Croatian Journal of Philosophy
Vol. XXV, No. 75, 2025
https://doi.org/10.52685/cjp.25.75.9
Received: July 7, 2025
Accepted: September 15, 2025

Linguistic Plausible Deniability: The Catalyst for Political Manipulation

MIRELA FUS-HOLMEDAL*
Norwegian University of Science and Technology, Trondheim, Norway

Risky politically manipulative speech has unexpectedly been on the rise. This paper investigates the role that the phenomenon of linguistic plausible deniability plays in the increasing prevalence of politically manipulative speech through dogwhistles, racial figleaves, and generic stereotypes. The paper unfolds in three main stages. First, it suggests that these linguistic devices share the phenomenon of plausible deniability, which, by offering cover for their overtness, mitigates (some) risks of such political speech. Second, it argues that the plausible deniability of these linguistic devices makes them powerful tools for politically manipulative speech as it helps it to spread more efficiently and appear more acceptable. Finally, it elevates the ethical and political dimensions of language to a more central position within the philosophy of language by discussing two normative claims stemming from conceptual engineering: (i) we should combat such pernicious political manipulation, and (ii) we should exploit the effects of plausible deniability for beneficial purposes.

Keywords: Linguistic plausible deniability; political manipulation; dogwhistles; racial figleaves; generic stereotypes; conceptual engineering.

* I would like to acknowledge the anonymous reviewers for their valuable input and suggestions, which significantly contributed to enhancing the quality of this paper. For useful feedback on earlier versions of the paper, I would like to thank the members of *Generics Online Group*; as well as the audiences of the *Hate Speech, Fake News and Freedom of Speech Conference*, 2023 July (University of Rijeka, online); *The 11th European Congress of Analytic Philosophy,* ESAP, 2023 August (University of Vienna, Austria); the *Philosophy of Language and Linguistics Conference,* 2023 September (Inter-University Center, Dubrovnik, Croatia); the *Vitenskapsteoretisk forum*, 2023 October (NTNU, Trondheim, Norway); the *Centre for Language Research*, 2024 June (University of Rijeka, Rijeka, Croatia); the *Perceiving Voice and Speaker Project Seminar*, 2024 December (University of Inland Norway, Lillehammer, Norway).

1. Introduction

Linguistic political manipulation can be broadly defined as the use of linguistic devices by politicians, often devised by their strategists, with the aim of gaining political advances. The propagation of -isms and -phobias, such as racism, sexism, Latinophobia, and Islamophobia, through politically manipulative speech is often seen as a calculated means to an end, exploiting racial resentment or implicit biases to achieve political advances. However, engaging in such tactics overtly carries certain risks for politicians and their strategists, as the audience may reject or condemn their speech, ultimately leading to a decrease in political advances.

Yet risky politically manipulative speech has unexpectedly been on the rise. Take, for instance, Trump's 2016 presidential campaign, which showcased the dissemination of both covert and overt racism through linguistic devices like dogwhistles, racial figleaves, and generic stereotypes. However, as Saul (2017a: 97–98) rightly points out, overt racism in political campaigns was previously 'widely thought to be socially unacceptable and death to a nationwide political campaign'. What has changed? What enables the relatively successful introduction of elements of such overtness in recent political discourse? Is there solely a shift in what is socially permissible or acceptable?

This paper takes as a standpoint that recent shifts in political communication norms and media environments may have increased the strategic value of plausible deniability, making it more instrumentally valuable now than before. It then investigates the role that the phenomenon of linguistic plausible deniability[1] plays in the increasing prevalence of politically manipulative speech through dogwhistles, racial figleaves, and generic stereotypes. The paper unfolds in three main stages. First, it suggests that these linguistic devices share plausible deniability, which, by offering cover for their overtness, can mitigate[2] (some) risks of such political speech[3]. Second, it argues that the plausible deniability of these linguistic devices has additional consequences that can be readily utilized for political manipulation. Finally, it discusses some normative considerations connected to the plausible deniability of these linguistic devices and the political manipulation that arises from these consequences.

[1] *Plausible deniability* hereafter refers specifically to linguistic plausible deniability, unless otherwise stated. See section 2 for more on this phenomenon.

[2] While plausible deniability may be significant, it is not the only factor in mitigating risks of these linguistic devices. Success may also vary, as effectiveness depends on factors such as biases, racial resentment, and the social and cultural background of the audience.

[3] Plausible deniability can potentially offer cover for both covert and overt forms of political speech. However, its effectiveness in relation to overt versions further strengthens the main argument of this paper.

The aim of this paper is to bring together and further elucidate the phenomenon of plausible deniability in the context of politically manipulative speech through linguistic devices such as dogwhistles, racial figleaves, and generic stereotypes[4]. However, this paper does not seek to argue for any particular view of plausible deniability nor these linguistic devices, as their existence is widely acknowledged in the literature. Instead, the paper assumes their presence and provides examples commonly recognized in the field. Its objective is to establish a general connection between these three linguistic devices, plausible deniability, and their role in politically manipulative speech. Moreover, it is worth noting that while the logical structure of generic stereotypes is considered to be relatively complex yet stable, figleaves and dogwhistles can be seen as possessing a more flexible logical structure. This flexibility contributes to the fact that plausible deniability associated with the latter two often requires extensive context and historical understanding to be considered credible. Although the topic of the logical structure and context of these linguistic devices warrants closer examination, it falls beyond the scope of this paper and will not be addressed in detail.

The paper proceeds as follows. Section 2 introduces the phenomenon of plausible deniability and provides motivation for its utilization in the context of politically manipulative speech. Sections 3-5 suggest that linguistic devices such as dogwhistles, racial figleaves, and generic stereotypes allow for plausible deniability, and that plausible deniability helps mitigate some potential risks associated with the messages conveyed through these linguistic devices. Section 6 argues that the plausible deniability of these linguistic devices makes them powerful tools for politically manipulative speech as it helps it to spread more efficiently and appear more acceptable. Furthermore, the section discusses two normative claims stemming from conceptual engineering: (i) we should combat such pernicious political manipulation, and (ii) we should exploit the effects of plausible deniability for beneficial purposes.

2. Plausible deniability

Plausible deniability, in broader terms, refers to the ability of individuals, especially those in positions of authority or power, to deny any knowledge, responsibility, or involvement in actions carried out by others within their organization or hierarchy. It is used in various domains including law, military, government, international relations, politics, espionage, intelligence operations, corporate governance, information security, programming, privacy, computer networks, cryptography, re-

[4] It is worth noting that other linguistic and non-linguistic phenomena, such as physical and linguistic micro-aggressions, accents, physical dogwhistles, propaganda, euphemisms, and speaker voice impressions, could also lead to politically manipulative acts that can be plausibly denied.

ligion, as well as informal, everyday or private speech and actions. For instance, a politician might strategically maintain plausible deniability by avoiding any direct involvement or knowledge of the controversial decision or action, despite being in a position to know.

A common thread in both linguistic and non-linguistic forms of plausible deniability is that denial often hinges on a lack of evidence directly linking those in question to the actions in question (linguistic and/or non-linguistic), even if they were personally involved or intentionally chose to remain ignorant about them. They rely on the assumption that their skeptics will be unable to prove otherwise due to the absence of compelling evidence, making their denial seem credible or believable. In some cases, when such (linguistic and/or non-linguistic) actions in question may involve wrongdoing or illegal activities, the absence of concrete evidence makes it difficult or even impossible to take any legal or punitive action against them based solely on accusations.

Within the philosophy of language, the phenomenon of plausible deniability can be broadly defined as a speaker asserting a certain proposition that she could, if challenged, later plausibly deny and claim that the proposition she asserted is, in fact, some other proposition (for some formulations, see Walton 1996; Pinker 2007; Lee and Pinker 2010; Fricker 2012; Stanley 2015; Peet 2015, 2024; Khoo 2017; Camp 2018; Mazzarella et al. 2018; Mazzarella 2021; Dinges and Zakkou 2023; Lemeire ms.).

To illustrate, let's consider a widely discussed example of plausible deniability involving a driver trying to bribe a police officer after being caught for running a red light by using language that allows for plausible deniability (see Pinker 2007: 437). We can contrast two cases to demonstrate how this bribery could occur[5]. In the first case, the driver could ask the police officer if they 'can take care of the ticket on the spot'. In the second case, the driver could explicitly say to the officer: 'I will give you money if you let me go without a ticket'. While the latter case would be considered explicit bribery, the former one would arguably be seen as a (more) implicit form of bribery. In contrast to the second case, what allows the first case to be considered a successful example of implicit bribery is its potential for plausible deniability. Specifically, if challenged by an incorruptible police officer, the driver in question could plausibly deny engaging in bribery. For instance, if challenged, the driver could deny that by uttering 'Can we take care of the ticket on the spot?' she intended to assert proposition p, where p represents something like 'Let me give you the money instead of paying the fine'. As a plausible defense, she could offer an alternative interpretation of her original statement, suggesting that she meant to assert proposition q, where q represents something like 'I would prefer the option of paying the fine on the spot using my credit card and a mobile terminal' (see Mazzarella 2021; Lemeire ms.).

