Proceedings of SALT 31: 703–719, 2021

Dogwhistles: Persona and Ideology*

Robert Henderson
University of Arizona

Elin McCready
Aoyama Gakuin University

Abstract A dogwhistle is a piece of language that sends one message to an outgroup while at the same time sending a second (often taboo, controversial, or inflammatory) message to an ingroup. We propose an analysis of dogwhistles in the setting of social meaning games that treats them as signaling the persona of the speaker, and in some circumstances enabling an enrichment of the conventional meaning of the expression through the connections of social personas and ideologies, which we model formally. We show that this account improves over pitfalls encountered in other accounts (Henderson & McCready 2019b; Khoo 2017; Stanley 2015), which includes in some of our own previous work. We further show how this formal framework allows, not just a account of dogwhistles, but opens up a way to analyze a variety of sociopragmatic phenomena like unconscious bias and epistemic hypervigilance.

Keywords: dogwhistles, sociolinguistics, game theory, pragmatics

1 Intro

George Bush's 2003 State of the Union address contains the following line.

(1) Yet there's power—wonder-working power—in the goodness and idealism and faith of the American people.

To most people this sounds like, at worst, a civil-religious banality, but to a certain segment of the population the phrase *wonder-working power* is intimately connected to their conception and worship of Jesus because it comes from a well known evangelical hymn. When someone says (1), they immediately know the speaker is an evangelical Christian, and that they are not expressing civil-religious banalities, but instead making a veiled religious appeal, even a sectarian one.

On a 2014 radio program, Representative Paul Ryan said the following.

* We thank Nicholas Asher, Tatjana Scheffler, and Malte Willer for help and comments on the current manuscript, as well as audiences at Probability and Meaning 2020, the Gothenburg Dogwhistle Roundtable, Stanford's Cognition and Language Workshop, the University of Cologne, ZAS Berlin, Tübingen University, the Ohio State University, and Queen Mary University of London for their incisive comments and questions.

©2021 Robert Henderson, Elin McCready

(2) We have got this tailspin of culture, in our inner cities in particular, of men not working and just generations of men not even thinking about working or learning the value and the culture of work.

He was criticized shortly after for making a "thinly veiled racial attack": the phrase *inner-city* is code or euphemism for Black neighborhoods (especially stereotypically racialized views of such neighborhoods). Thus, to a audience aware of the dogwhistle, Ryan was not making reference to neighborhoods across the city filled with lazy people, but to Black neighborhoods in particular.

These examples illustrate the notion of a *dogwhistle*—that is, language that sends one message to an outgroup while at the same time sending a second (often taboo, controversial, or inflammatory) message to an ingroup. We do not pick these two examples arbitrarily, but because they illustrate a distinction between two kinds of dogwhistles that we have introduced in previous work (Henderson & McCready 2019b), namely *identifying* dogwhistles and *enriching* dogwhistles. The intuition is that dogwhistles like *wonder-working power* in (1) mostly signal facts about the speaker's social / sociolinguistic persona. All listeners know that *wonder-working power* denotes that subset of powers that has the capacity to produce miracles, but only a subset of the audience may recognize that this is a stock evangelical phrase, and so the speaker is assuming that particular religious identity. In contrast, the dogwhistle *inner-city* in (2) has its denotation altered in virtue of being detected as a dogwhistle. An unsavvy listener may take it to refer neighborhoods in the city, but a listener who detects the dogwhistles knows that it specifically refers to Black neighborhoods in the city—that is, its semantic content is enriched in some way.

In previous work (Henderson & McCready 2019b), we have defended an account in which enriching dogwhistles are a kind of special case of identifying dogwhistles. That is, enriching dogwhistles are conventionally associated with multiple semantic meanings, but accessing certain meanings is dependent on identifying the speaker as bearing a certain social persona. While we think the rough outlines of this account are correct, namely that enriching dogwhistles are a special case of identifying dogwhistles that precede by identifying the speaker's persona, we have come to believe that conventionalizing the enriched meanings of enriching dogwhistles is wrong. The primary goal of this paper is thus to propose a new account of enriching dogwhistles that shows how enrichments can be generated by allowing sociolinguistic personas to be linked to ideologies. When a speaker uses a dogwhistle, and is detected, the listener will be able to enrich the meanings of certain expressions in virtue of the fact that speaker is likely to have ideological commitments associated with the persona they bear in virtue of using the dogwhistle. An important advantage of the proposal, in addition to the empirical advantages discussed in section 4.2, is that it provides a partial rapprochement with the purely

pragmatic account of dogwhistles developed by Khoo 2017 which strongly argues against conventionalizing the meaning of dogwhistles.

