C O P *i* L CAMBRIDGE OCCASIONAL PAPERS *in* LINGUISTICS
Volume **17**, Article 9: pp. 157–172, 2025 | ISSN: 2050-5949

Linguistic Insights into Implicit Bias: Identifying Unconscious Mechanisms in Discourse

PAMELA GITANI
UNIVERSITY OF CAMBRIDGE

1 Introduction

Implicit bias can be defined as people's automatic tendency to associate certain traits with members of a particular social group (Saul 2013: 244). It is assumed that implicit bias triggers judgement errors which are particularly pernicious, notably because they drive decisions which are founded on inadequate criteria, such as unconsciously assessing the quality of job candidates based on their social category rather than their competence. The present work makes an original contribution to the issue by approaching it from a linguistic point of view. It addresses the following research questions:

(RQ1): How does implicit bias reveal itself in discourse?

(RQ2): Does implicit bias pertain to utterance meaning?

Although this essay uses the methods of argumentation and conceptual analysis, the question of whether and how the linguistic expression of implicit bias can be tested empirically is kept in mind. In Section 1, I explain how implicit bias has been characterised in the philosophical literature, and propose an operational definition for its manifestation in discourse. Section 2 covers the differences between Minimalism and Contextualism, and elucidates why Contextualism may be a better framework for analysing implicit bias. The limitations of my approach are tackled in Section 3. Section 4 consists in the analysis proper, which aims to answer RQ1 and RQ2. Questions for future research are exposed in Section 5, and the main conclusions are summarised in Section 6.

2 Implicit Bias in Discourse

2.1 What is implicit bias?

Saul (2013: 244) defines implicit bias (henceforth IB) as a tendency to 'automatically associate concepts with one another'. This bias is assumed to be grounded in unconscious prejudice against certain social categories (e.g. based on race, gender, sexual orientation, etc.). One prototypical instantiation of implicit bias arises in job application processes (Saul 2013: 244–45). CV studies have demonstrated that participants tend to have more negative evaluations of candidates with foreign-sounding names than those with a typically white-sounding name, even if the content of the CV is identical. In that sense, IB is closely related to Fricker's (2007) notion of testimonial injustice: a type of injustice which occurs 'when prejudice causes a hearer to give a deflated level of credibility to a speaker's word' (Fricker 2007: 1). In the case of the assessment of CVs, there is no speaker's word proper, but the document testifies to the candidate's skills and professional trajectory, which can be taken less seriously because of the reviewer's (unconscious) prejudices. IB also goes beyond testimonial injustice in a more pernicious way: namely, because it is ubiquitous and can arise 'when we think we are evaluating evidence or methodology' (Saul 2013: 248). Saul's main thesis is that IB constitutes distortions to our judgement which cannot easily be rectified by self-reflection. She argues that if we let unconscious prejudice drive our decisions, then we are probably not doing as well as we could from a moral point of view (Saul 2013: 256). It is therefore our responsibility to exercise action and counter the effect of such biases.

Although Saul's definition of IB is largely supported by philosophers, it is not entirely accurate (Holroyd & Sweetman 2016). This stems from the confusion around the meaning of 'implicit' (De Houwer 2006: 12, Hahn, Judd, Hirsh & Blair 2014: 1370, Holroyd 2015). Research on IB originated in social psychology: it pertains to the use of implicit measures for uncovering people's attitudes towards certain social categories, and the effect of those attitudes on judgements and behaviours (Fazio & Olson 2003: 301, Machery, Faucher & Kelly 2010: 229). 'Implicit' thus refers to the fact that subjects are unaware of what is being measured, rather than of their own attitudes (De Houwer 2006: 13). The advantage of such measures is that they 'are likely to be free of social desirability concerns' (Fazio & Olson 2003: 301). Most experimental studies on IB draw comparisons between implicit associations and explicit beliefs. If the results of an Implicit Association Test (IAT), which looks into the rapidity at which participants associate concepts with one another (e.g. a specific race with some characteristics or constructs), go unreported in reflective statements targeting the participants' beliefs, then it is assumed that the subjects may unconsciously let their bias affect their judgement or behaviours (Holroyd & Sweetman 2016: 85). That said, it is important to note that IAT uncovers automatic or spontaneous associations between concepts which are not necessarily unconscious.