[5] For the contrast between the two cases, see Mazzarella (2021).

To contrast, certain utterances lack plausible deniability due to e.g. their logical structure or the context in which they are uttered, leaving no room for plausible deniability interpretations[6]. Let's take the universally quantified statement 'All Blacks are violent' as an example of implausible deniability. If challenged, one cannot reasonably deny it by claiming to mean that some Blacks are not violent. When considering plausible deniability in terms of context, there is no context in which the statement 'All X are Y' allows for an interpretation that includes 'Some Xs are not Y'. Furthermore, it is also possible to speak of varying degrees of plausible deniability. Some utterances may be deniable, but they are not particularly plausible, even though they are more deniable than statements that are completely undeniable. For instance, in the case of a police officer, one could argue that the statement 'Can we take care of the ticket on the spot?' is perhaps not the most plausible but is still more deniable than the statement 'Let me give you the money instead of paying the fine', which is entirely undeniable (unless, perhaps, one intends it as a joke). There will be many grey areas in between. For the purposes of this paper, it is sufficient to acknowledge that some utterances are plausibly deniable while others are implausibly deniable.

It has also been suggested that an important prerequisite for the speaker to plausibly deny the content of her utterance is that the audience relies on the context of the utterance to be able to recover the content the speaker wants them to recover. Furthermore, if challenged, the speaker can redirect the focus to another plausible context in the vicinity and claim that the audience recovered the content *p* instead of *q* because the audience relied on the context which was not the context of the utterance she made (see Camp 2018; Mazzarella 2021; Lemeire ms.). For instance, in the case of the driver trying to bribe the police officer, the driver may emphasize the context of paying off the officer instead of considering the alternative context where the driver intends to pay using a mobile terminal. The salience of credit card and mobile payment in this alternative context makes the alternative content *q* plausible enough to deny the content *p* (see Mazzarella 2021: 10). Apart from relying on the context in the vicinity, plausible deniability could also be related to the meaning of concepts. For example, in the case of the phrase 'can we settle it on the spot', the speaker may argue that her intended meaning of "settle" is different from the conventional interpretation. However, such alternative interpretations are typically less convincing, since (private) meanings are often disregarded or difficult to change quickly within the constraints of pragmatic norms and conventions. Additionally, the presuppositions, common ground, and other pragmatic parameters at play could also influence the plausibility of deniability.

[6] For the phenomenon of so-called "implausible deniability", see e.g., Lee and Pinker (2010: 793); Camp (2018: 48); Berstler (2019: 27-28); Dinges and Zakkou (2023).

It is, however, important to reiterate that this paper does not aim to develop a specific account of plausible deniability. Thus, instead of offering a well-designed definition of plausible deniability, this paper adopts a heuristic approach to maintain neutrality regarding the specific accounts of plausible deniability associated with the linguistic devices discussed below. By utilizing the notion of a "communicative message", this paper avoids delving into discussions about the precise nature of what is being plausibly denied (e.g., propositions, assertions, semantic or pragmatic content, concepts, context, etc.).

> *Plausible deniability heuristic:* A speaker communicates a message by using a specific linguistic device that enables her to plausibly deny this message when challenged, claiming that the message she communicated is, in fact, a different one.

This heuristic has some inherent limitations, as it assumes that the plausibility of denial is due to the specific linguistic device used. It is, nevertheless, hoped that the heuristic will serve as a guiding principle in supporting the hypothesis that the identified commonalities among the linguistic devices used in political manipulation examined in this paper, namely dogwhistles, racial figleaves, and generic stereotypes (see sections 3-5), are influenced by the phenomenon of plausible deniability.

When it comes to the main motivation behind utilizing plausible deniability, literature emphasizes the intent to avoid the risks associated with openly communicating a certain message, such as asserting content *p*. For instance, in the example of the driver, her motivation for using a linguistic device that involves plausible deniability for her implicit bribery is to mitigate the risk of being punished for bribing a police officer. Furthermore, it is worth noting that the speaker's motivation to utilize plausible deniability of a linguistic device can also arise after the communicative message has already been delivered and the speaker becomes aware of its potential riskiness. The incentive to employ plausible deniability exists regardless of whether the speaker intentionally used the linguistic device with that purpose in mind before being challenged or if she only realized its potential for plausible deniability after being challenged.

Moving on to the domain of political manipulation, one can imagine similar motivations for why certain linguistic devices that allow for plausible deniability would be preferable in political speech, especially if they can mitigate risks that could lead to the politicians losing current or scaring away their potential supporters. Politicians might sometimes use linguistic devices with plausible deniability for the purposes of political manipulation, without being aware that what they are spreading is deeply racist. Instead, they might only care to spread those messages because they believe it will get them certain political advances or help them win the election, without being interested in spreading racism. In other words, the full consequences of their politi-

cal manipulation might go beyond their initial intentions. They might not care about the message being conveyed; their main objective could be to politically manipulate their supporters into accepting something that aligns with their implicit biases, and to evade accountability for if challenged, all for the sake of gaining more political advantage.

Moreover, plausible deniability can extend from speakers to their audience. It can protect not only politicians but also their supporters, an effect that those engaging in political manipulation (or their strategists) might be especially motivated to exploit (often without their supporters being aware of it). For example, being able to plausibly deny an overtly risky racist message could not only mitigate the risks of being considered a racist for the politician who communicated such a message but also for their supporters. This feature of being able to plausibly deny communicating or supporting a racist message comes in handy because often politicians and especially their supporters do not want to be perceived as explicitly racists (though they might be implicitly biased to be such) nor would want to identify themselves as such.

To sum up, certain linguistic devices with plausible deniability allow individuals to mitigate the risks of their communicative message by enabling them to plausibly deny the communication of a risky message when challenged. Using such linguistic devices is particularly advantageous in the realm of political manipulation. In sections 3-5 it will be suggested that linguistic devices such as dogwhistles, racial figleaves, and generic stereotypes used in political manipulation involve plausible deniability, which contributes to their selection as means of political manipulation.

3. Dogwhistles[7]

Some existing accounts of dogwhistles explain them in terms of implicit presuppositions (see Langton 2012), conversational exercitives (see McGowan 2004, 2012), perlocutionary speech acts and effects (see Saul 2018), and not-at-issue content (see Stanley 2015). This paper draws mainly on Saul's (2018) account of dogwhistles. The focus of this paper is not bound to a particular account of dogwhistles, however. This section is about plausible deniability as a more coarse-grained property of dogwhistles. In that sense, the plausible deniability of dogwhistles is seen as a higher-order phenomenon, not tied to either semantics or pragmatics *per se*[8].

According to Saul (2018), dogwhistling as a form of political[9] manipulation can occur through intentional or unintentional dogwhistles,

[7] It is, however, worth noting that Saul's discussion of dogwhistles and the aim of her paper is more sophisticated than discussed here.

[8] Analogous remarks apply to the case of racial figleaves, where Saul's work on racial figleaves is utilized for the same general purposes (see section 4).

[9] For examples of dogwhistles outside the political domain, see Saul (2018) and Witten (ms.).

both in overt and covert ways. While Saul (2018) uses words such as "deniability" and "challenging" that support the phenomenon of plausible deniability, her paper does not explicitly discuss this phenomenon. This section utilizes her account to show how one can tie these different types of dogwhistles to plausible deniability and how plausible deniability plays a role in choosing political manipulation through dogwhistles.

> *Case #1*[10]*. Dogwhistle:* Our inner cities are a disaster. You get shot walking to the store. They have no education, they have no jobs. I will do more for African Americans and Latinos[11] than she [Hillary Clinton] can ever do in ten lifetimes. All she has done is talk to the African Americans and to the Latinos.

Political manipulation through dogwhistle utterances allows politicians (or their strategists) to send one message to the general electorate and another coded message to the target electorate that the general electorate could challenge (see Goodin and Saward 2005; Lopez 2014; Stanley 2015; Saul 2018; Witten ms.). In Case #1, an utterance expressing a dogwhistle "inner cities" carries a coded meaning that associates black and brown communities with negative attributes such as "disastrous", "dangerous", "uneducated", "jobless". If the politician is challenged or accused of racism, one possible defense for the politician (and those who support his political speech) can be to claim that no inherent connection between black and brown communities and these negative attributes has been implicated. Instead, they can argue that their intention was simply to highlight the difficult circumstance faced by these groups and the need for assistance, something they claim their opposition cannot provide.

