DOGWHISTLES, TRUST AND IDEOLOGY

Elin McCready
Aoyama Gakuin University
mccready@cl.aoyama.ac.jp

March 4, 2020

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. When someone says (1), they hear (2).

(2) Yet there's power—Christian power—in the goodness and idealism and faith of the American people.

In a 2016 Reddit AMA Green Party presidential candidate Jill Stein was asked about the party's platform vaccines and homeopathy. She said:

(3) By the same token, being "tested" and "reviewed" by agencies tied to big pharma and the chemical industry is also problematic.

Even though Stein said she thought vaccines work, across the internet she was accused of being an -vaxxer due to phrases like *big pharma*, which to people familiar with alternative-medicine discourses know is demonized as selling poison for profit. They heard:

(4) By the same token, being "tested" and "reviewed" by agencies tied to big pharma and the chemical industry, who sell unsafe vaccines to make a buck, is also problematic.

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

(5) 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 by fellow Representative Barbara Lee for making a “thinly veiled racial attack”. This is because the phrase *inner-city* is code or euphemism for African American neighborhoods (espcially stereotypically racialized views of such neighborhoods). Many people heard Paul Ryan say:

(6) We have got this tailspin of culture, in our African American neighborhoods 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.

All three of 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.

- Dogwhistle language has been explored quite a bit in political science and political economy (e.g., Calfano and Djupe 2008; Goodin and Saward 2005; Hurwitz and Peffley 2005; Mendelberg 2001), and even in their experimental literatures.

The linguistic literature on dogwhistles is practically non-existent, but there are proposal coming from philosophy:

- Stanley 2015 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.

- Khoo 2017 provides a purely pragmatic account, where dogwhistles involve certain default inferences.

- Finally, there are proposals, though slightly less fleshed out (e.g., Saul 2018), which takes dogwhistles to be simple gricean implicatures.

We think none of these proposal is correct, though exploring them is important because they expose certain tensions.

- We will see through arguments against the CI account that dogwhistles cannot involve conventionalized TC-meaning (either at-issue or not-at-issue)

- But, arguments against the pragmatic accounts of Khoo 2017 and Saul 2018 will include the fact that dogwhistles require some kind of conventionalization.

After exploring these previous accounts, we propose our own combining aspects of McCready 2012, Burnett 2016; Burnett 2017 which we think better accounts for their core properties, while resolving this tension about conventionalization.

- We further show our account has additional applications for other kinds of social meaning

In broad strokes, we make the novel proposal that dogwhistles come in two types.

- The first type—*identifying dogwhistles*—concerns covert signals that the speaker has a certain persona, which we model by extending the *Sociolinguistic Signalling Games* of Burnett 2016; Burnett 2017.

- The second type—*enriching dogwhistles*—involves sending a message with an enriched meaning whose recovery is contingent on recognizing the speaker's covertly signalled persona, which requires a further extension.

The conventional implicature account

Stanley 2015 argues that dogwhistle language involves a conventional non-at-issue component along the lines of more familiar expressions like slurs, honorifics, etc.

- 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 “African Americans are lazy”.

- In general, terms which carry both AI and NAI components can be referred to as *mixed content bearers*.

We disagree with this characterization.

Knowledge argument.
The requirements for knowing the meaning of dogwhistles differ from ordinary mixed content bearers.

- Take the case of pejoratives. Can a speaker know what *kraut* means without knowing it is derogatory? No.

- Conversely, can a speaker know what *welfare* means without knowing this association with Cadillacs, etc. (Stanley p. 158-9)?
  - We think the answer is: Yes. The whole idea of a dogwhistle is that the (so-called) NAI component is not accessible to some speakers.
  - So the NAI part is not part of conventional meaning.

Objection!
Maybe we're just dealing with different dialects?

- This view might explain the effect of dogwhistles in mixed company, but does not explain the use of dogwhistles with an in-group.
  - Under a dialect account, dog-whistle language should also be what is used when talking to an in-group because this is just what the words mean for the audience.
  - But dogwhistles, by definition, are not needed when talking to an in-group, which wouldn't make sense if the subtext of dogwhistle were part of its conventional meaning for the in-group.

‘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.

(7) A: Angela Merkel is a kraut.
B: What do you have against Germans?
A: #I don’t have anything against Germans. Why do you think I might?

Such dialogues are fine with dogwhistles; in the following, there seems to be no entailment that A has the relevant attitude.

(8) A: Elin is on welfare.
B: What do you have against social programs?
A: I don't have anything against social programs. Why do you think I might?