Functional definitions such as Saul's (2013) may be useful for discussing IB in general terms (Holroyd & Sweetman 2016: 81). However, considering that the bias itself is necessarily unconscious may be an overstatement. In light of what has been exposed in the previous paragraph, I suggest amending Saul's definition by relying on the following criteria:

i. Implicit bias is founded on stereotypes (positive or negative) pertaining to certain social categories;

ii. The association between concepts is automatic (but not necessarily unconscious);

iii. People may be unconscious of the influence these associations have onto their judgements and behaviours.

2.2 The linguistic expression of implicit bias

The primary goal of this research is to determine how implicit bias arises in discourse (RQ1). Experimental studies on IB usually appeal to linguistic form at two levels: (i) in the presentation of at least one of the concepts making up pairs in IAT; (ii) in the statements of explicit belief which the subjects are asked to assess.

However, this does not tell us whether spontaneous associations of the type BLACK = HOSTILE (ii in Subsection 2.1) can unconsciously arise in discourse (iii)[1]. Since IB stems from stereotypes (i), looking at how these are conveyed in language offers a satisfactory starting point. Consider the following examples:

(1) 'We advertised for a new nanny.'
(from Jaszczolt 2006: 203, after Levinson 2000.)

(2) 'I like my black friend Martin.'
(from Hahn et al. 2014: 1370)

(3) 'You won't be happy living in this neighbourhood'
(adapted from Camp 2018: 43)

It is reasonable to assume that readers of Example (1) are very likely to spontaneously picture a female referent. The case of Example (2) is slightly different: it could be read as an instantiation of racial prejudice, to the extent that specifying Martin's race is superfluous. Lastly, Example (3) may suggest that the addressee will not feel safe in that neighbourhood, for instance, because there are too many immigrants.[2]

Recall that I do not equate stereotypes with biases but rather consider them as constitutive of biases. Indeed, one could (i) be aware of the existence of a stereotype, (ii) harbour this stereotype, (iii) convey the stereotype in discourse (consciously or not), or any combination of the three. Relying on the three defining criteria from Subsection 2.1, I take the automatic associations between concepts to be the representation of a stereotype. I take the linguistic expression of IB to be an unconscious mechanism which involves inserting a stereotype into discourse. When I say that a person is biased, it means that s/he harbours a stereotype; s/he may be aware of holding these views but is not necessarily aware of all the situations in which that bias resurges. Table 1 summarises the key components of this operational definition.

[1] I am intentionally setting aside the question of whether language pertains to judgement or behaviour, as it is not crucial for my discussion. For a better understanding of the semantic/affective distinction within IB and its impact on judgements and behavioural outcomes, see Holroyd & Sweetman (2016).

[2] In the original example ('Perhaps you would feel more comfortable locating in a more… transitional neighborhood'; Camp 2018), the utterance is attributed to a realtor and is addressed to potential buyers. The latter are assumed to belong to a marginalised group. Therefore, the target of the stereotype in Camp's example are the buyers, not the local majority.

Component	Characteristics
Stereotype	Automatic association between concepts, e.g. BLACK = HOSTILE
Biased individual	Person who harbours a stereotype
Insertion of stereotype	Unconscious mechanism

Table 1 Characteristics of each component of the linguistic expression of implicit bias

At this stage, we do not know which mechanisms underlie the insertion of stereotypes into discourse (RQ1). We also do not know whether the fact that stereotypical views can be unconsciously revealed in discourse implies that this process contributes to the meaning of what has been uttered (RQ2). To address these questions, we need to clarify what utterance meaning is and via which processes it is arrived at. This is what the following section purports to do.

3 Meaning, Inferences, and Associative Processes

There are three main dimensions at play in the distinction between the linguistic meaning of a sentence, and its meaning in use (Recanati 2003: 5). These are *sentence meaning*, *what-is-said*, and *implicatures*. The way these three dimensions are grouped together yields two different approaches to utterance meaning: Minimalism and Contextualism. In this section, I briefly describe the main tenets of each theory and explain why I favour Contextualism over Minimalism for the analysis of IB. I focus on what different versions of Contextualism (i.e. Relevance Theory; Wilson & Sperber 2012 and Default Discourse Semantics; Jaszczolt 2006) consider as the primary meaning of an utterance, and on the type of processes (i.e. inferential vs associative) involved in the determination of primary meaning. I then select the version of Contextualism which I deem most adequate for the analysis in Section 5.

