Dear Dr. Morris,

Thank you for your encouraging response and for the opportunity to further revise our manuscript to address your remaining concerns. My coauthor and I have endeavored to add clarity and precision to the manuscript, and we believe it is now improved.

I want to start by addressing what I perceive as a possible clash of perspectives on the manuscript. From reading your review, it may appear that we are attempting to simplify a very complex and nuanced set of processes regarding identity signaling and cooperative assortment. In other words, we are subtracting nuance. This is, of course, true to some extent. We are building a formal model, and formal models simplify. However, from our perspective, the project is actually *adding* nuance and psychological realism. This is because we see ourselves as building on a formal modeling literature related to tags and ethnic markers and their use for assortment on norms. Those models focus on coarse-grained, overt signals of identity that allow people to avoid dissimilar individuals and assort with similar individuals. In our work, we are attempting to say, “Not so fast.” Within groups of people who tend to cooperate with one another, there is still quite a bit of variation. People interact with many different people with different sets of norms in different contexts. Existing theories of signaling and assortment are not wrong, but should be expanded to account for this variation. I actually think this is quite consistent with your ideas about polyculturalism. Our paper is about how individuals view each other in complex ways and thereby adjust their signaling strategies in response to their likely audiences and the contexts for both signaling and assortment. The model is only a small part of the larger project to formalize these more detailed ideas about identity signaling (see my 2019 Behavioural Processes paper for a more complete view), but I believe it is an important step in that direction.

You will find our responses to specific comments below, with the original review in **black** and our response in **blue**. We hope you find this response and the accompanying revision satisfactory to warrant publication in Psychological Review.

Sincerely,

Paul Smaldino

From an economic perspective, it seems like you are modeling coordination rather than cooperation. Coordination requires matching: heads heads or tails tails. Cooperation is about forgoing opportunism and collaborating for the greater good. Of course, there is much evidence that positive interaction intentions are increased by homophily, but cooperative outcomes can come from differences. I would take the point even further than R3 did: economic analyses of cooperation emphasize that synergy comes from differences in priorities (that allow integrative tradeoffs) or differences in capacities/resources (that allow gains to trade). Even, say, in an interpersonal informal context, there are certain differences that foster a high joint-gain outcome. Identical twins are well suited to coordinate with each other, but less so to cooperate with real synergy.

It is true that we am primarily modeling coordination. However, coordination is intrinsically connected to generating the benefits of cooperation (see, for example, discussion in Calcott, 2008), so I don’t view these as separate. Cooperation does not require differences to generate synergy. In fact, some of the most fundamental aspects of generating cooperative benefits are joint attention, common ground, and shared norms. It is certainly true that some cooperative endeavors benefit from diversity, division of labor, and the like. However, as I have argued here and elsewhere, one still requires some baseline amount of similarity – that is, coordination – to achieve those cooperative benefits. We now explain this more clearly in the paper’s Introduction (In the section “Identity Signals Are For Cooperative Assortment”) in the following footnote:

“Many cooperative endeavors benefit from diversity and division of labor. Even in these cases, however, a baseline of similarity on shared norms, goals, and expectations are required to achieve those benefits (Bicchieri, 2005).”

You have run a simulation and obtained some results. Some of these results seem to fit the patterns generally observed in the world and some don't. Covert signaling is favored when it is needed and when it is efficient. That jibes. However, as R3 noted, "The finding that covert signals are more likely to be found in diverse communities where individuals are open and give people the benefit of the doubt seems to be somewhat at odds with the historical examples given."

We did our best to address R3’s comments in our previous revision, and to clarify the scope of the theory/model. We believe the theory of covert signaling presented in this paper does a good job of at least partially explaining much of signaling behavior. The comment is also difficult to address without any specific examples.

However, the results of a simulation are not interesting if its assumptions miss some of the most important dynamics of the real interaction, or involve processes that are patently unrealistic. Obviously, no model fits all the conditions that appear in the world, but does this model fit the canonical examples that you present as the motivating phenomena? If your results emerge from qualitatively different dynamics than those that we see in human societies, then they will not be of interest to psychologists, and would be better suited to some other kind of journal interested in parsimonious simulations. When neither the assumptions nor the results of a simulation bear a close fit to empirical reality, its usefulness to psychology is in doubt.