We begin by reviewing previous accounts of dogwhistles, in particular, that of Stanley (2015), who provides a semantic / pragmatic proposal, where dogwhistles are Pottsian CIs, contributing an at-issue component for the outgroup audience and a non-at-issue component that potentially only the ingroup is sensitive to. We additionally consider the account in Khoo 2017, which provides a purely pragmatic account, where dogwhistles involve certain default inferences. We see these two accounts as opposing poles. A Stanley-style account involves maximal amounts of conventionalization albeit in the not-at-issue domain. In contrast, a Khoo-style account is highly deflationary—dogwhistles involve no conventionalized content, nor Gricean implicature, but bog-standard default inferences interlocutors make all the time, whether involving so-called dogwhistles or not.

Our approach ends up somewhere in the middle. As we will see as we develop our own account, dogwhistles do not involve conventionalized truth conditional meaning (contra Stanley and our own previous work on enriching dogwhistles), but instead only involve a kind of conventionalization enabled by social meaning. These social meanings can be linked to ideologies which help generate the kinds of inferences we see with enriching dogwhistles, which Khoo aims to understand using default inferences. Our account differs from his, at the deflationary pole, in virtue of being much thicker, involving conventionalized social meanings and well as a theory of ideologies, rather than just default inferences. As we will argue, we believe this thicker theory is warranted, and we will consider some extensions that we can make once we admit ideologies into our framework and connect them to personas. In particular, it will allow us to understand a widespread phenomenon, especially in online spaces, which we call *epistemic hypervigilance*.

2 Previous work

A major question for dogwhistles is whether the dogwhistled content is conventionalized or not, and in what way. Stanley 2015 proposes a conventional implicature account in which dogwhistle language involves a conventional non-at-issue component along the lines of more familiar expressions like slurs, honorifics, etc. For instance, just as a slur like *kraut* would have AI-component “German” and a NAI-component "I hate Germans", a dogwhistle like *welfare* would have AI-component “the SNAP program” and a NAI-component “Blacks are lazy”. In general, terms which carry both AI and NAI components can be referred to as *mixed content bearers*. The problem is that a mixed-content account of dogwhistles is untenable because there are strong arguments that dogwhistled content is not conventionalized, whether we are talking about the NAI or AI dimensions.

The argument concerns 'what is said' by a dogwhistle. The use of dogwhistles is prompted by a desire to 'veil' a bit of content, but still to convey it in some manner. Deniability is essential. If a bit of content is conventional, it's not deniable any longer. This can be seen with pejoratives, which clearly carry conventional NAI content. We see this from the following example using the pejorative *kraut*, which is clearly a mixed-content bearer.

(3) A: Angela Merkel is a kraut.
B: What do you have against Germans?
A: # I don't have anything against Germans. I'm just talking about Merkel's nationality.

In contrast, similar dialogues are fine with dogwhistles. In the following, there seems to be no entailment that A has the relevant attitude.

(4) A: Elin is living high on the hog on welfare again.
B: What do you have against poor people?
A: I don't have anything against poor people. I'm just saying Elin is on welfare and I saw her buying steak at the store.

By this test, dogwhistles can be concluded not to be conventional. Moreover, it makes sense that dogwhistles should fail such a test. When people are called out for using a dogwhistle, they commonly deny that they were dogwhistling. This would be impossible if the dogwhistle bore conventional content, even NAI content, reducing the efficacy of dogwhistling as a conversational strategy.

The argument from examples like (4) shows that dogwhistles cannot bear conventionalized truth-conditional content. At the same time, there are problems with accounts which take there to be no conventionalized aspect to dogwhistles. We see such an account in the inferentialist approach to dogwhistles in Khoo 2017. In this account dogwhistles involve default inferences. That is, the speaker claims that x is C and the interpreter believes that C's are R's, then the interpreter will conclude that x is R; it's this kind of inference that Khoo thinks that dogwhistles license. More concretely, if the interpreter believes that *inner-city* neighborhoods are Black neighborhoods, then the speaker saying that people who live in inner-city neighborhoods lack a culture of work licenses the inference that people who live in Black neighborhoods lack a culture of work.

Notice that this kind of inferentialist account relies on the (at-issue) content of the dogwhistle itself and the background beliefs interpreters to license a constellation of inferences about things related to that content. An advantage of such an account is that dogwhistles are now deniable. A speaker can just deny that they hold the background beliefs which lead to the inference.

The inferentialist account makes sense of the fact that dogwhistles are deniable, but it has problems. In particular, Khoo's inference follows from the expression's truth conditions. Thus, any expression with the same truth conditions should dogwhistle. This is not true. A phrase like *downtown neighborhoods* doesn't dogwhistle like *inner city* does; the situation is similar for *welfare* and paraphrases like *assistance to the poor*. This suggests that while dogwhistles must not bear conventionalized content, some expressions are singled out as something like "dogwhistle expressions", and so there is some kind of conventionalization.