According to Minimalism, sentence meaning and what-is-said form together the literal meaning of the utterance. This literal meaning contrasts with speaker meaning; only the latter is a matter of speaker intentions. An argument which has recently been brought up in defence of Minimalism is linguistic liability: the fact that there are some contexts in which speakers may be held liable for the literal meaning of their utterances (Borg & Connolly 2022). Although Borg herself (2004: 3), as a defender of Minimalism, has claimed in prior work that a theory of linguistic meaning should not purport to be a theory of communication, Borg & Connolly (2022: 13) argue that the possibility of strict linguistic liability is an indication that the minimal literal content is consciously available to conversational participants. In non-minimalist frameworks such as Contextualism, what-is-said belongs to speaker meaning and is pragmatically enriched. It is part of what is intended by the speaker, rather than being a property of the sentence, and reflects what intuitively seems to be said (Recanati 2003: 18). Given the nature of IB and its ethical implications, it seems logical to rely on a theory which reflects as closely as possible how conversational interactants intuitively make sense of utterances. This leads me to focus on pragmatic theories and exclude Minimalism (but see Section 6).

In Recanati's Contextualism (also referred to as Truth-Conditional Pragmatics, henceforth TCP), the pragmatic what-is-said (i.e. the primary meaning) has to be consciously accessible, since it is 'a matter of intention-recognition' (Recanati 2003: 14). Sentence meaning together with contextual ingredients contribute to the determination of primary meaning via *primary pragmatic processes*; these are associative (i.e. involve the activation of associatively related concepts), and unconscious. *Secondary pragmatic processes*, on the other hand, intervene in implicature-generation; they are inferential and conscious. On this account, inferences thus always involve a conscious process from premises to conclusion and yield implicatures, which constitute a separate thought. These inferences can also be spontaneous, which does not mean that they are unconscious (Recanati 2002: 118).

Relevance Theory (henceforth RT) also adopts a version of what-is-said which is pragmatically enriched – this is called an *explicature* (Carston 2004: 819; Carston 2007: 18; Carston 2013: 2). The fundamental difference between RT and TCP lies in the nature and psychological reality of pragmatic processes. According to RT, the idea that primary pragmatic processes are associative rather than inferential is problematic, notably in cases of non-literal uses of language such as irony: in those cases, Carston (2007: 30) argues, what counts as primary meaning and how is it arrived at, if it cannot be inferentially derived from some consciously accessible content? For Relevance Theorists, inferences are ubiquitous and not necessarily conscious; they claim, contra Recanati (2002), that communication cannot be as direct as perception. In his reply to Carston, Recanati (2007: 51) admits that metarepresentational elements – which presuppose an inferential process – can occur at the primary level but are constitutive of secondary processes. But this leads him to what Jaszczolt (2015: 769) describes as 'formidable complications of the theoretical apparatus', notably because he essentially maintains the same distinction between primary meaning and implicature. In Default Discourse Semantics (henceforth DDS; see Jaszczolt 2006, 2011, 2015), Jaszczolt takes an even more radical view on primary meaning, by freeing it from any syntactic constraint: on this account pragmatic processes can but do not necessarily develop the logical form. DDS aims to reflect the inner workings of communication in a more psychologically real way, by 'model[ling] utterance meaning as intended by the Model Speaker and recovered by the Model Addressee' (Jaszczolt 2015: 744). It distinguishes between four processing mechanisms: syntactic processing, conscious pragmatic inference, cognitive defaults, and social/cultural/world-knowledge defaults. Jaszczolt likens defaults to automatic and unconscious interpretations, which contrast with conscious, inferential processes. This seems to concur with Recanati's idea that not all communication is inferential. The key difference, however, is that the conscious/unconscious nature of the processing mechanisms in DDS does not determine the level of meaning (primary vs secondary) that goes through.