3.1 Intentional dogwhistles

Saul's (2018) definition of overt intentional dogwhistles, adopted from Kimberly Witten (ms.), is as follows: 'A[n overt intentional] dogwhistle is a speech act designed, with intent, to allow two plausible interpretations, with one interpretation being a private, coded message targeted for a subset of the general audience, and concealed in such a way that this general audience is unaware of the existence of the second, coded interpretation' (Witten ms.: 2). As an example of an overt intentional dogwhistle, consider George W. Bush's use of the phrase 'wonder-working power' to signal to fundamentalist Christians. While the general audience, specifically non-fundamentalists, may perceive this phrase as ordinary political language, it carries a coded message for its target audience, the fundamentalists, referring to 'the power of Christ'. Saul

[10] https://www.vox.com/policy-and-politics/2016/10/19/13336894/third-presidential-debate-live-transcript-clinton-trump

[11] One could remark that Trump is making this dogwhistle more explicit by naming the social groups the dogwhistle inner cities stand for, i.e. he uncoded the coded.

(2018) argues that this dogwhistle conveys two coded messages to its target audience: (i) it aligns with their religious language (idiolect), and (ii) it signifies group membership by speaking their idiolect.

On the other hand, covert intentional dogwhistles are more complex. They are often connected to the Norm of Racial Equality[12] (see Mendelberg 2001), which gained prominence after the 1960s when overt racism became increasingly unacceptable to most supporters[13]. However, what has remained largely unchanged among the target audience are implicit biases and a belief system referred to by psychologists as 'racial resentment' (see Mendelberg 2001) or 'symbolic racism' (see Tesler and Sears 2010). These dogwhistles are often not consciously recognized by both the general and target audience, partly because they appear to be unrelated to race. As a result, the audience does not oppose them in the same way as they would with explicitly racist dogwhistles. Furthermore, these dogwhistles are 'lending deniability if confronted with racism accusations' making them less risky than more overtly racist dogwhistles (see Saul 2018: 365). As an example of covert intentional dogwhistles, consider the Willie Horton advertisement used in George H. W. Bush's campaign against Michael Dukakis. At the time, Dukakis was leading Bush in the opinion polls. The ad, which was part of negative campaigning criticizing the prison furlough program, featured a black man Willie Horton, a furloughed convict who had raped a woman and stabbed a man in their home. Notably, the ad did not explicitly mention Horton's race. However, following the airing of the ad, Dukakis's lead in the opinion polls began to decline significantly. The ad was later labeled as "racist" and sparked extensive discussion. Subsequently, Dukakis's standing in the polls started to recover. Saul (2018: 366) argues that 'as the possibility of racism was raised, the ad ceased to function wholly on an implicit level. Viewers began to consider the possibility that something racial might be going on. And at this point, Dukakis started to rise in the polls again—some indication that the ad had ceased to be effective once race was explicitly under discussion'.

3.2 Unintentional dogwhistles

Another extremely prevalent category of dogwhistles are unintentional dogwhistles. This category encompasses instances where a dogwhistle is inadvertently transmitted, even though those who transmit it are not doing so intentionally and may not be aware of the dogwhistle in question. Nonetheless, these unintentional dogwhistles can produce the

[12] Mendelberg (2001), who introduced the norm, associates it with 'implicit political communication', while Lopez (2014) links it more directly to dogwhistles.

[13] Mendelberg's research indicates that prior to 1930s, American political discourse allowed explicit use of pejoratives for black people, asserting their inferiority to white people, and supporting legal discrimination in the form of enforced segregation or refusal to hire black people. However, The Norm of Racial Inequality began to erode from the 1930s to the 1960s (see Mendelberg 2001: 67).

same effects as the original intentional dogwhistle. According to Saul's (2018: 368) working definition, unintentional dogwhistles refer to the 'unwitting use of words and/or images that, used intentionally, constitute an intentional dogwhistle, where this use has the same effect as an intentional dogwhistle'. These unintentional dogwhistles can be either covert or overt, often contingent on the initial intentional dogwhistle. As an example of spreading an unintentional dogwhistle, consider the case of individuals such as reporters and TV producers who were responsible for re-showing and disseminating[14] the aforementioned Willie Horton ad from Bush campaign, without being aware that it was initially a covert[15] intentional dogwhistle they were propagating. Another example is the creation of a covert intentional racial dogwhistle by the Republican Party in the 1980s, resulting in the association of "government spending" with racial minorities. This association has since been unintentionally perpetuated in everyday discussions about a country's expenditure (see Saul 2018: 367–369).

3.3 Plausible deniability

Regardless of their subtype, both overt and covert dogwhistling involve coded communication that allows for two plausible interpretations. This characteristic provides different pathways for plausible deniability. For example, according to Khoo (2017), the discussion of deniable norm-violations and the ambiguity of code words like "inner city" involves examining their potential racial and non-racial interpretations. Khoo suggests that the social context and implications associated with these code words play a more significant role in their meaning than their semantic definition. The social meaning can introduce genuine ambiguity, particularly with racial connotations. Khoo's analysis revolves around the idea that code words, despite their plausible deniability, can still have racial effects. He proposes a minimal inference-driven account, which emphasizes that the social implications and racial associations of these code words contribute to their impact. This perspective acknowledges the potential for genuine ambiguity while recognizing the predictive power of plausible deniability and the racial effects that can result from the use of such code words.

In the context of political manipulation through dogwhistles, plausible deniability enables politicians and their target audience to deny the presence of a risky coded message, such as one that would label them as racists, when challenged. For instance, if confronted, they can explicitly deny endorsing or accepting the coded and risky message. Instead, they can endorse the second plausible interpretation, a non-risky message intended for the general audience and claim that it

[14] Saul (2018: 368) refers to them as "amplifier dogwhistles".

[15] For an example of an overt unintentional dogwhistle, see Saul's (2018) discussion of the Dred Scott dogwhistle.

was the only intended message. The plausible deniability of both overt and covert dogwhistles can be viewed as an improvement over unambiguously racist statements, as it provides politicians with a defense strategy. This factor may explain why politicians choose to employ dogwhistles. The following paragraphs examine in more detail how this dynamic unfolds in overt and covert dogwhistling.

In the case of an overt intentional dogwhistle[16], both the politicians and the target audience (as opposed to the general audience) are aware of the coded message. However, due to factors like implicit bias or racial resentment, one or both parties may not be aware of the message's riskiness. Nevertheless, regardless of their initial awareness of the riskiness, they can choose to deny the coded message when challenged. Even Saul highlights the deniability aspect of the overt Dred Scott dogwhistle used by Bush: 'Bush intends to have his anti-abortion message recognized, and recognized as intended. At the same time, though, use of a code phrase gives allows [sic] him to avoid placing his contribution on the record—thus achieving deniability' (Saul 2018: 371). It is also worth noting that while Saul had reservations about the effectiveness of an explicit racist dogwhistle because the target audience would likely recognize it as racist and resist it (see Saul 2018: 365), she admits that 'its efficacy would vary from voter to voter, but the deniability it would bring might well allow for a substantial degree of success. When [she] initially drafted [her] paper, [she] thought an explicit racial dogwhistle would fail, but [she is] now (post-Trump) not at all convinced' (Saul 2018: fn. 8).

In the case of a covert intentional dogwhistle, only the politicians (and their strategists) are aware of the coded message, while the target audience is not. Like in the case of overt intentional dogwhistles, both the politicians (and their strategists) and the target audience may not be aware of the riskiness of the coded message due to factors like implicit bias or racial resentment. Saul points out that these dogwhistles 'would appear on its face to be innocuous and unrelated to race—lending deniability if confronted with racism accusation' (Saul 2018: 365). For instance, in the case of the Willie Horton ad, there were 'no overtly racist assertions that are easily pointed to. And politicians can, and did, easily deny that there was racism in the ad or in their intentions' (Saul 2018: 381). However, like with overt intentional dogwhistles, regardless of their initial awareness of the message's riskiness, both the politicians and the target audience can choose to deny the coded message when challenged. Saul explains that 'it will indeed be *conversationally* challenging to make what has been covert explicit. People will reject what challengers say, and deny that it is true. Sanity may be, and often is, called into question. Challengers will be accused of having

[16] Similarly, once the spreaders become aware of the riskiness of their message, unintentional dogwhistles exhibit the same kind of plausible deniability.

a political agenda' (…) 'the intended audience of the speech acts will probably insist that the analysis is wrong and deny the existence of the covert material' (Saul 2018: 381).