With this two arguments we see that we have to strike some kind of middle course with respect to conventionalization. Dogwhistled content is not part of conventional truth-content content, so speakers are able to avoid (complete) responsibility for what they convey. At the the same time, dogwhistles can be identified as such, even if not bearing conventional content, so they must be conventionally distinguished from other expressions with the same truth conditions. We solve this problem in the next section through an account which has dogwhistles conventionally denote, not in the truth conditional domain, but in the domain of social meaning. This explains why dogwhistles are conventionalized as such, solving the substitutability problem we identified for an invited inference account in the style of Khoo 2017. We then show how to deal with enriching dogwhistles, which we argue arise when listeners use an ideology associated with a sociolinguistic persona to license inferences that enrich the truth conditional content of the dogwhistle. Because social personas are only probabilistically assigned, a speaker can always deny the persona, and thus any downstream enrichments, accounting for the fact that dogwhistled content is deniable.

3 Our account

In recent work, Burnett (2016, 2017) pioneers the use of Bayesian signaling games to model identity construction through sociolinguistic variation. We take identifying dogwhistles to work via an only slightly more complex sort of sociolinguistic identity construction than the kind Burnett (2016, 2017) discusses. After building the basic formal system for identifying dogwhistles, we show that enriching dogwhistles are a special subtype of identifying dogwhistle which interact with ideological backgrounds.

Burnett's Social Meaning Games which have the following simplified architecture (which we modify / elaborate further below):

- Players
  - a speaker S

  - a listener L

- Actions for players

  - The speaker chooses a persona p from the space of personas P
  - Based on their persona, the speaker chooses a message $m \in M$ to send to the listener.
  - Based on the message, the listener chooses a response $r \in R$, which in the simplest case we can identify with selecting an element of P—i.e., identifying the speaker's persona.

- Utility functions for players

  - U_S/U_R—functions from $P \times M \times R$ to $\mathbb{R}$, which represents payoffs for every possible combination of actions.

We want the dogwhistle effect to arise from listeners being unaware (or uncertain) about the close connection between some bit of language an a persona. That is, we want listeners to have beliefs about a speaker's persona, but also beliefs about how personas and messages are connected. To model this, we assume listeners have prior over P, but also beliefs about $P(m|p)$—namely how closely messages are linked to particular personas.

We can now update a listener's belief about the speaker's persona given their message by doing Bayesian inference.

(5) $P(p|m) \propto P(p)P(m|p)$
'The probability of a persona given a message is proportional to prior probability of the persona and the likelihood of sending that message given that persona'

This is a extension of Burnett (2016, 2017) only in that she takes social meanings to be fully lexicalized, i.e., the likelihood of $P(m|p) = 1$ when p and m are consistent. It is a very mild extension, though, which we take to be a nice feature of our analysis. It strikes us as a positive that we can model dogwhistles with the same ingredients use to model third-wave vacationist sociolinguistics more broadly.

The final ingredient of the analysis is to better specify the utility functions. The listener's utility is straightforward—it is maximized by extracting as much information from a message as possible about a speaker's persona—that is, by doing doing Bayesian inference as just described. Utility for speakers is more complex because unlike in many signaling games, the speaker doesn't just pick messages based on some type assigned by nature—i.e., they don't just *report* their personas. Instead,

speakers have preferences for different personas, some of which may be dependent on how the listener would react to that persona. Thus, we must allow for speakers to "construct" a persona in concert with their listeners. Speakers want to present themselves in a certain way, but speakers will also be sensitive to whether listeners will approve of that persona or not. In adversarial contexts, a speaker might have to juggle presenting a safe persona with a persona they might prefer to present (or prefer to present to another audience that might be listening)—this is when dogwhistle language become useful.

Along these lines, we follow Burnett (2017) and Yoon, Tessler, Goodman & Frank (2016) in assuming that the utility calculation takes into account the message's social value, which is given by two functions:

- The speaker has a function ν_S that assigns a positive real number to each persona representing their preferences.
- The listener has a function ν_L that assigns a real number (positive or negative) to each persona representing their (dis)approval.