This section shows that even within Contextualism, some issues around primary meaning remain unresolved. This could be imputed to the meaning overlap across terms such as 'automatic', 'unconscious', 'unintended', 'implicit', although these are not actual synonyms[3]. However, one framework needs to be selected for the rest of the discussion. Although I recognise the appeal of the abandonment of the syntactic constraint in DDS, I will set this theory aside, notably because it does not take a clear stance on what counts as an inference and on which pragmatic mechanisms contribute to primary meaning. Indeed, in my view, a clear account of the distinction between inferential vs associative is required in order to understand whether IB processes map onto utterance interpretation processes. Additionally, DDS does not exclude the possibility that meanings develop the logical form of sentences. Therefore, I will maintain the traditional distinction between pragmatic what-is-said and implicatures, as in RT and TCP. Next, it seems that what some call 'inferring', others call 'reasoning' (Recanati 2002: 81; Mercier & Sperber 2017: 51). Since the question of whether communication is essentially inferential has not yet been resolved, and since I consider that the distinction between processes which appeal to reasons and those which do not matters more than the label chosen, I will settle on the term 'inference'. This leads me to adopt TCP and its distinction between inferential and associative processes. The motivation for this is that associative processes in TCP are reminiscent of the processes IAT measures tap into (see Subsection 2.1). This means that similar paradigms could potentially be used in future experimental research to answer RQ2.[4]

4 Limitations and Rebuttals

As mentioned in the introduction, this paper, though purely theoretical, purports to lay the ground for empirical research. However, some of the utterances analysed in the following section were constructed for the purpose of the analysis. I also chose to adopt an 'omniscient' perspective on what goes on in the minds of speaker and addressee, meaning that I establish at the start who is biased and who is not. These manoeuvres can easily give the impression that I am artificially creating an object of study just for the sake of studying it. Although I am aware that my method of inquiry is imperfect, I would object to the idea that this scientific endeavour is altogether pointless, since this work is exploratory. The representation provided in Section 5 should therefore be taken as a working hypothesis for the investigation of IB in language, rather than an actual depiction of what goes on in the mind of conversational interactants.

[3] For instance, for Recanati (2003: 42) 'automatic' does not equal 'unconscious'.

[4] I do not mean to suggest that TCP should only be selected based on how easily it can be tested. My assumption is rather the following: IAT is usually taken to be an accurate measure of (spontaneous) associative processes; therefore, if an experimental paradigm can successfully reflect how IB and stereotypes are processed and if this process seems to be associative, then it will be easier to answer RQ2. However, the question remains as to whether this paradigm would allow for a distinction between purely associative processes, and automatic/spontaneous (but conscious) inferences (i.e. should such a distinction exist).

5 Analysis

This analysis consists in two steps: uncovering the mechanisms which reveal IB in discourse (RQ1), and determining whether IB constitutes a pragmatic enrichment of what-is-said, a separate thought (i.e. an implicature), or neither (RQ2). The parameters involved in the analysis are listed in Table 2.

Component	Characteristics	Research Questions
Stereotype	Automatic association between concepts, e.g. BLACK = HOSTILE	RQ1
Biased individual	Person who harbours a stereotype	RQ1
Insertion of stereotype	Unconscious mechanism	RQ2

Table 2 Parameters used for answering RQ1 and RQ2

To answer RQ1, we first need to identify the stereotype. Table 3 displays utterances (1)-(2)-(3) from Subsection 2.1 in the first column. The second column shows the association of concepts making up the stereotype; the third one represents in linguistic form how the information pertaining to the stereotype can be conveyed.

Note that this information (i.e. last column) is not necessarily conveyed. For instance, the collocation of 'black' and 'friend' in (2) could be used to disambiguate the referent (i.e. if the speaker has another friend called 'Martin'); and (3) could convey that this neighbourhood is not an option because, for instance, it would significantly lengthen the addressee's commute to her workplace. This hints at the fact that background knowledge about conversational participants, context, co-text, and even paralinguistic elements such as prosody, might either block or help to convey a stereotype. What Table 3 also shows is that there seems to be a cline from (1) to (3): in (1), the stereotype seems to be conveyed through the very

Utterance	Stereotype	Information conveyed
(1) 'We advertised for a new nanny.'	NANNY = FEMALE	We advertised for a female nanny.
(2) 'I like my black friend Martin.'	BLACK = HOSTILE	I like my friend Martin, even though he's black.
(3) 'You won't be happy living in this neighbourhood.'	IMMIGRANT= UNDESIRABLE	You won't be happy living in this neighbourhood – you won't feel safe because there are too many immigrants.