I agree with this. However, I also don’t believe there is cause for concern in this case. The model captures foundational aspects of human communication and interaction. I have presented this work to numerous audiences of both psychologists and anthropologists, and while it is acknowledged that there are limitations to the scope of the model (as with all models), the story it tells is generally viewed as insightful and compelling. I have also used the model as a basis for an large scale empirical experiment (manuscript currently in prep) and the results map very well onto the model predictions.

The approach we take with our modeling is more in line with that taken by evolutionary anthropologists and ecologists than those traditionally used by psychologists, which might explain some of the discomfort here. For example, our approach builds on that used by people like Boyd and Richerson in their influential book “Culture and the Evolutionary Process,” which uses models far simpler than ours to carve out an extremely powerful and influential theoretical framework. Part of our motivation for submitting this to Psychological Review is a strong conviction that psychologists \*should\* be interested in this sort of approach and the insights therein, and also because we have added additional psychological and sociological realism that that the scope aligns well with topics of interest to psychologists. The model is not and cannot be a complete theory of covert interactions, because to do so would require a book-length treatment. But I am strongly convinced that this is important work with real-world implications.

1. *Are the costs of overt signaling the same for all groups?*Many empirical facts (and commonsense) suggest that the costs of overt signalling may be higher for groups that lack prestige, that are stigmatized, or that are even illegal.

· The literature on power finds that people with high power act (even induced) show amplified self-expression and decreased sensitivity to audiences. People with low power have more active social perception and manage impressions by “covering” their identities more.

· Political dissidents in repressed societies such as Tibet or North Korea find ways to express themselves that allow for plausible deniability, whereas those in the Netherlands or Germany are more overt.

· There is longitudinal evidence from changes in gay culture: In the UK the cryptolect Polari fell out of use as homosexuality was decriminalized. In the US, the use of subtle fashion signifiers like earrings and caps has declined.

As R3 notes, it doesn't seem too hard for you to explore this in your modeling, and the work feels incomplete and perhaps erroneous without it.

Yes, we did not vary the costs between groups in the minority-majority condition (the only one that uses explicit group structure). It is true that in reality costs can differ greatly for different groups. Given that the costs should fall more strongly on the minority individuals, this indicates that, if anything, covert signaling should be MORE preferentially favored in those groups with differential costs. The model is therefore a strong test of the hypothesis. I also think that it is also unreasonable to propose that all potentially meaningful variation is included in one modeling study. I would argue that we are in fact extremely thorough in our exploration of the model’s parameter space. It is worth noting that we are well aware of this limitation, and have been in communication with my colleague John Bunce at MPI-EVA who has studied power imbalances between groups using related methods for an eventual collaboration to explore just these issues. I believe a study of differential costs would involve detailed study in its own right, and is beyond the scope of the current analysis. All that said, we have edited the section on minority status to acknowledge the existence of and likely importance of differential costs. Specifically, we inserted the following text:

“Our model assumes that the costs of disliking are equal for both majority and minority individuals. In reality, costs can differ between groups, and are likely to be higher for minority individuals (Bunce and McElreath, 2018). Our earlier analyses of costs indicate that the effects of minority status, diversity, and differential cost should all be additive. That is, under higher costs, we should expect even more covert signaling among minority individuals, relative to majority individuals.”

The political dissidences example is actually consistent with a higher cost of being detected in the former countries (Tibet, North Korea), and hence is also consistent with the model. The idea that signaling strategies would change over time, in response to cultural shifts --- e.g., signals become more widely known, costs and opportunities for assortment change --- is an important feature to add to future versions of the model. This was already discussed in the Discussion section, but we have rewritten the relevant paragraphs for added clarity. See in particular the paragraph beginning with “Our theory requires that covert signals remain obscured to dissimilar individuals.”

The real world is more nuanced than the model. Of course, the model is a simplification, but such simplifications are essential to make any theoretical progress (among other places, this position is well articulated in Kieran Healy’s 2017 paper in Sociological Theory, “F\*ck nuance.”) We have done our best to acknowledge the scope and applicability of the model, as well as its current limitations. Ultimately, formal theories of social behavior cannot progress without some amount of abstraction and loss of nuance. With all modesty, the model is already more realistic and dynamic than 99% of economic models of signaling or coordination (I am the sole author of a forthcoming textbook on modeling social behavior for Princeton University Press, so this is not merely an idle claim, but one that comes from a deep familiarity with the modeling literature).