We can now calculate the speaker's utility. The utility is dependent on the affective values of the range of personas consistent with the message and the likelihood that the particular persona is recovered given the message, as follows:

(6) $U_S^{Soc}(m,L) = \sum_{p\in[m]} P(p|m)+\nu_S(p)P(p|m) + \nu_L(p)P(p|m)$

When only one listener is addressed, dogwhistles reduce to ordinary social meaning; the speaker should choose a signal which maximizes U_S^{Soc}. This will be the message conveying a persona that the speaker and listener jointly like the most, weighted by the probability that the listener will actually assign the speaker that persona based on the message. Dogwhistles come into their own when speakers address groups of individuals with mixed preferences over personas, different priors for the speaker's persona, and different experiences about the likelihood of a persona given a message. The simplest way to assign utilities to the group case is to sum over all listeners; we will assume this metric in the following.

(7) $U_S^{Soc}(m,G) = \sum_{L\in G} U_S^{Soc}(m,L)$

With this utility function, the basic prediction is that speakers will use language that maximizes their social utility with respect to a group of listeners. For the dogwhistle case, this happens when using the dogwhistle allows gain of higher social utility than otherwise with respect to the entire group—i.e., when the dogwhistle gives benefit for some 'savvy' listeners while avoiding deficits that would come from speakers disliking the persona but oblivious to the dogwhistle.

In formal terms, oblivious listeners are oblivious to the fact that $P(m|p)$ is high for some offensive persona p and dogwhistle m. This means that while they would greatly penalize the speaker (in terms of utility) for having persona p, that penalty can be discounted because they are not likely to assign the speaker p based on hearing the dogwhistle m. In contrast, ingroup members who are aware of the dogwhistle, and likely approve personas linked to the dogwhistle, will both reward the speaker (in terms of utility) for having persona p, and have a high likelihood of assigning the speaker p given the message m.

This is precisely how we would model identifying dogwhistles like *wonder-working power* in (1). This evangelical catchphrase would be recognized by evangelicals, who would assign Bush the 'evangelical' persona, a persona that both speaker and listener value, yielding a high utility payoff. In contrast, listeners who might disapprove of evangelicals would not recognize that this phrase was related to that particular faith, and so while they would otherwise punish Bush in terms of utility for using a phrase associated with the offensive persona, Bush can discount this penalty because he is not likely to be detected by these unsavvy listeners. In this situation, Bush's social utility is maximized by using the identifying dogwhistle.

4 Enrichment

Having briefly presented our account of identifying dogwhistles, we now are in a position to turn to the main object of this paper, enriching dogwhistles. To make sense of how enrichment works in dogwhistles, we believe that we must achieve a new kind of synthesis. We think the Khoo-style account we critiqued earlier, which focuses on enriching dogwhistles, fails short for not providing an integrated treatment of identifying dogwhistles. Since we think that identifying dogwhistles are the simpler case, it follows that a treatment of enrichment should be an extension of how identifying dogwhistles are analyzed: for us, this means that social meanings must be involved. At the same time, our previous account of enriching dogwhistles in terms of pragmatic enrichments falls short as it failed to make the precise connection between enrichment and personas clear.

We thus propose an account on which recognition of speaker persona invites certain kinds of inferences, which result in alterations of the meaning recovered by 'savvy' interpreters. In particular certain kinds of personas, mainly those associated with ideologies and political stances, 'project' sets of beliefs and values. Such projections enable certain kinds of invited inferences which, we claim, ground the phenomenon of enriching dogwhistles. Note that this approach allows a partial reconciling with Khoo, at least for enriching dogwhistles. That is, dogwhistle invite inferences, but those inference do not follow from the semantic content of what was said. The inferences follow from the social meanings of the relevant expressions,

which, under our proposal is linked with ideologies. Our new account thus improves on the problems of both Khoo and our earlier account: we have an explanation of what ties inference to expression, and of what grounds the persona-inference connection.

We take enriching dogwhistles to be *identifying dogwhistles*$^+$ in the following sense. On use of the dogwhistle, a savvy listener detects it and assigns the speaker a relevant persona. Those personas are associated with ideologies, which come with background assumptions. The listener, consciously or not, learns what ideological grounds the speaker is speaking on. The savvy listener can then draw inferences about the speaker's intended content. But sometimes, when the listener is not conscious of this process, those grounds invite the listener to make similar assumptions, and the listener gets sucked into the ideological maelstrom.

To make all this precise, our tasks are twofold. First, we must make clear and work into our formal model what ideologies are and how they can be entangled with personas. Second, we need to "bridge the gap" between speaker and hearer, that is, understand how recognizing the ideological grounds on which the speaker is speaking can cause the listener to make inferences about the speaker's communicative intent, and, sometimes, influence listener behavior. We turn to these tasks now.

4.1 Ideologies: formal treatment

Ideologies as we use them in this paper are construed as ways of viewing the world, usually with a political dimension (or resulting in political effects, as with religion). Ideologies thought of like this are obviously complicated, and it is not our main purpose to give a full formal treatment of them and their effects here. Our goal is smaller: just to show how ideologies associate with personas, and what effects recognizing a persona and its related ideology or ideologies have on linguistic interpretation. This task is much simpler.