Table 3 Utterances (1)–(3), related stereotypes, and how these are conveyed

use of the lexeme 'nanny', while in (2) and (3) it emerges from a causal link. (2) differs from (3) in that the speaker's views are expressed through a (perceived) oxymoron (i.e. instantiated in the phrase 'black friend'), while in (3) the stereotype is conveyed in an even less direct way (i.e. some sort of embedding of causal links of the type '*x* because *y* because *z*', with *y* and *z* referring to the concepts making up the stereotype).

Let us now assess whether the process which reveals the stereotype could be considered as unconscious and therefore qualify as IB. Table 4 focuses on utterance (1) and shows how the stereotype that 'only women can be nannies' can be conveyed[5]. Several variants of (1) consisting in slight changes in co-text or lexical choice are provided. This is done to test the extent to which the stereotype is systematically conveyed whenever 'nanny' is uttered. The other columns focus on the first three parameters listed in Table 2: the information pertaining to (i) is given; (ii) results from the manipulations of the example; the mechanism (iii) which exposes the stereotype is described in the fourth column. The last column establishes whether the process is unconscious (this also relates to iii).

Except for (1g), anaphora is the primary linguistic mechanism which reveals the stereotype NANNY = FEMALE. However, anaphora alone does not suffice (see Footnote 8). Moreover, in some cases, the revelation of the stereotype can only be achieved interactionally, as in (1c) and (1d). (1g) stands out for two reasons. First, the speaker's bias is revealed through a specific collocation rather than anaphora. Second, it can also be read as an anticipation on the part of the speaker that the addressee will spontaneously picture a female referent. This sheds light on the fact that this socio-cultural stereotype may be so entrenched that it is nearly impossible to block it[6]. Assuming that most members of the linguistic community are aware of this, it may be that the phenomena displayed in Table 4 are borderline cases of IB, because the insertion of the stereotype is not always unconscious. Regarding examples (2) and (3), similar mechanisms would apply, although anaphora is understandably not one of them, since (2) and (3) do not involve gender. Turning to RQ1, it seems that it is always possible to find contexts in which the stereotype is revealed, and that the mechanisms involved in the process can be (but are not always) unconscious[7]. Therefore, I consider that the utterances (1)-(2)-(3) fall within the definition of IB I have provided.

[5] Similar tables for (2) and (3) can be found in the Appendix. I only focus on (1) here for reasons of space and because it is the most complex example.

[6] Of note: Levinson (2000: 223) would consider those instances as concept enrichment via a stereotype.

[7] It is actually because the activation of the stereotype may be unconscious that more co-text is needed to fully grasp the phenomenon (e.g. if the speaker whose utterance conveys a stereotype is being called out by the addressee). One should also note that after being called out, the original speaker can (genuinely or not) deny that s/he is biased. This further complexifies the analysis of IB in discourse. Other methods may make it possible to uncover the nature of the process without the need to rely on conversation analysis (see Section 6).

Example	Variants of (1)	Biased individual (i)	Revealer (ii)	Mechanism (iii)	Unconscious process?
(1a)	A: We advertised for a new nanny. she should be available every day from 3 to 6 pm.	Speaker	Speaker	Anaphora in co-text (same speaker).	Yes[8]
(1b)	A: We advertised for a new nanny. B: I hope she'll be better than the previous one!	Speaker? + Addressee	Addressee	Anaphora in addressee's turn. No reaction to B's response, so nothing can reveal A's bias.	Yes (for addressee)
(1c)	A$_{t1}$: We advertised for a new nanny. B: I hope she'll be better than the previous one! A$_{t2}$: So do I.	Speaker Addressee	Speaker + Addressee	Anaphora in addressee's turn and not rejected by A in $_{t2}$.	Yes (for both)
(1d)	A$_{t1}$: We advertised for a new nanny. B: I hope she'll be better than the previous one! A$_{t2}$: Or he. Men can be nannies too.	Addressee	Addressee	Anaphora in addressee's turn rejected by A in $_{t2}$.	Yes
(1e)	A: We employed a new nanny.	Speaker	/	Taken on its own, (1e) does not reveal anything because it is assumed that the speaker knows the gender of the referent.	/
(1f)	A: We employed a new nanny. B: Oh nice! How is she?	Speaker? + Addressee	Addressee	Anaphora in addressee's turn (similar to (1b) and (1c)). Only additional turns could reveal A's endorsement of the stereotype.	Yes (for addressee)
(1g)	A: We employed a male nanny.	Speaker?	/	The collocation of 'male' + 'nanny' reinforces the idea that there is something unusual about male nannies.	Unclear, more co-text and context needed.

