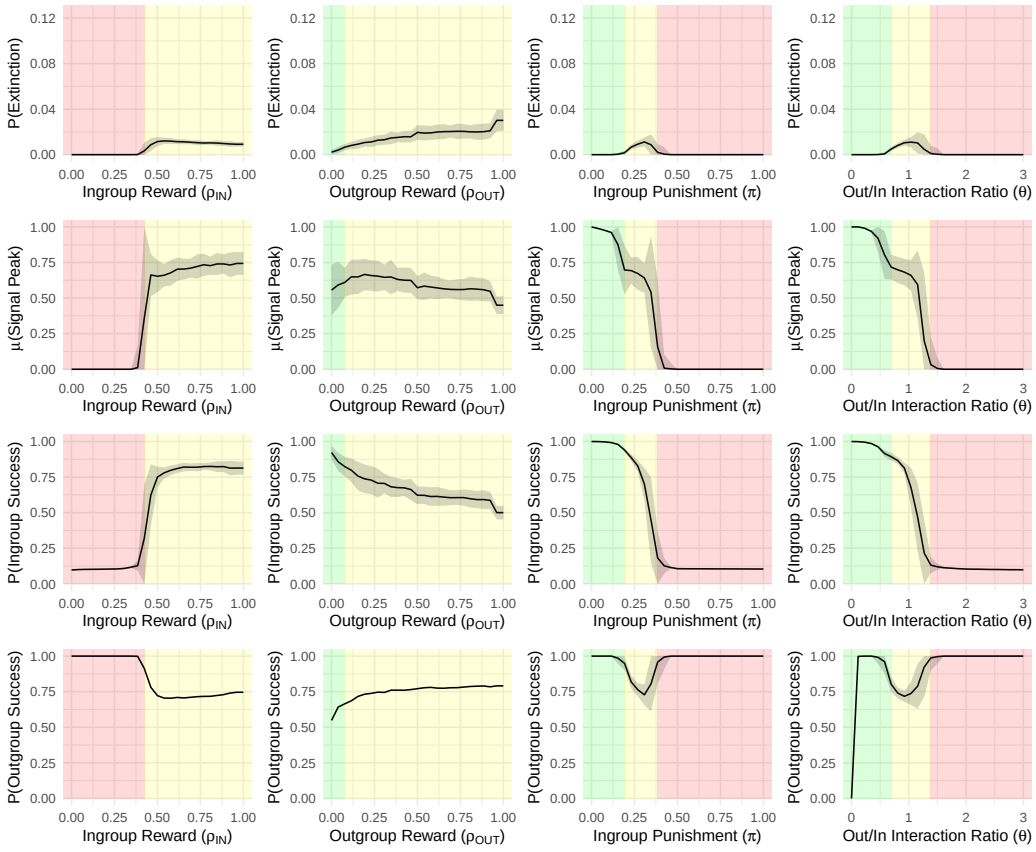


FIGURE D2. Sweeps of theoretically significant parameters holding all others constant as in Figure 4. Unlike Figure 4, which initializes with neutral conditions, wherein all ingroup signals are 0.1, these simulations initialize with saturation-like conditions, wherein all ingroup signals are 1. The patterns observed are similar to Figures 4 and D1, but differ notably due to artifacts of model design. First, P(Extinction), or rate of cycling, is lower throughout the regions of parameter space that produce cycling. Second, the region of parameter space that produced cycling in companion plots has partly given way to suppression. The cause of both discrepancies is similar and stems from the mechanism for signal extinction and novel signal emergence. Both in a real system and our model the ingroup will develop a novel signals unable to use existing ones. However, in our model novel signals emerge only after reaching a prevalence threshold in the total population, here set to an average informational value of 0.15. In some cases, ingroup reward is not sufficient to overcome ingroup penalty at the interaction rate to reach that 0.15 threshold, and thus no signals can be renewed and the system remains in a state of suppression. Similarly, the rate of extinction is low even when cycling because it takes a substantial portion of the 1000 time steps to reach the threshold and enter cycling. However, these are both consequences of the model design and number of time steps; cycling is a stable attractor from this suppression-like state.

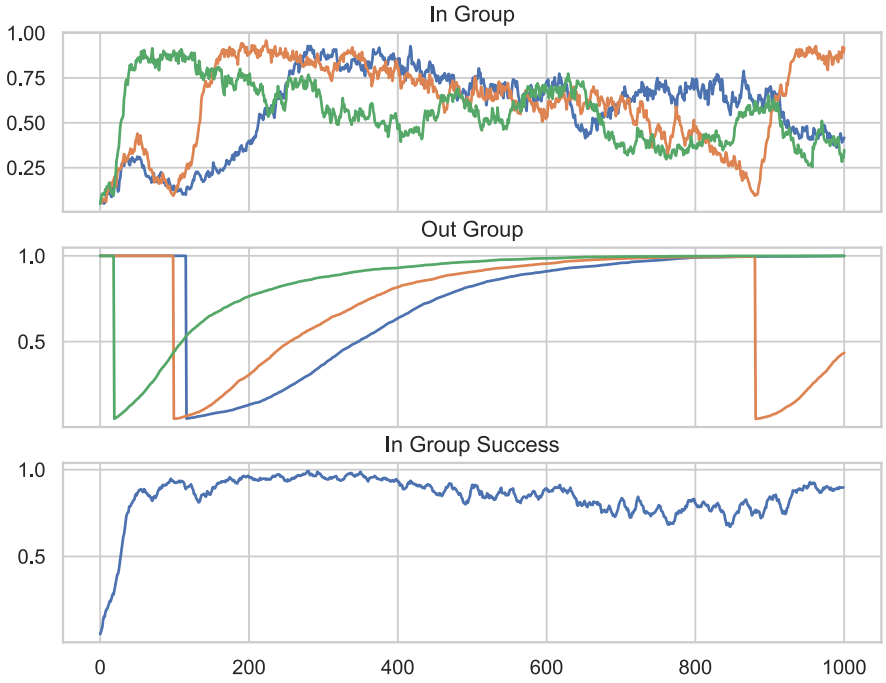


FIGURE D3. One run from initial suppression-like conditions, i.e., initial ingroup signals are set to 0.1 and initial outgroup signals are set to 1. The rate of outgroup/ingroup interaction is set to 0.7. All other parameters are set to the defaults given by Figure 1.