4. Racial figleaves

Racial figleaf[17] utterances provide cover for more overt political manipulation than dogwhistles but also represent an improvement over unambiguously racist statements as they offer politicians a defense strategy. This section utilizes Saul's (2017b, 2024) work[18] to demonstrate that the effectiveness of racial figleaf utterances is consistent with plausible deniability. It also discusses their role in overt political manipulation.

> *Case #2. Racial figleaf:* When[19] Mexico sends its people, they're not sending their best — they're not sending you.[20] They're not sending you. They're sending people that have lots of problems and they're bringing those problems with us. They're bringing drugs. They're bringing crime. They're rapists. And some, I assume, are good people.

Political manipulation through racial[21] figleaves involves additional utterances that serve to provide cover for both racist-sounding statements within a political speech and the politicians themselves. In Case #2, a racial figleaf utterance, 'And some, I assume, are good people', is used alongside an explicitly racist-sounding statement, 'They're rapists', to deflect from the publicly unacceptable act of being a racist. If challenged on their racism, politicians (and their supporters) may use the figleaf utterance to argue that they did not mean that *all* of them are rapists, as they explicitly acknowledged that *some* of them are good people.

4.1. Racial figleaves: varieties and complexities

Most notably, Saul introduced the term 'racial figleaf' with the following definition: 'an utterance made in addition to an otherwise overtly racist one, that serves the function of calling into question the racism of the speaker and the utterance' (Saul 2017b: 98). Racial figleaves can

[17] See Saul (2017b: fn. 1) for, non-linguistic, human figleaves, and Saul (2024) for figleaves more generally.

[18] See fn. 7 of this paper.

[19] https://archive.nytimes.com/www.nytimes.com/politics/first-draft/2015/06/16/choice-words-from-donald-trump-presidential-candidate/

[20] "You" can be interpreted as another dogwhistle if one understands it as a coded reference to individuals who harbor resentment or implicit biases against "problematic foreigners".

[21] It's important to note that figleaves exist in other domains beyond race (see Saul 2024).

be categorized as synchronic[22] or diachronic[23], depending on whether the figleaf utterance is provided roughly at the same time or substantially later than the problematic utterance. Some well-known varieties of racial figleaves include the *Denial figleaf* ('I'm not a racist but….'), the *Friendship Assertion figleaf* ('Some of my best friends are … , but . . . [racist utterance]'), and the *Mention figleaf* ('I feel like saying [racist utterance]'). While the first two are often met with skepticism and their acceptance may depend on the audience, the *Mention figleaf* offers more complexity and room for plausible denial to be accepted (see Saul 2017b: 103–107).

Another more complex and arguably more credible racial figleaf is illustrated in Case #2. The typical interpretation of "they" in this speech is that it includes a *generic* statement that sounds racist, rather than a *universal*[24] statement, regarding Mexicans, Mexican (illegal) immigrants, or Mexican (illegal) immigrants sent by the Mexican government[25], which allows for some ambiguity. Unlike universal generalizations, generics have specific truth conditions that do not require all instances to possess the same characteristic for the statement to be considered true. The combination of this feature of generics with the additional figleaf utterance asserting that some of them are assumingly good people enables an interpretation in which both the generic statement and the figleaf utterance can be seen as true (see Saul 2017b: 105). It's important to note the confusion that arises from other instances of the pronoun "they" in the speech, such as 'They're bringing crime', which could also ambiguously refer to one of the aforementioned (sub-)groups of Mexicans or to those who sent them, e.g., the Mexican government. Furthermore, one could argue that the racial figleaf utterance 'And some, I assume, are good people' provides similar cover for

[22] Example of a synchronic racial figleaf: 'I´ve always had a great relationship with the blacks.' https://www.huffpost.com/entry/donald-trump-blacks-lawsuit_n_855553

[23] Example of a diachronic racial figleaf, uttered in an interview that took place after Trump had made several utterances considered to display antiblack racism: 'I have great African-American friendships. I have just amazing relationships, and so many positive things have happened'. https://edition.cnn.com/2015/12/13/politics/donald-trump-antonin-scalia-affirmative-action/

[24] It is worth noting that Tim Kaine, during a debate with Mike Pence, accused Trump of saying that *all* Mexicans are rapists. Pence denied this and mentioned Trump's qualification about some Mexicans being "good people". Fact-checking websites found Kaine's universal quantification to be false. Despite the clarification, some Trump supporters interpreted his statement as slandering all Mexicans, while Pence was able to defend Trump based on the specific wording. For an explanation of this rhetorical move in terms of specific characteristics of generics, see McKeever and Sterken (2021).

[25] For example, Saul shifts and expands the interpretation of "they" between narrower and broader readings, ranging from: 'Mexicans who are sent are rapists' (see Saul 2017b: 104) and 'The Mexicans who come to the United States are rapists' (see Saul 2017b: 105, 111, 112) to 'Mexicans are rapists' (see Saul 2017b: 97, 109, 113).

utterances that label "them" as problematic, criminals, drug users, or rapists.

4.2 Plausible deniability

According to Saul, a racial figleaf effectively counters the inference that the speaker has expressed overt racism, blocking the claim 'The speaker is racist' (see 2017a: 107). For example, the *Denial figleaf* directly denies the claim 'The speaker is racist', but it often fails to convince the audience of its sincerity, especially if the speaker has a history of making racial remarks. The *Friendship Affirmation figleaf*, such as saying 'Some of my best friends are black', attempts to refute the claim 'The speaker is racist' by suggesting that a racist wouldn't have close black friends. The *Mention figleaf* strategically includes the potentially racist utterance within quotation marks, making it more challenging to draw the inference that 'The speaker is racist'. However, it is worth noting that, *pace* Saul (2017b), one may argue that the effectiveness of racial figleaf utterance being consistent with plausible deniability is not because it needs to (always successfully) block the implicature that the statement is racist (or that the speaker harbors racist beliefs) but because it attempts to soften or deny racism by defusing charges of racism, as suggested in section 5[26].

As argued by Saul (2017a), a racial figleaf doesn't necessarily require a direct denial to prevent the inference that 'The speaker is racist'. It is sufficient for the figleaf to create confusion and uncertainty[27]. This understanding is based on a thin version of the Norm of Racial Equality, which aligns with the Ideology of Personalism. According to this ideology, racism is solely a matter of individual beliefs, intentions, and actions[28]. This thin version of the norm allows for the coexistence of racial resentment, which can be measured by agreement with statements like 'Blacks should do the same without any special favors, as Irish, Italian, Jewish, and many other minorities overcame prejudice and worked their way up[29]'. In Saul's words:

> Due to the Norm of Racial Equality, politicians attempting to exploit racial resentments need to be able to deny that this is what they are doing. Of course, it is far easier to make a convincing denial if you have avoided mentioning race. This is a significant advantage of using an implicit appeal/covert dogwhistle. However, figleaves can be used to provide deniability even when one has been more explicit. Indeed, as we have seen, this deniability

[26] I would like to acknowledge the anonymous reviewer for highlighting this concern.

[27] It is possible that Trump believes both Mexican immigrants are generally rapists and murderers, and that some of them are good people. Such a belief would explain his statements without an intention to hide racism, as he genuinely holds both views (see Saul 2017b: 104).

[28] Hill's (2008) work highlights the individualistic nature of racism.

[29] See Tesler and Sears (2010) who examine racial resentment and its relation to specific beliefs.

> may come in the form of simply denying racism, as in a Denial figleaf. However, the more subtle figleaves offer more possibilities. (Saul 2017b: 109)

On Saul's account, racial figleaf utterances serve as tools for politicians and their target audience to deflect, weaken, or create uncertainty around the potentially risky message that could label them as racists when confronted. This aligns with the idea that racial figleaves are linguistic devices that allow for plausible deniability. However, the effectiveness of different racial figleaves in providing plausible deniability can vary based not only on the type of linguistic device used but also on an individual basis. While none of these figleaves may be 100% effective, they can still be effective with certain groups while potentially less effective with others. For example, they may be less effective with the group targeted by the overt utterance, but their effectiveness does not necessarily need to be decisive[30].