For this less ambitious purpose, it seems to us that two elements are needed to make sense of ideologies as they relate to personas and dogwhistles, especially in the context of enrichment. First, ideologies indicate affect: the (dis)approval of various actions or people, and the ideology holder's way of valuing various aspects of the world. Second, ideologies bring in more global assumptions about the world: what kinds of things are true, what sorts of stereotypical associations hold between various groups, what properties different kinds of people have, and so on.. Thus, to understand what effects assigning ideologically related personas to discourse agents has, we need at minimum a way of valuating actions and individuals and a way of introducing beliefs and world knowledge to our models.

4.1.1 Assigning affective values

In order to model the way ideologies affect judgments about approval and disapproval, we need a function that can assign affective values to objects relevant to ideologies and personas. We will use ρ ('rate') for our new function. This function takes individuals as input and yields real numbers as value: we allow both positive and negative real numbers here, as with the listener valuation function v_L on personas presented in the last section. Values yielded by ρ must include attitudes toward particular individuals such as Trump, where value assignments may be relatively extreme, but also attitudes toward behaviors, groups of people, and properties. Consequently we need a function which takes such things as input. We are going to take the individual case as basic and treat the other kinds of inputs needed – actions, properties, groups – as individuals by making use of the kind-mapping function '$\cap$' (Chierchia & Turner 1988; Chierchia 1998). This function is more standardly used to model nominalizations (self-predication, as in Chierchia & Turner 1988) and bare nominals in languages like Chinese and Japanese (Chierchia 1998). We use this function to produce kind terms corresponding to properties more generally, which, we take it, also includes ways of being ('revolutionary', for example).

(8) Being nice is nice.
$nice(\cap nice)$

(9) a. inu-wa hoeru
dog-Top barks
'Dogs bark.'
b. $barks(\cap dog)$

4.1.2 Ideologies: epistemic bases

Ideologies assign value, which we model using ρ; they also comprise sets of beliefs about how the world is: the kinds of things that make it up, the properties of kinds of people, systems, and objects, and the causal elements that induce and condition change. The truth-evaluable elements which make up an ideology are modelable as sets of propositions. We call each set of this kind the *basis* of an ideology. We use the notation $\mathscr{B}$ for ideological bases.

What sorts of propositions form the basis of ideologies? The answers to this question are as various as ideologies themselves. For instance, anti-vaxx ideology takes it as a given that vaccines have negative effects, and that they are promoted by pharmaceutical companies as a part of exploitative capitalist strategies. QAnon ideology takes the existence of a conspiracy with bizarre goals as a given. Racist

ideologies involve beliefs about the relative value and superiority of ethnic groups, and so on. Ultimately the content of ideological bases is as various as the ideologies themselves. In this context, we are restricting our attention to ideologies with political content, because of our subject matter: dogwhistles as usually construed are confined to political discourse, as opposed to other kinds of covert communication. For some nonpolitical ideologies, it might be that we need a substantially different kind of formal model which puts less emphasis on propositions (for instance certain kinds of aesthetic views which we might take to be ideological).

All these beliefs can function to bridge gaps in reasoning and connect things that without the ideology would be nonobvious. For instance, it might not seem plausible to the nonsubscriber to white supremacist ideology that when a nonwhite person is hired for a university position, the reason must involve affirmative action. But to the white supremacist of a certain stripe, it will be obvious, because of the beliefs they hold as a result of subscribing to the white supremacist ideology. We will argue that it is these sorts of beliefs, and ideological bases in general, that trigger enriching dogwhistles.

In our model, then, ideologies related to personas have the form $\iota = \langle \rho, \mathscr{B} \rangle$ and so consist of pairs of affect-assigning functions and ideological bases.

We should clear up two points before proceeding.

First, the propositions comprising the basis of an ideology can be somewhat indeterminate and vary from individual to individual depending on where they have acquired their ideological beliefs. So we must think in terms of related but possibly non-identical ideologies, which we can view as ideological equivalence classes. We thus define the basis of an ideology as the set of beliefs common to all its variants (here Π is a projection function).

(10) $\Pi_2(\iota) =_{df} \bigcap \Pi_2(\iota')$, where $\iota' \sim \iota$.

Second, we need to make one key assumption about the relation between persona and belief. What kind of personas are available for an individual? That is, in a linguistic context, what kinds of personas can a speaker assume or signal? We assume here that speaker personas are required to be sincerely assumed, ie that the basis of that persona correlates with the speaker's actual beliefs. This is an analogue of Gricean Quality for the domain of social meaning, which we will call *Social Sincerity*. Formally speaking, this amounts to requiring the personas compatible with the speaker's utterance, $\mathbf{emf}(u)$, to associate with bases which have some relationship to the speaker's beliefs.