*2. Assumption that all signaling is honest.*

Life is a constant process of identity negotiation. People often front identities that they don't hold in the eyes of others, wittingly or unwittingly. In economics (the work on signaling by Spence) this is the heart of signaling -- it goes along with the assumption that a signal is less costly to send when it is true than when it is false. In the case of sexual orientation, the phenomena of gays passing for straight is still widespread in much of the world (in the workplace and in marriages) and queerbaiting is an increasingly-heard accusation in the art world. In both cases it can be costly to be found out, albeit less in the former case given the majority/power position of straight people, even if its reversed within particular industries or circles.

I can understand your assumption, possibly, if you are just trying to explain conversational implicatures. But your paper is quite unclear about the scope of the phenomena you are trying to model. When it comes to social identities, there is a constant process of identity negotiation in which new entrants try to make identity claims and gatekeepers of the community vet them and sanction them.

I agree with most of this. The focus on honest signaling is a modeling necessity to keep the whole thing analyzable. Deception and phenomena like “passing” are super important features of human signaling, and we do not mean to suggest or imply otherwise. And indeed, covert signals may implicitly involve some deception, at least insofar as they allow people to avoid detection (though not in the sense where people claim to be something they are not). The present work is in the tradition of honest signals for coordinating on conventions/norm, and is in fact quite a bit more nuanced than most of those. It is an attempt to build more psychology into those theories. The progress is necessarily incremental, because social dynamics are complicated. That said, the scope is greater than conversational implicatures. It extends to fashion, language, and other markers (e.g. accents).

We tried very hard to be clear with our model assumptions and the limitations of our conclusions. I am not sure what else we can do on that front. In particular, we had already discussed the limitation of restricting our discussion to honest signals – there is an entire paragraph in the Discussion dedicated to this limitation. We believe that many signals are in fact honest.

*3. Assumption that an actor sends the same signal regardless of the audience’s characteristics.*

In a broadcast Anderson Cooper may choose words that mean different things to different parts of the audience. A politician using dogwhistle codes may do the same. But in dyadic interactions we can choose different words for each new audience. That's what Relevance Theory is chiefly about -- conversations -- and it assumes the speaker elaborately tailors the message, and the audience takes such tailoring metacognitively into account. Are you modelling a series of communications to a broad audience? Are you assuming that all audience's characteristics are invisible to the speaker? This starts to seem like a very different phenomenon than conversations.

Yes, the model is specifically focusing on signals that are, at least in theory, perceivable by a broad audience. The example that now introduces the paper is about Twitter, which is about as public as one can get. The model uses classic feature of evolutionary modeling, which is to assume behavioral strategies that are minimally cognitive, so as to assess the relative fitness of the behaviors themselves. We discuss this in terms of the “behavioral gambit” in the Discussion section. It is possible that individuals might adjust their signaling strategies based on their \*perceived\* audiences (I am currently working on this idea with a grad student with a background in computational linguistics). In theory, the model could apply to dyadic conversations if they also include these kinds of signals (such as wearing certain clothes, making certain encrypted jokes). In the revision, we now spend more time talking about the assumptions concerning audience makeup. For example, in the Introduction, we have altered the following sentence (addition in bold):

“A potential solution is the use of overt or conspicuous displays---often called tags or ethnic markers---**that are perceivable by a broad audience**.”

And at the beginning of the section “A theory of covert signaling,” we add the following to our definition of covert signaling:

“Implicit in this definition is the assumption of public communication---that is, of a broad audience with characteristics that are beyond the control of the signaler.”

*4. Assumption that the cultural learning operates across the whole population*

"Individuals engage in success-biased copying, whereby the signaling and receiving strategies of individuals with higher payoffs are preferentially copied." Can you clarify what are the receiving strategies? Is it the generous/churlish distinction?

Yes, the receiving strategies are generous vs. churlish. We have now added parentheticals to clarify this:

“Individuals engage in success-biased copying, whereby the signaling **(overt vs. covert)** and receiving **(generous vs. churlish)** strategies of individuals with higher payoffs are preferentially copied.”

A lot of the work in this kind of model is done by the cultural learning process. But does prestige or success learning operate across the lines of groups? How do people learn from the covert signaling styles of outsiders if they can't even detect those covert signals in the first place? These seem like assumptions that are not co-tenable.