Generalizing, we can identify a dialogue-based test for conventional content: in a dialogue in which participant A says 'X', where $[\![X]\!]$ is a mixed content bearer with AI content Y and NAI content Z and participant B responds with 'It's not cool to say Z', it is incoherent for A to respond 'I didn't say that Z" if Z is conventional content.

- By this test, dogwhistles can be concluded not to be conventional.

The inferentialist account

Khoo 2017 argues that dogwhistles involve default inferences:

- 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.

- If the interpreter believes that *inner-city* neighborhoods are African American 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 African American neighborhoods lack a culture of work.

- This is a kind of invited inference account which relies on the (at-issue) content of the dogwhistle itself and the background beliefs interpreters have which license a constellation of inferences about things related to that content.

The inferentialist account makes sense of the fact that dogwhisles are deniable, but it has problems.

Non-substitutability argument

Khoo's inference follows from the expression TCs. Thus, any expression with the same TCs should dogwhistle.

- This is not true. A phrase like *downtown neighborhoods* doesn't dogwhistle like *inner city* does. The same for *welfare* and paraphrases like *assistance to the poor*

This suggests that while dogwhisles must not bear conventionalized content (see the arguments against the CI account), some expressions are singled out as something like "dogwhistle expressions", and so there is some kind of conventionalization.

Aren't dogwhistles just (manner) implicatures?
Another kind of pragmatic account takes dogwhistles involve, not content-based inferences, but classical gricean inferences based on the form of what was said.

- This get to the idea that dogwhistles are deniable and involve something about the expression itself, but it's surprisingly hard to make work along multiple channels.

Consider Saul's attempt to understand the George W. Bush's "Dred Scott" dogwhistle as a relevance implicature:

"One can certainly tell a story of how they'd be calculated: He's stating his opposition to Dred Scott. But everyone opposes Dred Scott, and that's not relevant to the question he was being asked. He must be trying to convey something else—that he is opposed to abortion, like those other people who talk about Dred Scott." (Saul 2018, p. 7)

But wait, who is computing this relevance implicature? Remember we have two popluations, those who hear the dogwhistle and those who don't.

- It seems like the population that doesn't know about the dogwhistle should have a relevance implicature triggered, but that is precisely what we don't want from a codeword.

- In contrast, the people who know about "those other people [opposed to abortion] who talk about Dred Scott"—i.e., those who hear the dogwhistle—don't have to compute a relevance implicature. It *is* relevant. That's how they talk about the issue in -abortion discourse.

If we think of relevance implicatures as arising to keep a conversation coherent under pressure from the speaker's choice of language, then any relevance approach will run into this kind of problem because dogwhistles, by definition, exploit a split between naive and savvy populations of listners.

- Savvy listeners in general will generate less relevance implicatures because they know more about the grounds of the conversation than the naive listeners

- But this is backwards because we want the savvy listeners to be generating the enriched meanings for the dogwhistles.

If not a relevance implicature, it seems like the only other choice is manner, given that dogwhistles involve word choice. Once again this is surpisingly hard to make work.

- Consider again the "inner-city" dogwhistles in "We have got this tailspin of culture, in our inner cities in particular, of men not working. . . "

- We want to generate, for the savvy listeners, the enriched meaning: "We have got this tailspin of culture, in our African-American neighborhoods in particular, of men not working. . . "

But wait, what is the competitor c to "inner-city" that the listener could have said, whose refusal to say allows us to conclude the speaker meant "African-American neighborhoods"?

No good options! I think the enriched meaning is just not deriable by manner implicaturs

- That said, it seems like what is actually in pragmatic competetion “inner-city” and it’s enriched meaning itself, “African-American neighborhoods”.

- But what kind of manner implicatures are generated here? We can get implcatures about the dogwhistle effect itself, but they are not standard Gricean implicatures.

The ‘what is said’ argument reprise

A manner implicature should go something like this—The speaker said *inner-city*. Assuming they are co-operative and following manner, they would have said *African-American neighborhoods*. . . . . **PAUSE**. . . . . .

- Already we have a problem. Classic manner implicatures involve non-standard ways of saying the same thing. This assumes that *African-American neighborhoods* and *inner-city* mean the same thing (e.g., *stop the car* vs. *made the care stop*), but we have already seen that dogwhistles must not have this conventionalized content given that they are deniable.

Let’s grant this equivalence, though. The manner implicature is still non-standard.