(11) *Social Sincerity*
$\forall s, u, \pi[utter(s)(u) \wedge \pi \in \mathbf{emf}(u) \wedge \iota_\pi \rightarrow MOST(p \in \Pi_2(\iota_\pi))(Bel(s, p))]$
'If a speaker utters a sentence compatible with persona π, they believe a

significant number of the propositions comprising the basis for π.'

Two comments on this principle. It is relatively weak in the sense that it simply requires the speaker to hold most of the beliefs associated with the ideology. We might fix this by using a different, stronger, quantifier, even a universal quantifier; this strikes us, however, as too strict, for people can certainly sincerely project ideological personas without accepting every aspect of the ideology they project. Another option is to use a different underlying theory, for instance a contextually determined parameter for sincerity in the manner of Kennedy 2007 on vague predicates or McCready 2015 for reliability of information source. This seems promising to us, but we will not pursue it here. The principle as stated also treats all beliefs in $\Pi_2(\iota)$ identically, but likely some of these beliefs are more 'core' to the ideology than others. which could be modeled by weighting them as in the belief revision literature on entrenchment (Gärdenfors 1988). We leave further consideration of these issues for future work.

With these formal elements in place, we can return to enrichment.

4.2 Enrichment via ideology

We argue that enrichment is a multistep process, the first part of which is shared with identifying dogwhistles. First, the listener identifies the speaker's persona on the basis of their utterance; this is of course the same as with identifying dogwhistles. Next, the listener identifies any ideology associated with the persona and calls up its ideological basis. This step is dependent on the Social Sincerity assumption, for without it one cannot assume that the persona the speaker presents with is actually one whose basis they are willing to adopt (even for the purposes of the conversation). If the basis together with the utterance content allow inferences to be drawn, enrichment is the result.

Let us consider one case in detail: the classic *inner city*. On our view, this is analyzed as follows. Suppose, for the (quasi)racist persona and corresponding ideology ι communicated by this DW to a savvy listener, $\mathit{live\_inner\_city}(x) > \mathit{black}(x) \in \Pi_2(\iota)$. This extra premise licenses an inference from 'inner city people don't work' to 'Black people living in inner cities don't work', which is the enriched meaning. This statement of course can then be used to understand the speaker's political views, draw conclusions about their policy decisions, etc. We can already see how this view improves on Khoo and our previous work. For Khoo, the inference depends purely on semantic content. This means that there's no way to explain how semantically coextensive terms trigger/don't trigger these inferences; but for us, the mediation through persona, which is only enabled by dogwhistles for savvy listeners, makes coextensive phrases act differently in the inferences they trigger

via ideologies. It also improves on our previous work (Henderson & McCready 2019b). There, we claimed that enriching dogwhistles involve a process of pragmatic enrichment a la Recanati 2003 mediated by dogwhistles. But we had nothing to say about how this mediation took place, or what the relationship between dogwhistle and enrichment looked like. In this new analysis, a principled explanation is available: dogwhistles allow savvy listeners to recover personas, personas are associated with ideologies, and certain ideologies, together with the content of the speech act, trigger inferences. Thus, taking the relation between persona and ideology seriously gives the necessary ingredients for an explanation that satisfies our desiderata.

Clarifying the link between dogwhistle and enriched meaning is not the only reason to adopt a theory like ours. Another crucial pieces of evidence that we need to distinguish invited inference and dogwhistle recognition comes from Hurwitz & Peffley (2005). In this study, they showed that white people in studies on racial dogwhistles are more likely to unconsciously assign stereotypes to racial minorities on a post-test, but that you can correct this by telling them that the dogwhistle is racist before they hear it. However, Black people in these same studies showed no effect when hearing the dogwhistle in assigning racial stereotypes, yet they are aware that the word in questions is, in fact, a dogwhistle.

We draw from this study the lesson that we must separate “hearing the dogwhistle” from “making inferences based on the dogwhistle”. Linking personas with these background ideologies can help us do this. When a white person hears a racial dogwhistle, they learn that the speaker, in virtue of their persona, is willing to have a conversation from a "white chauvinist" standpoint. Without objection, the default effect might be that this ideology becomes the ground for the conversation, dragging along all the inferences associated with that ideology—hence, the effect we see on implicit bias post-tests. In contrast, when white participants are explicitly warned about the dogwhistle, resisting the associated ideology becomes salient. Listeners still "hear" the dogwhistle, but they are primed to not slide onto the ideological grounds the persona associated with that whistle invites. We see a similar effect with Black participants. They "hear" the dogwhistle, but by default resist grounding the conversation in a racist ideology by. This makes sense for obvious reasons.