Political figures, including Donald Trump, have effectively utilized racial figleaves to shape public opinion or advance discriminatory policies. By employing these figleaves successfully, they can navigate the boundaries of acceptable speech and, if challenged, maintain a certain level of defense, regardless of their true intentions, racial resentments, or biases. For instance, one could argue that Trump managed, in the 2016 elections, to tap into the prevalent anti-Latino immigrant sentiment in the country and amplify its reach. He did so by employing overt political manipulation using racial figleaves, which potentially garnered him additional votes. Surprisingly, Trump also experienced an increase in support from Latino voters in the 2020 election compared to the 2016 election. This increase in support could be attributed to factors such as his emphasis on economic growth, job creation, as well as his stance against socialist and leftist regimes. However, it is also worth considering that a good faith interpretation of his speeches, which may have included elements of plausible deniability, could have played a role in attracting Latino voters.

5. Generic stereotypes

Generic stereotypes most commonly take the form of a bare plural statement 'Ks are F', such as 'Blacks are violent[31]'. They can appear on their own or in conjunction with other linguistic devices like racial figleaves, which were discussed in the previous section. This section employs the concepts of majority and explanatory generalizations, as well as the ideas of defensive shifting and defensive weakening, to motivate plausible deniability of generic stereotypes. It also highlights their role in overt political manipulation.

Case #3. Generic stereotype: Mexicans are rapists.

[30] Similar can be noted for dogwhistles and generic stereotypes.

[31] For further reference, see Beeghly (2015).

Political manipulation through the use of generic stereotypes exposes the audience to generic beliefs, reinforcing prejudice and implicit biases, and perpetuating harmful or false stereotypes, such as racism or sexism (see Haslanger 2011; Langton, Haslanger and Anderson 2012; Rhodes, Leslie and Tworek 2012; Wodak, Leslie and Rhodes 2015; Leslie 2017; Saul 2017a; Wodak and Leslie 2017; Ritchie 2019; Fus 2021). In Case #3, when a generic stereotype about Mexicans is expressed, it can lead the audience to implicitly adopt the belief that there is a connection between being a rapist and belonging to the group of Mexicans. Generic sentences, unlike universally quantified ones like 'All Mexicans are rapists', allow for exceptions, making them easier to accept and harder to refute (Langton, Haslanger and Anderson 2012). If challenged or accused of racism, politicians (and their supporters) may attempt to claim that the intended meaning was not to imply a non-accidental connection[32] between being a rapist and being a Mexican, or that only some, not all, Mexicans are rapists, as Case #2 explicitly states.

It is worth noting that Case #3 can be viewed as an implicit generalization in Trump's political speech, building upon the previous Case #2. Specifically, as mentioned in section 4.1, his statement 'They are rapists' has been widely perceived as a derogatory and inflammatory generalization, raising questions about whether it pertains to Mexicans in general or specific subgroups within the Mexican population. If one does not agree with this interpretation of Case #2, it is still conceivable that a contemporary politician like Donald Trump could make such a generic stereotype. In case it may still be challenging to imagine it being uttered on its own, without a racial figleaf or a similar device, it is more plausible to consider a politician uttering a generic stereotype such as 'Muslims are terrorists'. For the purposes of this paper, it is sufficient to demonstrate that at least some generic stereotypes could be utilized.

5.1 Majority generalization and explanatory generalization

On the one hand, generic stereotypes can be understood as expressing a majority generalization. For example, they imply that most members of a specific social group, such as Blacks, possess a particular characteristic, like being violent. On the other hand, they can also be interpreted as expressing an explanatory generalization or in-virtue-of-kind-membership. For instance, they may attempt to explain a supposed biological essence, suggesting that Blacks are inherently violent

[32] This is not to claim that, ontologically speaking, racism requires essentialism. Statistical discrimination and statistical stereotyping can still be racist, especially when the alleged statistical connection is groundless. Instead, the denial through defensive shifting or weakening, as argued in 5.1 and 5.2 respectively, is supposed to capture the moves that speakers can employ to defuse charges of racism/sexism/etc., for the purpose of, hopefully, keeping the audience on their side.

in-virtue-of their membership in the social kind 'blacks' (Noyes and Keil 2019; Prasada and Dillingham 2009). However, when attempting to explain a biological essence, speakers engage in what is known as psychological essentializing, which should be distinguished from metaphysical essentializing (Leslie 2013, 2014, 2017). In this context, they treat a certain social group as if there is something inherent in the biological essence of women that makes them less capable of abstract thinking or as if there is a shared cultural or social essence that leads to immigrants being poor (Vasilyeva and Lombrozo 2020; Lemiere ms.).

It is also important to emphasize that generic stereotypes are commonly understood to convey both a majority generalization and an explanatory generalization (Haslanger 2014; Bian and Cimpian 2017; Rosola and Cella 2020; McKeever and Sterken 2021; Lemiere ms.). For instance, Rosola and Cella point out that accepting the statement 'women are bad at abstract thinking' leads one to believe that women, as a group, are generally poor thinkers, attributing this to their presumed 'women are poor thinkers and that this applies to almost every woman, due to their – supposed – underlying nature' (Rosola and Cella 2020: 743).

Given the above, in the context of political manipulation, when a politician (and their supporters) asserts a generic stereotype such as 'Mexicans are rapists', they are asserting both that *most* Mexicans are rapists and that they are inherently dangerous *in-virtue-of* their membership in the social group of Mexicans. If either the politician or their supporters are challenged for falsely claiming that *all* or *most* Mexicans are rapists, they can plausibly deny that they intended to make such a claim. They can argue that there is something *in-virtue-of* being Mexicans that predisposes them to being rapists, even if most of them are not or will never be rapists. It is worth noting that challengers can interpret generic stereotypes either as majority or explanatory generics, and those who are challenged can equally plausibly deny either interpretation. Which interpretation is denied depends on which one is perceived as riskier, and which one is being challenged. The takeaway is that the use of a generic form allows for plausible deniability, unlike more explicit quantification statements such as 'All Mexicans are rapists', 'Some Mexicans are rapists', or 'Most Mexicans are rapists', which do not offer the same level of plausible deniability.

5.2 Defensive shifting and defensive weakening

Langton, Anderson, and Haslanger (2012) employ the concept of defensive shifting to elucidate the two interpretations of generic stereotypes mentioned above. They propose a majority generalization interpretation, which is applicable to generics like 'Cabs are yellow' and 'Barns are red' (Prasada and Dillingham 2009; Prasada, Khemlani, Leslie and Glucksberg 2013). Additionally, they propose a characteristic (or explanatory) generalization interpretation, which suggests that generics

convey that 'those members which do have the property, are disposed to have it by virtue of the fact that they are members of the kind. (…) It does not follow, however, that all or most members of the kind have the property' (Langton, Haslanger and Anderson 2012: 763). Given these two distinct interpretations, Langton, Haslanger, and Anderson (2012) argue that when confronted with accusations of asserting prejudiced generic stereotypes, the speaker can defend themselves by shifting to either of the two interpretations:

> Does [Latinos are lazy] assert a majority generic or a characteristic [i.e. explanatory] generic? Interpret [it] as a majority generic. To combat it, one provides many counterexamples. However, the speaker can then suggest that, although many Latinos aren't lazy, they tend to be—thus embracing the characteristic generic. Instead interpret [it] as a characteristic generic. To combat it one provides evidence that, say, Latinos show no greater tendency towards laziness than any other group. The speaker can then suggest that, although it is not part of the nature or essence of Latinos to be lazy, most are. This slide back and forth between different interpretations of the utterance allows speakers to avoid taking responsibility for the implications of their claims. (Langton, Haslanger and Anderson 2012: 764)

According to them, defensive shifting occurs when, in order to mitigate the risks of their utterance, the speaker who has been challenged for making a generic generalization shifts between the majority and explanatory interpretations, depending on which interpretation they have been challenged on.

Lemeire (ms.), however, argues that Langton, Haslanger and Anderson's (2012) account of defensive shifting overlooks a crucial aspect of the plausible deniability phenomenon. As mentioned above, generic stereotypes are typically interpreted as both majority and explanatory generalizations, such as 'Tigers are striped' or 'Ravens are black', rather than solely as majority generalizations like 'Cabs are yellow' or solely as explanatory generalizations like 'Bulgarians are good weightlifters'. According to Lemeire (ms.), when challenged[33], the speaker (and her supporters) is usually challenged on one interpretation, either the majority generalization or the explanatory generalization. She can then plausibly deny the interpretation she is challenged on (e.g., the majority generalization) and rely on the unchallenged one (e.g., the explanatory generalization), even though she intended to communicate both. Consequently, Lemeire (ms.) and Bowker, Fus-Holmedal, Lemeire, Thakral (ms.), however, argue, such denials represent instances of '*defensively weakening* [of] the content of one's utterance, rather than as defensively shifting between two alternative interpretations'. Consider, for example, a statement made by senior Oxford academic Nick Bostrom: 'Blacks are more stupid than whites[34]'. Despite his sub-

[33] See Lemeire (2021) for two strategies for responding more forcefully to generically formulated stereotypes.