The ‘what is implicated’ argument

The speaker said *inner-city*. Assuming they are cooperative and following manner (speaking plainly without codewords), they would have said *African-American neighborhoods*. They didn’t. It must be because they cannot say *African-American neighborhoods* without violating one of the other maxims. *But which one? And what is the resulting implicature supposed to be?*

- It seems like there is no good choice for what maxim this would be, but it seems like what the listener concludes of the speaker is something like: *It’s not **safe** for the speaker to say “African-American neighborhoods”*

- Thus, the listener infers that the speaker believes that “We have got this tailspin of culture, in our

African-American neighborhoods in particular, of men not working. . . ”

- Additionally, the listener concludes (based on **safety**) a bunch of meta-conversational things like: (i) Speaker is being semi-cooperative, (ii) the speaker believes the audience is a mix of savvy and naive listeners, (iii) the speaker belives the audience will have mixed reaction to their implicated meaning.

This seems fine, but it is clearly not a vanilla Gricean implicature. Even if we grant the troublesome assumption that will allow the manner implicature to go through, we have to make reference to this new principle, **safety**, and then we generate a bunch of meta-conversational implicatures about the conversational participants, including the speaker’s cooperativity.

Core properties to be accounted for:

- Dogwhistles are not part of conventional content, so speakers are able to avoid (complete) responsibility for what they convey.

- Dogwhistles can be identified as such, even if not bearing convetional content.

- Dogwhistles are semi-cooperative—that is, they are meant to be under-informative to one segment of the audience, while communicating a particular message to another.

- While deniable, dogwhistles are risky. Being detected using a dogwhistle by the wrong party should be costly.

Dogwhistles come, we think, in two types.

Type 1 (Identifying Dogwhistles): The content sends one message to all audience members, while the whistle transmits the speaker's true identity to a sub-audience.

- The Stein and Bush cases above probably best fit in this category.
  - Stein's "Big Pharma" just means large, faceless pharmaceutical corporations (parallel to "Big Agriculture", etc.), but she flagged herself a vaccine denier because that phrase is primarily used in vaccine-denial (and alternative medicine) discourse.
  - Bush's "wonder-working power" probably doesn't convey some secondary message about the power at hand, but instead just flags him as an evangelical because only they talk like that.

Type 2 (Enriching Dogwhistles): The content sends one message to all audience members, while the whistle sends places an addendum on that message for a sub-audience.

- The Ryan case above best fits this category. His use of “inner city” conveys to all audiences a geographical location inside cities, but then to a sub-audience, it specifically picks out African American neighborhoods in those cities.

- Of course, Ryan’s utterance will also allow a listener to infer things about Ryan’s identity as in Identifying examples—this is especially true if the whistle is detected.

We take each of these cases in turn, starting from the simpler Identifying dogwhistles and then expanding into the Enrching dogwhistles.

In recent work, Burnett 2016; Burnett 2017 pioneers the use of Bayesian signaling games to model identity construction through sociolinguistic variation.

- We take identifying dogwhistles to be only slightly more complex verions of sociolinguistic identity construction through variation of the kind Burnett (2016) and Burnett (2017) discuss.

- Enriching dogwhistles will be an extension of these games where amended messages are sent to a sub-audience that work in concert with the kind of identity construction we see in indentifying dog-whistles.

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.

- The speaker's utility is maximized by picking a message that sends the most information to the listener about the persona they want them to assign to them.

- The listener's utility is maximized if they extract the most information they can about a speaker's persona given their message.

We now elaborate on these ingredients and model the behavior of Identifying dogwhistles.

- The set of personas P is a set of maximally consistent sets of properties.
  - For instance, in the Stein case (e.g., "agencies tied to **big pharma** and the chemical industry is also problematic"), the relevant properties might be: -VAX, +VAX, -CORPORATE, +CORPORATE
  - Maximally consistent subsets of these properties would be:
    $\{$-VAX, -CORPORATE$\}$,
    $\{$-VAX, +CORPORATE$\}$,
    $\{$+VAX, -CORPORATE$\}$,
    $\{$+VAX, +CORPORATE$\}$

- Messages $m \in M$ may have their normal denotational meaning $[\![m]\!]$, but for the sake of Identifying dogwhistles, messages also have a social meaning, which they take from P, written $[m] \in P$.
  - While a message m is associated with a particular persona, we often work with a related object $c(m) = \{n \in M | m \cap n \neq \emptyset\}$
  - We can think of $c(m)$ as denoting all of the personas that are consistent with m
  - Thus, assuming
    [*Big Pharma*] $= \{$-VAX, -CORPORATE$\}$,
    we also have
    c(*Big Pharma*) $= \{\{$-VAX, -CORPORATE$\}$,
    $\{$-VAX, +CORPORATE$\}$, $\{$+VAX, -CORPORATE$\}\}$
  - That is, using *Big Pharma* is consistent with any persona that is not
    $\{$+VAX, +CORPORATE$\}$

With this in mind, games now have the following elaborated action structure.