5 Epistemic vigilance and hypervigilance

We now want to switch gears and consider a practical problem in the interpretation of dogwhistles. It is common to see people on Twitter and elsewhere reacting quickly and aggressively to the use of dogwhistles, or what they call dogwhistles, some of which fall into our category of dogwhistle and some of which don’t. But

are these quick and aggressive reactions the right ones to take? Are there better approaches, either from a human relations perspective, or, more relevantly for us here, from the perspective of interpretation? More generally, when observing a speaker using an expression we know to be a dogwhistle, what is the proper reaction? The complicating factor is that it is possible that the expression is being used innocently: speakers often hear dogwhistles without recognizing them, and may pick up expressions from political discourse without knowing their dogwhistley quality. Indeed, the fact that dogwhistles have (at least in principle) completely innocent and non-persona-signaling uses is precisely why they can be used as dogwhistles in the first place. Some uses of dogwhistles, then, are innocent. But how can these be distinguished from cases where dogwhistles are genuinely used to deceive, as covert signals of identity?

There might be no universal or proper procedure for deciding this question, any more than there are universal considerations that allow us to decide whether a person is being sincere, or cooperative, or any other kind of property that one might hope to make judgements about on the basis of speech. We instead want to offer a diagnosis of what happens when expressions are judged to be dogwhistling, and when such judgements might be made. A first observation is that for a dogwhistle to be useful, the speaker must judge the speech situation to be one in which it is to their advantage to covertly signal. In other words, the speaker must take themself to be in a context which is, at least potentially, one where revealing their true social – or ideological, given what we have said above – identity could have negative effects.

This means that for a dogwhistle to be rationally used, the speaker must take themself to be in a zero-sum situation with respect to social identity, at least with respect to a subset of their audience. A consequence of this observation is that, for a listener to judge a particular use of an expression as a dogwhistle, she must think that the speaker is intending to deceive; more specifically, she must take the speaker to believe the interests of some discourse participants not to be aligned. The result is that judging an expression a dogwhistle already imputes hostile intent to the speaker. If this judgement is wrong, it can both create arguments and disagreement where there might have been none: an innocent speaker accused of strategically hiding racist or other bad social attitudes will often react negatively and the situation may easily escalate. This kind of judgement can also dispose the listener/interpreter to systematically misconstrue other utterances of the speaker due to shifts in the probabilities she assigns to speaker personas and consequently to what she guesses the speaker's intentions to be. This kind of Bayesian-style triangulation of intended interpretation is found widely elsewhere: for instance, it is discussed extensively with respect to 'underspecified emotive adjectives' like *fucking* in mcc).

Still, there are reasons to try to catch people out in their use of dogwhistles. Clearly, in cases where dogwhistling strategies are genuinely being used to hide

repugnant ideologies, one should do one's best to ferret out the culprits. But being too quick to assign malign intent to speakers is also a danger. The upshot is that while it is important to be epistemically vigilant about dogwhistles, it is perhaps equally important not to be hypervigilant, for doing so can create communication breakdowns and eliminate the possibility of real and genuine communication.

We should therefore be careful about stepping off the deep end into hypervigilance; and in order to enable this kind of care, it is useful to try to determine the conditions under which hypervigilance arises. A first obvious reason for hypervigilance lies in experience, for instance in a buildup of bad interactions with users of a particular potential dogwhistle who ultimately consistently turn out to be genuine bad actors, something which is perhaps a common experience for active Twitter users. This kind of experientially based change boils down to shifts in probabilistic priors on the basis of interaction, which is already expressed in our model, with its heavy use of Bayesian reasoning. Another reason lies in values: in certain cases, it is genuinely beneficial to the speaker to search out dogwhistles, looked at from the perspective of utilities. We want to consider the latter case in a bit more detail.

Recall that in our system utilities are assigned via a combination of informative content and value assigned to particular social meanings. But we haven't had anything to say in this paper about how exactly these value assignments are determined. What is the right way to do this? Clearly, what values we let ρ assign to particular personas for a speaker will depend on the kinds of views the speaker has and how those interact with what is projected by the persona, particularly in terms of the ideologies they are associated with: tradition/radicalness, political views, social groupings, etc. This could be spelled out in many different ways. Here we want to explore one simple metric: similarity, ie. 'I like people who are like me.'

The idea is just that the speaker's own ideology assigns value to various kinds of actions, properties, memberships in social groups and so on, and the degree to which these value assignments are similar to those assigned by the projected ideology will determine a degree of affinity. If this is right, we can assign affective values on the basis of similarity metrics between speaker and hearer personas, something already proposed in Henderson & McCready (2019a) and used there to show how people might make judgements about trustworthiness on the basis of affinity rather than just on epistemic reliability (see also **?** for more on this idea). Of course, similarity is only one aspect of value assignment, but it seems to us to be a key one.