[34] https://thetab.com/uk/oxford/2023/01/12/senior-oxford-uni-academic-argues-blacks-are-more-stupid-than-whites-in-unearthed-emails-29768

sequent apology, he does not appear to renounce his comment regarding relative IQ. Instead, he attributes the disparity to social inequality resulting from unequal access to education, resources, and basic healthcare, rather than a genetic predisposition.

5.3 Plausible deniability

The upshot, then, is that no matter whether generic stereotypes involve defensive shifting[35] or defensive weakening, the accounts discussed above connect them to features of plausible deniability. Specifically, these accounts can be applied to the context of political manipulation to explain how politicians can spread risky (e.g., racially oppressive) messages[36] by incorporating elements of plausible deniability through the use of generic stereotypes in their speeches. This enables politicians and their supporters to mitigate the risks associated with their messages when challenged. Consequently, the presence of plausible deniability in generic stereotypes helps explain why they are favored as more overt forms of political manipulation, shedding light on the shift from covert to more overt strategies.

6. Further consequences and normative considerations

It appears that plausible deniability, as a political phenomenon, often relies on widespread public disagreement about what counts as racist speech—disagreement shaped by a thin interpretation of the Norm of Racial Equality (see section 4.2). Similar dynamics may apply to issues such as sexism, transphobia, and related forms of discrimination, which may be underpinned by similarly thin interpretations of their corresponding equality norms. If the public disagrees about what counts as racism/sexism/transphobia/etc., it is going to be hard for charges of racism/sexism/transphobia/etc. to stick, except in the most obvious cases. There will be political incentives to engage in racist/sexist/transphobic/etc. oppressive speech, especially when addressing audience with simplistic, actively ignorant views[37] about what racism/sexism/transphobia/etc. is and how it manifests in speech/thought/behavior.

This section presents some further arguments for why the plausible deniability of linguistic devices, such as dogwhistles, racial figleaves,

[35] A similar explanation in terms of defensive shifting and weakening could also be applied to dogwhistles and racial figleaves, further reinforcing the idea that they all share plausible deniability.

[36] This point applies regardless of whether the content is conversationally implicated or context-dependent semantic content.

[37] I would like to thank the anonymous reviewer for formulating the importance of this type of active ignorance. While it is important to acknowledge the potential significance of this type of active ignorance in plausible deniability, the limitations of this paper restrict further exploration. At the very least, the paper assumes a connection between active ignorance and the aforementioned thin Norm of Racial Equality.

and generic stereotypes, makes them effective tools for politically manipulative speech. Additionally, it briefly discusses normative considerations from the perspective of conceptual engineering that arise because of these consequences.

6.1 Efficient spread and perceived acceptability

The successful use of linguistic devices with plausible deniability in overt political manipulation can lead to a more efficient spread of the manipulative message. Plausible deniability allows both the transmitters and the recipients of the risky communicative message (e.g., a racist message) to plausibly deny its riskiness. This is especially important in cases where those spreading the message, such as politicians and their supporters, would choose not to transmit or accept it if they were aware of its riskiness. Plausible deniability provides a cover for those intentionally or unintentionally transmitting the risky message, enabling them to deny the risky message when challenged. Moreover, it can incentivize those who are aware of the message's riskiness to deliberately use these linguistic devices, knowing that they can deny it if challenged, thus facilitating the more efficient spread of the risky message.

Furthermore, politically manipulative speech utilizing linguistic devices with plausible deniability can, when successful, appear more acceptable despite its elements of overtness. Plausible deniability protects both the transmitters and the recipients of the risky communicative message by allowing them to (in certain cases) plausibly deny the risky interpretation. If accused of spreading a risky message, such as racism, politicians and their supporters can more easily deflect the accusation by leaning on the non-risky interpretation. Plausible deniability also relies on the principle of charity, which assumes that the speaker's intended message must be the best possible interpretation, in this case, a non-risky one that would not label them as racists. Alternatively, the principle of charity can be extended to one's intentions, so challengers should give the benefit of the doubt to politicians and their supporters once they have pointed out the plausible deniability of their message. Plausible deniability offers a rational alternative interpretation that suffices for charitable interpretation, allowing manipulators to escape more easily. This can further motivate political manipulators to exploit the principle of charity for their purposes.

It is worth noting again that other factors, such as implicit bias and racial resentment, may also facilitate the more efficient spread of the communicative message. These factors can make it harder for politicians and their supporters to perceive that they are spreading and accepting racism, for example. Furthermore, while politicians or their supporters may not initially be aware of the riskiness of their message or its potential for plausible denial, once challenged and made aware of the riskiness, they can still benefit from the plausible deniability of such linguistic devices. They may deny the interpretation that labels

them as racists because they do not want to explicitly commit to racism, even though they may be implicitly biased against different races.

6.2 Some normative considerations

The aforementioned consequences raise important normative considerations that can be addressed through the philosophical method of conceptual engineering. According to this approach, it is recognized that certain representational devices are deficient (descriptive claim) and should be improved (normative claim) (see Cappelen 2018; Burgess, Cappelen and Plunkett 2020).[38] In the context of this paper, one can argue that overt political manipulation through linguistic devices involving plausible deniability is deficient when it leads to socially, morally, or politically harmful effects[39] (descriptive claim), and thus, it should be improved (normative claim).

> *The first normative claim:* We should combat overt political manipulation through linguistic devices with plausible deniability when it results in pernicious effects.

It could be argued that the urgency for improvement is even greater in this case because, as discussed in section 6.1, overt political manipulation through linguistic devices with plausible deniability can appear more acceptable and spread more efficiently. At the same time, plausible deniability makes it challenging for individuals opposing such politically manipulative speech (e.g., anti-racists) to change the minds of politicians, their strategists, or their supporters regarding the dissemination and acceptance of such speech.

Addressing this deficiency requires comprehensive ameliorative projects that extend beyond the scope of this paper. While a systematic approach to ameliorating politically manipulative speech is lacking, recent literature has proposed some solutions for combating harmful effects both within and outside the context of political manipulation. Specifically, for dogwhistles, see Khoo (2017), and Saul (2018); for racial figleaves, see Saul (2017a); and for generics, see Leslie (2017), Saul (2017b), Ritchie (2019), Lemiere (2021), and Fus (2021: ch. 6).

So far, I have suggested that utilizing linguistic devices with plausible deniability for overt political manipulation results in morally, politically, or socially harmful effects, such as the spread of racism or sexism (descriptive claim). However, the plausible deniability of these linguistic devices is not inherently good or bad. On the contrary, some of its consequences, as described in section 6.1, can perhaps be harnessed to achieve morally, politically, or socially beneficial effects.

[38] Given that the deficiency and amelioration in this case do not pertain to concepts, a more suitable term at a higher-order level could be what Fus-Holmedal (2024) refers to as "philosophical engineering".

[39] For objectionable effects of the semantic value, such as socially, morally, or politically objectionable effects, as well as cognitive effects and effects on theorizing, see Cappelen (2018: 33–34), and Fus (2021: ch. 6).

In the context of conceptual engineering, Cappelen (2018) introduces a category of conceptual engineers he calls Exploiters of lexical effects. He argues that: 'There are of course Exploiters with good intentions, but the overall effect of their exploitation is to contribute to and encourage a use of language that undermines what we should treasure the most about it: the continuous exchange of ideas. Exploiters are in effect anti-intellectualist opportunists that contribute to a destruction of genuine communication' (Cappelen 2018:133-134). Similarly, politicians, their strategists, and their supporters can be seen as exploiters of the effects of plausible deniability in linguistic devices such as dogwhistles, racial figleaves, and generic stereotypes. However, one may argue, their aim should be not solely to promote their political agenda and gain power, but also to achieve morally, politically, or socially beneficial effects.

> *The second normative claim:* We should exploit the beneficial effects of overt political manipulation through linguistic devices with plausible deniability.

The underlying assumptions behind this second normative claim are that there can be morally good (political) exploiters or manipulators, and that it is permissible to exploit or manipulate to achieve beneficial goals. Given these assumptions, we should utilize the plausible deniability of linguistic devices such as dogwhistles, racial figleaves, and generics, when they can lead to morally, politically, or socially beneficial effects such as promoting social justice, minority rights, or gender equality (see Cappelen 2018; Fus-Holmedal 2024).