Yes. It is a cultural evolutionary model, so it makes sense that the learning process drives the model dynamics. One could imagine that prestigious or successful individuals are more likely to teach or otherwise be targets for social learning, and that what is learned is general signaling/receiving strategies. The specific signals do not have to be transmitted, only the tendencies to signal in a covert vs. overt way (and likewise to default to neutral or negative attitudes). This does not require that the specific signals are detected by learners. For example, children learn how to speak without learning how to generate every specific sentence. To clarify this, we have added the following text to the model description under “Payoff-biased copying”:

“Note that payoff-biased copying need not imply the transmission of any particular signals, which would need to be learned from specific individuals and would change over time in a process not modeled here. Rather, what is transmitted are the tendencies to signal in a covert vs. overt way, and likewise to default to neutral or negative attitudes when acting as a receiver.”

**Smaller concerns:**

You just assert and don't explain your conclusion that machines will not be able to pick up on covert signals. Why not? Kosinksi's "gaydar machine" research seems to suggest the opposite. Is it your view that the facial signals picked up were all overt? What about the many prior papers by him and others finding that digital traces and actual behavioral residue allows inference of people's sexual orientation, political orientation, and so forth?

Kosinksi’s “gaydar machine” relies on a number of fallacies of both statistics and social science, and its success is almost certainly an artifact of the training data. Your colleague Andrew Gelman has written a compelling critique of that work (Gelman et al. 2018, “Gaydar and the fallacy of decontextualized measurement”). Our more general claim about AIs not being able to detect covert signals is perhaps a hope and not a definitive conclusion, but we frame it explicitly as such (“we believe it unlikely”) and so stand by our proposal.

The words churlish and generous don’t seem optimal to describe the constructs you model. Churlish means rude or mean. Generous means giving. These suggest dispositional strategies: never cooperate or always cooperate. What you are describing are conditional strategies--whether or not people cooperate/coordinate in uncertainty about the other’s characteristics. "Tolerant/intolerant" would be better because they reference attitudes toward the other's traits. You could define them as about ambiguity tolerance -- an important construct in political psychology. "Trusting” and “nontrusting” might better capture it because they refer to conditional strategies -- orientation to uncertainty about the other’s characteristics.

There is always a danger in mapping words used colloquially to technical meanings. It is possible that there are better choices for the receiving strategies. However, we have already used these terms in our previous publication on covert signaling (Smaldino et al. 2018, Scientific Reports). I think it would be worse to introduce new terms inconsistent with the previous paper, so we have opted to keep them.

You provide no evidence that covert signaling is “ubiquitous.” I don't think that you need that claim.

This is a fair point. We have changed this to “Yet the existence of covert signals suggests some adaptive function.”

The issue is not interactions with “strangers” (individuals with whom you are unaquainted) but with “outsiders” (people who are not in your cultural group and thus less able and less inclined to coordinate).

It’s both. The issue is that you don’t know what sort of person a stranger is (an insider or outsider) until you exchange signals. You want to identify other insiders, without highlighting your differences to outsiders.

It's not always the case that criminal organizations demand overt signals. In many cases criminals, like spies, depend on not being detected as such.

Absolutely. It was not our intention to suggest otherwise, only to highlight one example in which it may be strategic to highlight differences. We now clarify this in a footnote:

“On the other hand, covert signaling may be quite common among criminals, as their livelihood depends on not being detected as such (Gambetta, 2009).”

It's an overstatement to say that ethnic markers lack ambiguity – they are always more known to insiders than outsiders. Do you know what the different color turbans mean in Bangalore, or different length Djeballas mean in Marakech?

This is true, though in that case the markers are still not ambiguous for members of the local cultural community. I don’t know what they mean, but I am also extremely unlikely to be in a position where that matters. The model is specifically addressing members of an diverse population of potentially interacting people (it also provides for the possibility that not everyone recognizes even overt signals when R < 1). While it may be a slight overstatement to say that ethnic markers are always completely unambiguous, clarity is nevertheless their purpose. They lack the sort of strategic ambiguity present in, say, a political dog whistle. That said, we address your concern by adding the possibility of the sort of error your suggest to the relevant passage:

“**By design**, overt signals like ethnic markers lack ambiguity. They are visible to all, and **most potential receivers will know** what they mean.”