Our hypothesis is that people are sensitive to dogwhistles to precisely the degree they have strong feelings about the ideology. If their feelings are highly positive – or, otherwise put, if the two ideologies are highly similar, given that we use similarity metrics to model 'ideological support' – the listener can learn that the speaker is sympathetic; if they are highly dissimilar and thus highly negative, then the speaker is to be avoided, or even combatted. All this is very easily modeled in our theory:

once affective value is incorporated into utilities, learning someone holds an ideology with high positive or negative utility for the interpreter is useful to know, and directly recoverable from the outputs of ρ.

We can think of the results of this procedure as an index of the degree of approval the speaker and interpreter assign to each other's ideologies; this degree is already a part of the utilities yielded by the interpretation procedure we outlined in previous sections. Thus, to the extent that speaker persona is similar/dissimilar to listener persona, utility is high: this gives greater incentive to search out dogwhistles when they are present. The result can be hypervigilance. We think there is a mathematical result to be found here, somewhat akin to the credibility result for cheap talk games of Farrell (1993), where a signal is credible to the degree that the interests of the sender align with that of the receiver: here, perhaps, listeners interpret possible dogwhistles as dogwhistles to the degree that doing so reflects a utility change, ie to the extent that personas are (dis)similar along some dimension. The precise statement of this result we leave for future work.

6 Conclusions

To summarize, we have argued against several existing accounts of dogwhistles, including some of our own previous work. The puzzle is that dogwhistles must not convey conventionalized truth-conditional content, yet certain expressions seem to be specified, by convention, as dogwhistles, and the use of these expressions appears to lead to truth-conditional enrichments. We solve this problem with a novel account of dogwhistles based on Social Meaning Games (Burnett 2017, 2016) in which dogwhistles convey social personas, but can affect the at-issue content conveyed, which is influenced by the persona recovered. We then considered some implications of the account, in particular, for our understanding of unconscious bias and vigilance about dogwhistles.

References

???? .

Burnett, Heather. 2016. Signalling games, sociolinguistic variation and the construction of style. In *the 40th Penn Linguistics Colloquium, University of Pennsylvania*, .

Burnett, Heather. 2017. Sociolinguistic interaction and identity construction: The view from game-theoretic pragmatics. *Linguistics and Philosophy* .

Chierchia, Gennaro. 1998. Reference to kinds across language. *Natural Language Semantics* 6. 339–405. doi:10.1023/A:1008324218506.

Chierchia, Gennaro & Raymond Turner. 1988. Semantics and property theory. *Linguistics and Philosophy* 11. 261–302.

Farrell, Joseph. 1993. Meaning and credibility in cheap-talk games. *Games and Economic Behavior* 5(4). 514–31. doi:10.1006/game.1993.1029.

Gärdenfors, Peter. 1988. *Knowledge in Flux*. MIT Press.

Henderson, R. & E. McCready. 2019a. Dogwhistles, trust, and ideology. In *Proceedings of the 22nd Amsterdam Colloquium*, .

Henderson, Robert & Elin McCready. 2019b. Dogwhistles and the at-issue/non-at-issue distinction. In Daniel Gutzmann & Katharina Turgay (eds.), *Secondary content*, 222–245. Brill.

Hurwitz, Jon & Mark Peffley. 2005. Playing the race card in the post–willie horton era the impact of racialized code words on support for punitive crime policy. *Public Opinion Quarterly* 69(1). 99–112.

Kennedy, Chris. 2007. Vagueness and gradability: The semantics of relative and absolute gradable predicates. *Linguistics and Philosophy* 30(1). 1–45.

Khoo, Justin. 2017. Code words in political discourse. *Philosophical Topics* .

McCready, E. 2015. *Reliability in Pragmatics*. Oxford University Press.

Recanati, Francois. 2003. *Literal Meaning*. Cambridge University Press.

Stanley, Jason. 2015. *How propaganda works*. Princeton University Press.

Yoon, Elina J, Michael Henry Tessler, Noah D Goodman & Michael C Frank. 2016. Talking with tact: Polite language as a balance between kindness and informativity. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, Cognitive Science Society.

Robert Henderson
Communications 306
1103 E University Blvd
Tucson, AZ 85721
rhenderson@arizona.edu

Elin McCready
Department of English
Aoyama Gakuin University
4-4-25 Shibuya
Shibuya Tokyo 150-8366
JAPAN
mccready@cl.aoyama.ac.jp