Consider the positive generic stereotype like 'Mexican immigrants are hardworking and have strong family value' that politicians could utter in their political speeches to challenge negative stereotypes and contribute to a more positive perception of a particular group. In some cases, epistemic goals should be overridden[40] by morally, politically, or socially beneficial goals, such as promoting social justice. For instance, introducing a false generic statement like 'girls play football' (see Saul 2017a; Ritchie 2019) could be permissible since, as Saul (2017a: 13) observes, such generics can be 'very important weapons in our anti-prejudice arsenal', or as Ritchie (2019: 38) argues: 'even if it is a false generic generalization, there may be good reasons to assert it in certain political contexts'. Conversely, in other cases, utilizing certain linguistic devices with plausible deniability could serve to express more accurate statements[41]. For example, using descriptively accurate and

[40] For a more in-depth exploration of overriding epistemic goals with morally, politically, or socially beneficial ones, see Fus (2021: ch. 6).

[41] Ritchie (2019) identifies accuracy as one of the primary benefits of social generics. She argues that generics can more accurately describe systematic patterns of violence and discrimination than explicitly quantified claims (Ritchie 2019: 34). This accuracy is one explanation for their social and political effectiveness (Ritchie 2019: 38).

true generic statements like 'women are expected to want children' or 'Blacks face economic, legal, and social discrimination' (see Ritchie 2019: 36) could aid in combating social injustice by providing justification for politicians to introduce policies that could improve the conditions of these marginalized social groups. It is important to notice that such policies may not be justifiable if universal versions of these statements were promoted instead.

One might, however, question the second normative claim—namely, that plausible deniability can be used for good, politically just ends—since examples such as 'Blacks face economic, legal, and social discrimination', 'women are expected to want children', and 'Mexican immigrants are hardworking and have strong family values' are not instances of plausible deniability due to the absence of racist/sexist/transphobic/ implicatures to be denied. In response to this concern, I would like to offer two points. Although these types of statements may not seem, at first glance, to be cases of plausible deniability, as those uttering or accepting them would not be (as frequently) challenged on their racist/sexist/transphobic/ undertones, it does not necessarily mean that they cannot contain such implications. Specifically, there is an interpretation of the above examples suggesting that they stereotype certain social groups, i.e. Blacks, women, Mexicans and can as such be seen to carry racist/sexist/transphobic/ implications. Even if they do not carry them in a strict sense, it does not invalidate the perception of certain audience members, who may interpret them as carrying racist/sexist/transphobic/ implications. The *perceived* racist/sexist/transphobic/ implication is as important in these cases, as political manipulation, whether good or bad, thrives on the audience's interpretation and perception of what can be seen as racist/sexist/transphobic/. For instance, certain members[42] of the audience opposing racism/sexism/transphobia/ might be especially inclined to perceive or emphasize that such statements can be seen as hostile, thereby challenging the ameliorators by focusing on racist/sexist/transphobic/ implications (whether real or perceived), rather than on the non-racist/non-sexist/non-transphobic/ values and goals they aim to promote.

It is also important to recognize that the efficacy of ameliorative approaches is largely an empirical question. This pertains to the broader context of implementing conceptual engineering and, consequently, extends to both the first and second normative claims—whether to counter harmful overt political manipulation through linguistic devices with plausible deniability, or to leverage the beneficial effects of such manipulation. Consider the case of generics, where we are only beginning to unravel the mechanisms dictating the influence of generic

[42] It is also worth noting that not every member of the audience needs to challenge the speaker even though the linguistic device with plausible deniability allows for such challenges. For example, those who already openly adhere to racist/sexist/transphobic/ statements are typically not the ones to challenge the speaker using linguistic devices with plausible deniability.

language on essentialist beliefs, which in turn can contribute to the formation of social stereotypes related to race, gender, transphobia, etc. Recent studies by Foster-Hanson, Leslie and Rhodes (2022) shed light on how generics shape children's concepts. Their findings suggest that simply making a generic claim (e.g., 'girls like pink', 'girls are good at reading'), even if the content is neutral or positive, can lead to the adoption of views that might reinforce stereotypes, regardless of the absence of explicit negative content in the generics themselves. These conclusions prompt discernment when using generics, even for well-intentioned purposes, as positive stereotyping can lead to unintended biases. Furthermore, their studies suggest that ameliorative strategies, which involve responses that maintain 'the generic scope of reference—even if it challenges claims about the referenced features (such as, "no, that's not right about girls" or even "well, boys like dolls too")—are unlikely to limit the spread of essentialist beliefs' (Foster-Hanson, Leslie and Rhodes 2022: 4). Merely negating generic statements about gender would not be sufficient to undermine their influence. Instead, their studies propose two potential solutions to this issue. First, to mitigate the possible negative consequences of the generic, 'one would need to directly challenge the generic scope of the sentence by limiting it to a specific person. For example, when a child hears (or utters themselves) a generic statement about a gender category, a parent might ask which particular person the child is referring to (e.g., "What person do you mean? Yes, Jimmy does like trucks")' (Foster-Hanson, Leslie and Rhodes 2022: 26). Second, to counteract the prescriptive effect of certain generics, the parent might also expand it to 'a superordinate category (e.g., "Lots of kids like trucks")' (Foster-Hanson, Leslie and Rhodes 2022: 26). These findings may suggest that endorsing the first normative claim may not necessarily support the second normative claim, at least in the context of generics.

To summarize, the primary objective of this section was to emphasize the additional consequences of manipulative messaging through linguistic devices with plausible deniability, such as their efficient spread and perceived acceptability. Additionally, the section explored two normative claims that incorporate the phenomenon of plausible deniability within overt political manipulation and suggested potential avenues for exploration within the field of conceptual engineering, without endorsing any specific ameliorative approach.

7. Conclusion

This paper suggested that linguistic devices with plausible deniability have played a significant role in enabling politicians to reintroduce and maintain some elements of overt messaging in the recent era, which was thought to have declined in the 1960s, when the Norm of Racial Equality gained prominence. It has shown how these devices can contribute to the resurgence of certain overt characteristics from the pre-

1960s era but in a more subtle and seemingly plausible manner. As a result, contemporary political speech has become more overt than it was approximately ten years ago while remaining more covert than it was 80–100 years ago.

Furthermore, the paper explored the role of plausible deniability in overt political manipulation, focusing on linguistic devices like dog-whistles, racial figleaves, and generic stereotypes. It discussed the phenomenon of linguistic plausible deniability and demonstrated how these devices can facilitate risky political manipulation. The paper also discussed the consequences that arise from plausible deniability, highlighting the power of linguistic devices with plausible deniability as tools for political manipulation.

Moreover, it contributed to the elevation of ethical and political considerations in the philosophy of language by discussing normative aspects related to plausible deniability and politically manipulative speech from the perspective of conceptual engineering.

References

Baram, M. 2011. *Donald Trump Was Once Sued By Justice Department For Not Renting To Blacks* <https://www.huffpost.com/entry/donald-trump-blacks-lawsuit_n_855553> accessed 24 June 2023.

Beeghly, E. 2015. "What is a stereotype? What is stereotyping." *Hypatia* 30 (4): 675–691.

Berstler, S. 2019. "What's the Good of Language? On the Moral Distinction between Lying and Misleading." *Ethics* 130 (1): 5–31.

Bowker M., Fus-Holmedal M., Lemeire O. and R. Thakral. Manuscript. "Weakening Generic Sterotypes."

Bian, L. and A. Cimpian. 2017. "Are Stereotypes Accurate? A Perspective from the Cognitive Science of Concepts." *Behavioral and Brain Sciences* 40, E3. doi:10.1017/S0140525X15002307.

Burgess, A., Cappelen, H. and D. Plunkett. 2020. *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Burns, A. 2015. *Choice Words From Donald Trump, Presidential Candidate* <https://archive.nytimes.com/www.nytimes.com/politics/first-draft/2015/06/16/choice-words-from-donald-trump-presidential-candidate/> accessed 24 June 2023.

Camp, E. 2018. "Insinuation, Common Ground, and the Conversational Record." In D. Fogal, D. W. Harris and M. Moss (eds.). *New work on speech acts*. New York: Oxford University Press, 40–65.

Cappelen, H. 2018. *Fixing Language: An Essay on Conceptual Engineering*. Oxford: Oxford University Press.

Dinges, A. and J. Zakkou. 2023. "On Deniability." *Mind* 132/526: 372–401, https://doi.org/10.1093/mind/fzac056.

Foster-Hanson, E., Leslie, S.-J. and M. Rhodes 2022. "Speaking of Kinds: How Correcting Generic Statements can Shape Children's Concept." *Cognitive Science* 46: e13223. https://doi.org/10.1111/cogs.13223.