- The speaker picks a persona and a message—e.g.,
  $\langle \{$-VAX, -CORPORATE$\}$, *Big Pharma*$\rangle$

- The listener then identifies the speaker's persona based on their message from P:
  $\{$-VAX, -CORPORATE$\}$,
  $\{$-VAX, +CORPORATE$\}$,
  $\{$+VAX, -CORPORATE$\}$,
  $\{$+VAX, +CORPORATE$\}$—while knowing that the social meaning of *Big Pharma* rules out the persona
  $\{$+VAX, +CORPORATE$\}$

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.

- We want listeners to have beliefs about a speaker's persona. . .

- . . . but also beliefs about how personas and messages are connected.

That is, 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.

(9) $$P(p|m) \propto P(p)P(m|p)$$

'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'

- Note that we are working in a Bayesian RSA framework (Goodman and Frank 2016; Franke and Jger 2016; Franke and Degen 2016, among others), where (9) would be the 'literal listener'.

- This is a extension of Burnett 2016; Burnett 2017, who takes social meanings to be fully lexicalized, i.e., the likelihood P(m|p) = 1 when p and m are consistent.

The final ingredient we need to provide utility functions. For the listener it is straightforward—utility 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.

For speakers, Utility is more complex becuase unlike in many signalling 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 lister 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.

- 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; Yoon et al. 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:

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

When only one listener is addressed, dogwhistles reduce to ordinary social meaning; the speaker should choose a signal which maximizes U_S^{Soc}.

- Dogwhistles come into their own when speakers address groups of individuals with mixed preference 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.

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

With this utility function, the basic prediction is:

- Speakers will use language that maximizes their social utility wrt a group of listeners.

- For the dogwhistle case, this happens when using the dogwhistle allows gain of higher social utility than otherwise wrt the entire group,

- ie., 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.

Detailed formal example redacted for time reasons, but you can construct it easily—consider an audience with uniform priors over speaker personas, and an audience split in the likeliehood $P(m|p)$ that is equally split on a polarizing persona p.

To analyze Enriching dogwhistles we import the machinery of standard signaling games. Strategy:

- Use signaling games, assuming signals with two possible meanings, one an enriched version of the other

- Let recovery of the enriched version be tied to recognition of the relevant persona.

We need more components to model TC meaning.

- messages now denote pairs of truth-conditional meanings and social meanings: $\langle [[m]], [m] \rangle$.

- $T =$ a set of states t (worlds). Speaker strategies S_σ are now functions from pairs of states and personas to messages, and listener strategies L_ρ are functions from messages to such pairs.

A utility function for information retrieval

(12) $US(m, L) = US_{Soc}(m, L) + EU(m, L)$, where

$EU(m, L) = \sum_{t \in T} Pr(t) \times U(t, m, L)$, where

$U(t, m, L) > 0$ if $t \in L_\rho(m)$ and else $= 0$ (cf. van Rooij 2008).

i.e., the social meaning is always recovered, but if the listener fails to recover the proper truth-conditional meaning, no value is extracted from this aspect of the communication.

A more elaborated version of this function can be given by weighting the two components of the utilities with values δ, γ, giving $U_S(m, L) = \delta U_S^{Soc}(m, L) + \gamma EU(m, L)$

- δ indexes the value placed on the social meaning and γ the value of the truth-conditional meaning.

- Setting $\delta = 0$ gives stereotypical robot communication, where social meaning is disregarded.

- At the other extreme, setting $\gamma = 0$ gives 'post-truth'.

Put a pin in this, we'll return to this point.

The above seems correct for Identifying dogwhistles, where the two meanings are semi-independent. But more needs to be said for Enriching dogwhistle meanings.

- The reason is that, in these cases, proper recovery of intended (enriched) TC meaning is dependent on identifying the relevant persona.

- We are inclined to view this as a kind of pragmatic encroachment somewhat parallel to the cases discussed by e.g. Recanati 2003.

(13) mom to child on the playground
You're not going to die (*from that cut*).