Fricker, E. 2012. "Stating and Insinuating." *Proceedings of the Aristotelian Society Supplementary Volume* LXXXVI: 61–94.

Fus, M. 2021. *Assert This: 'Philosophers Are Engineers' (A Study of Philosophical Engineering and Generic Judgments*). PhD dissertation. University of St. Andrews and University of Oslo.

Fus-Holmedal, M. 2024. "In Defense of 'Philosophical Engineering': A Novel Terminological Dispute Resolution." In P. Stalmaszczyk (ed.). *Conceptual Engineering: Methodological and Metaphilosophical Issues*. Leiden, The Netherlands: BRILL | mentis, 135–159. https://doi.org/10.30965/9783969753026_008

Golshan, T. 2016. *Full transcript: Hillary Clinton and Donald Trump's final presidential debate* <https://www.vox.com/policy-and-politics/2016/10/19/13336894/third-presidential-debate-live-transcript-clinton-trump> accessed 21April 2023.

Goodin, R. and M. Saward. 2005. "Dogwhistles and Democratic Mandates." *Political Quarterly* 76 (4): 471–476.

Haslanger, S. 2011. "Ideology, Generics, and Common Ground." In C. Witt (ed.). *Feminist Metaphysics*: *Explorations in the Ontology of Sex, Gender and the Self*. Springer Verlag, 179–208.

Haslanger, S. 2014. "The Normal, the Natural and the Good: Generics and Ideology." *Politica & Società* 3: 365–392.

Hill, J. 2008. *The Everyday Language of White Racism*. Chichester: Wiley-Blackwell.

Khoo, J. 2017. "Code Words in Political Discourse." *Philosophical Topics* 45 (2): 33–64.

Langton, R. 2012. "Beyond Belief: Pragmatics in Hate Speech and Pornography." In I. Maitra and M. K. McGowan (eds.). *Speech and Harm: Controversies Over Free Speech*. Oxford: Oxford University Press, 72–93.

Langton, R., Haslanger, S. and L. Anderson. 2012. "Language and Race." In G. Russell and D. Graff Fara (eds.). *The Routledge Companion to Philosophy of Language*. NewYork: Routledge, 753–767.

Lee, J. J. and S. Pinker. 2010. "Rationales for Indirect Speech: The Theory of the Strategic Speaker." *Psychological Review* 117 (3): 785–807.

Lemeire, O. 2021. "Falsifying Generic Stereotypes." *Philosophical Studies* 178 (7): 2293–2312.

Lemiere, O. Manuscript. "The Strong yet Deniable Meaning of Generic Stereotypes."

Leslie S. J. 2013. "Essence and Natural Kinds: When Science Meets Preschooler Intuition." In T. Gendler and J. Hawthorne (eds.). *Oxford Studies in Epistemology*. Oxford: Oxford University Press, vol. 4, 108–165.

Leslie S. J. 2014. "Carving up the Social World with Generics." In J. Knobe, T. Lombrozo and S. Nichols (eds.). *Oxford Studies in Experimental Philosophy*. Oxford: Oxford University Press, vol. 1, 208–231.

Leslie S. J. 2017. "The Original Sin of Cognition: Fear, Prejudice, and Generalization." *Journal of Philosoph,* 8: 393– 421.

Lopez, I. 2014. *Dog Whistle Politics: How Coded Racial Appeals Have Reinvented Racism and Wrecked the Middle Class*. New York: Oxford University Press.

Mazzarella, D. 2021. "'I Didn't Mean to Suggest Anything Like That': Deniability and Context Reconstruction." *Mind & Language* 1–19.

Mazzarella, D., Reinecke, R., Noveck, I. and H. Mercier. 2018. "Saying, Presupposing and Implicating: How Pragmatics Modulates Commitment." *Journal of Pragmatics* 133: 15–27.

McGowan, M. K. 2004. "Conversational Exercitives: Something Else We Do With Our Words." *Linguistics and Philosophy* 27 (1): 93–111.

McGowan, M. K. 2012. "On 'Whites Only' Signs and Racist Hate Speech: Verbal Acts of Racist Discrimination." In I. Maitra and M. K. McGowan (eds.). *Speech and Harm: Controversies Over Free Speech*. Oxford: Oxford University Press, 121–147.

McKeever, M. and R. Sterken. 2021. "Social and Political Aspects of Generic Language and Speech." In Khoo, J. and R.K. Sterken (eds.). *The Routledge Handbook of Social and Political Philosophy of Language*. New York: Routledge, 259–280.

Mendelberg, T. 2001. *The Race Card: Campaign Strategy, Implicit Messages, and the Norm of Equality*. Princeton: Princeton University Press.

Noyes, A. and F.C. Keil. 2019. "Generics Designate Kinds but not Always Essences." *Proceedings of the National Academy of Sciences* 116 (41): 20354–20359.

Peet, A. 2015. "Testimony, Pragmatics, and Plausible Deniability." *Episteme* 12 (1): 29–51.

Peet, A. 2024. "The Puzzle of Plausible Deniability." *Synthese* 203 (156): https://doi.org/10.1007/s11229-024-04600-4.

Pinker, S. 2007. "The Evolutionary Social Psychology of Off-record Indirect Speech Acts." *Intercultural Pragmatics* 4 (4): 437–461.

Prasada, S. and E.M. Dillingham. 2009. "Representation of Principled Connections: A Window onto the Formal Aspect of Common Sense Conception." *Cognitive Science* 33 (3): 401–448.

Prasada, S., Khemlani, S., Leslie, S. J. and S. Glucksberg. 2013. "Conceptual Distinctions amongst Generics." *Cognition* 126 (3): 405–422.

Rhodes, M., Leslie, S.J. and C. Tworek. 2012. "Cultural Transmission of Social Essentialis." *Proceedings of the National Academy of Sciences (PNAS)* 109 (34): 13526–13531.

Ritchie, K. 2019. "Should We Use Racial and Gender Generics?" *Thought: A Journal of Philosophy* 8: 33–41.

Rosola, M. and F. Cella. 2020. "Generics and Epistemic Injustice." *Ethical Theory and Moral Practice* 23 (5): 739–754.

Safire, W. 2008. *Safire's Political Dictionary*. New York: Oxford University Press.

Saul, J. 2017a. "Are Generics Especially Pernicious?" *Inquiry*. doi:10.1080/0020174X.2017.1285995.

Saul, J. 2017b. "Racial Figleaves, the Shifting Boundaries of the Permissible, and the Rise of Donald Trump." *Philosophical Topics* 45 (2): 97–116.

Saul, J. 2018. "Dogwhistles, Political Manipulation, and Philosophy of Language." In F. Daniel, C. Matt and D. Harris (eds.). *Dogwhistles, Political Manipulation, and the Philosophy of Language*. Oxford: Oxford University Press, 360–383.

Saul, J. 2024. *Dogwhistles and Figleaves: How Manipulative Language Spreads Racism and Falsehood*. Oxford: Oxford University Press.

Scott, E. 2015. *Trump Hits Scalia Over Comments on Black Students* <https://edition.cnn.com/2015/12/13/politics/donald-trump-antonin-scalia-affirmative-action/> accessed 24 June 2023.

Stanley, J. 2015. *How Propaganda Works*. Princeton University Press.

Tesler, M. and D. O. Sears. 2010. *Obama's Race: The 2008 Election and the Dream of a Post-Racial America*. Chicago: University of Chicago Press.

Time stuff. 2015. *Here's Donald Trump's Presidential Announcement Speech* <https://time.com/3923128/donald-trump-announcement-speech/> accessed 21 April 2023.

Valentino, N., Hutchings, V. and I. White. 2002. "Cues That Matter: How Political Ads Prime Racial Attitudes During Campaigns." *American Political Science Review* 96 (1): 75–90.

Vasilyeva, N. and T. Lombrozo. 2020. "Structural Thinking about Social Categories: Evidence from Formal Explanations, Generics, and Generalization." *Cognition* 204: 104383.

Walton, D. 1996. "Plausible deniability and evasion of burden of proof." *Argumentation* 10 (1): 47–58.

Witten, K. Manuscript. "Dogwhistle Politics: The New Pitch of an Old Narrative."

Wodak, D., Leslie, S. J. and M. Rhodes. 2015. "What a Loaded Generalization: Generics and Social Cognition." *Philosophy Compass* 10 (9): 625–634.

Wodak, D. and S. J. Leslie. 2017. "The Mark of the Plural: Generic Generalizations and Race." In Taylor, P. C., Alcoff, L. M. and L. Anderson (eds.). *The Routledge Companion to the Philosophy of Race.*, accessed 21 April 2023, Routledge Handbooks Online.