- However, standard cases are entirely contextually conditioned, while these seem to be the result of a conventional association: once the persona is identified, the additional meaning becomes apparent to the interpreter.

This means that Identifying dogwhistles (etc) are actually a special case.

- In fact, they are likely an extreme instance of a general phenomenon.

There seem to be two steps in this kind of interpretation.

- The listener first recovers the speaker's persona on the basis of the utterance, and then uses the result to determine 'what is said'.

- In the present setting, this amounts to conditionalizing prior probabilities on the social meaning and using the posterior probabilities to recover the TC meaning.

This can be modeled by altering the expected utility computation for the TC part of (12) to reference posterior probabilities, as represented by Pr' in (14):

(14) $US(m, L) = US_{Soc}(m, L) + EU(m, L)$, where

$EU(m, L) = \sum_{t \in T} \mathbf{Pr'}(\mathbf{t}|\mathbf{p}) \times U(t, m, L)$, where

$U(t, m, L) > 0$ if $t \in L_\rho(m)$ and else $= 0$ (cf. van Rooij 2008).

Example.

Consider the utterance (5), with its Enriching dogwhistle.

- This utterance contains the phrase ‘inner cities’ which, on its dogwhistled interpretation, means ‘African American neighborhoods’.

- Without recognizing Paul Ryan’s persona, this interpretation seems to be very difficult to get; but, once the persona is recognized, it is very easy, given knowledge of the relevant signal.

A quick summary and then one extension. This paper has:

- Argued against a CI account of dogwhistles on which they introduce mixed content

- Distinguished two types of dogwhistle, both of which convey social personas but only one of which has at-issue content which is influenced by the persona recovered

- Modeled the two types using an extension and variant of Burnett's social meaning games

What is the best characterization of dogwhistles within existing domains of not-at-issue meaning?

- As we have argued, CIs are an improper characterization, for the meaning is not fully conventional.

- Rather, on our analysis, all the action in Identifying dogwhistles is in the domain of social meaning, while Enriching dogwhistles further build on the result of these inferences to alter or enrich at-issue content.

- They share with conversational implicatures the property of being cancellable (deniable), but differ from (standard views of) them in not following from (anything but an extremely nonstandard construal of) the Gricean Maxims.

- Dogwhistles seem to occupy a genuinely new niche in the characterization of not-at-issue meaning.

Topic shift: dogwhistles, reliability, trust.

Q1: Reliability Why do we trust what other people say, and form beliefs on the basis of their speech?

- One answer: they are taken to be *reliable*.

- Intuitively this means that the other person (or institution, or group) is taken to be reliable in what they say, at least with respect to a particular domain.

Question: What is reliability, and how can one be judged reliable?

Q2: Trust

What is the relationship between reliability and trust?

- Do we trust reliable and only reliable people?
- If not, which takes precedence: reliability or something else?
- Claim 1: we might still trust the unreliable in some circumstances.
- Claim 2: sometimes trust – ideological trust – trumps reliability; surprisingly, such trust might be rational.

One way to be authoritative, in the sense of having one's speech consistently believed, is to be a speaker who is judged reliable with respect to speaking truth.

- Reputation is key given that belief is a form of co-operation.

- cf. game-theoretic case: use of reputation in strategizing in repeated Prisoner's Dilemma (Nowak and Sigmund 1998).

How to model reputation with respect to reliability?

Histories

McCready 2015: Reputations derived in part from *histories*: sequences of objects $act \in A, A$ the set of possible actions for a given agent in a given (repeated) game.

- These objects are records of an agent's actions in past repetitions of the game.

- Game histories are n-tuples of sequences of records representing the history of the agent's actions at each decision point.

McCready 2015 uses these to model the action of hedges: remove objects from the 'permanent record.'

Reputations

A player's reputation in a game is derived from his history in that game.

- A player's reputation with respect to some choice as his propensity, based on past performance, to make a particular move at that point in the game.

- Such propensities are computed from frequencies of this or that move in the history.

- Specifically, the propensity of player a to play a move m in a game g at move i is:

- the proportion of the total number of game repetitions that the player chose the action m at choice point i.

Property-based heuristic

The above must be combined with other information about reliability.

- Fricker 2007: speakers make judgements about people's epistemic authority based on stereotypical information.

- e.g. their looks, gender, occupation, grooming, context . . .

This heuristic gives a first guess about reliability which is then modified by interaction.

- This can all be embedded in a more general model of information change.

Interim summary

Judgements of the reliability of an agent rely on . . .

- the history of the agent's communicative behavior:
  - is it truth-tracking?
  - is it honest?
- the agent's properties:
  - do they inspire confidence?
  - how do they interact with social stereotypes?
  - sometimes: unjust and incorrect results (*epistemic injustice*: Fricker 2007).
- the agent's linguistic behavior

But, back to the main question:

- Is their reliability wrt truth-tracking the only factor in determining whether to believe someone?

No.

Sometimes other considerations completely overshadow reliability.

- Trump: well-known to lie frequently, or at least not to care much about truth

- (DT being only the most obvious example here!)

- More generally: knowing someone is unreliable doesn't always lead people not to trust them.

But why would anyone ever trust an unreliable person?

Trust vs reliability

Idea: ground an analysis in the difference between reliability and trust.

- Reliable person: consistently truth-tracking
- Trustworthy person: acts in a way that furthers one's interests, broadly conceived
- Both lead to value in a game- or utility-theoretic sense
- $\Longrightarrow$ Reliability can inspire trust, but trust need not be grounded in reliability.

What, though, grounds trust, if not reliability?

Idea: social meaning/ideology expression, as used for dogwhistles.

Positive value

The previous assumed that we have a way to assign affective values to personas. On what basis?

- Many possibilities, but many are ideological in a broad sense (tradition/radicalness, political views, social groupings).

- One metric is similarity: ‘I like people who are like me.’

- Then we can assign affective values on the basis of similarity metrics between speaker and hearer personas.

Of course, this is only one aspect of value assignment; but it leads to trust.

Valuing communicative agents

The analysis of McCready 2015 was predicated on truth-tracking.

- The assumption adopted from van Rooij 2008 above means that payoffs are positive only if true information is recovered in cases of only truth-conditional communication.

- In repeated game settings, this means that truth-tracking/reliability is the *only* relevant consideration in deciding whether to believe/cooperate/continue to interact with an agent.

- But now we have social meaning to consider.

Idea: the (a) role of social meaning for hearers is to decide whether to trust.

Summary and implications

Main conclusion: Trust is distinct from reliability.

- We can trust the unreliable, and not trust the reliable;

- this is even a rational way to behave, if we value social matching over truth – which we might, if we care most about what kind of policies a politician might want to implement.

A lesson: if we want to re-rationalize politics, pointing out the falsehoods of politicians is not a productive method in general.

- Suggested positive strategy: show that the ideological presentation of those politicians is insincere;

- or that the ideology itself is flawed.

Next steps

Immediate next steps for this research program, now underway:

- Examination of precise conditions under which valuing ideology over truth is rational (including simulations);

- Broadening the notion of persona similarity to give a formal account of the notion of ‘standpoint’ as used in standpoint epistemology;

- Other metrics than similarity in persona evaluation (weighted by genre?).

- Integrating the social meaning story fully with history-based theory of reliability.

THANK YOU!!!!!!!!!!!!

*References

Burnett, Heather (2016). “Signalling Games, Sociolinguistic Variation and the Construction of Style”. In: *the 40th Penn Linguistics Colloquium, University of Pennsylvania*.

— (2017). “Sociolinguistic Interaction and Identity Construction: The View from Game-Theoretic Pragmatics”. In: *Linguistics and Philosophy*.

Calfano, Brian Robert and Paul A Djupe (2008). “God talk: Religious cues and electoral support”. In: *Political Research Quarterly*.

Goodin, Robert E and Michael Saward (2005). “Dog whistles and democratic mandates”. In: *The Political Quarterly* 76.4, pp. 471–476.

Hurwitz, Jon and 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”. In: *Public Opinion Quarterly* 69.1, pp. 99–112.

Khoo, Justin (2017). “Code words in political discourse”. In: *Philosophical Topics*.

McCready, Elin (2012). “Emotive equilibria”. In: *Linguistics and Philosophy* 35.3, pp. 243–283.

— (2015). *Reliability in Pragmatics*. Oxford University Press.

Mendelberg, Tali (2001). *The race card: Campaign strategy, implicit messages, and the norm of equality*. Princeton University Press.

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

Saul, Jennifer (2018). “Dogwhistles, Political Manipulation, and Philosophy of Language”. In: *New Works on Speech Acts*, pp. 360–383.

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

van Rooij, Robert (2008). “Game Theory and Quantity Implicatures”. In: *Journal of Economic Methodology* (15), pp. 261–274.

Yoon, Elina J et al. (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.