Table 4 Mechanisms at play and role of speaker/addressee in the revelation of stereotypes in examples based on utterance (1) 'We advertised for a new nanny.' Underlined = Anaphoric reference, **?** = Cannot be clearly revealed without additional turns, $_{t1}$ = First turn, $_{t2}$ = Second turn (same speaker as $_{t1}$)

Let us now move on to RQ2. Given the theoretical framework chosen, the remaining task consists in determining whether IB processes map onto the pragmatic processes delineated in TCP (at least, theoretically). This forces me to only focus on the addressee's side (i.e. on how a hearer recovers information pertaining to a stereotype which has been unconsciously conveyed by the speaker). This is shown in Table 5.

Utterance	**Process Involved in Recovering Stereotype**	**Primary/Secondary Process**	**Level of Meaning**
(1) 'We advertised for a new nanny'	Spontaneous activation of related concepts	Primary (associative)	What-is-said
(2) 'I like my black friend Martin.'	Spontaneous activation of black = hostile followed by clash of concepts (black vs friend)	Secondary (associative then inferential)	Implicature?
(3) 'You won't be happy living in this neighbourhood.'	Inference (causal links)	Secondary (inferential)	Implicature?

Table 5 Utterances (1)-(2)-(3): Comparison between processes involved in recovering stereotypes and TCP processes

Table 5 shows that the addressee may be aware of the representations which are activated, either because they are widespread (as in Example 1), or because they are part of an inferential process (as in examples 2 and 3). However, the fact that the information pertaining to a stereotype can be recovered by the addressee and therefore be part of her interpretation of the utterance, does not entail that she is aware that what guides her interpretation qualifies as a stereotype. This applies to speakers too: an utterer of (1) can even mean 'a female nanny', without knowing that this conveys a stereotype. This makes me wary of considering (2) and (3) as implicatures, since the content inferred by the addressee (i.e. something along the lines of 'some black people are surprisingly friendly') can correspond to (i) what the speaker actually thinks and intends to convey (implicature), (ii) what the addressee thinks and how it influences what she infers, or (iii) a putative thought which the addressee ascribes to the speaker.

Let us now answer RQ2. This in-depth analysis has shown that IB in discourse is a very complex phenomenon. Considering that what is instantiated in Table 5 fits within my definition of IB, it seems that stereotypes are ingredients of IB and can affect the processes involved in utterance interpretation, but that IB processes as such are orthogonal to the processes delineated in TCP. For this reason, I do not consider IB itself as part of the meaning of the utterance. But I do consider that analysing how IB arises in language raises two important points: first, that

[8] I consider the selection of the pronoun as an unconscious mechanism in this case, or at least a spontaneous one (see Footnote 7), because the referent is not yet known. However, one could also read (1a) as the speaker's overt preference for female nannies. The many readings these examples can receive are further evidence that context is needed to assess whether a stereotype is being conveyed.

stereotypes may unconsciously guide utterance production and interpretation, and second, that it is challenging to fully grasp the complexity of IB by only resorting to speaker intentions.

6 Taking stock: Implications for Future Research

This discussion has shown that adopting a linguistic perspective on implicit bias constitutes an interesting contribution to the debate, in that it forces us to think about the role played by stereotypes in utterance production and interpretation. I have characterised IB as an unconscious process which involves inserting stereotypes into discourse. Admittedly, when approached from a strictly philosophical point of view, what matters is not how the bias arises, but whether it is there and what one can do about it. But I would argue that looking into the processes which lead to the emergence of IB in language may illuminate the potential solutions to put in place to counter its adverse societal effects. Indeed, this observation, together with the fact that IB cannot reasonably be incorporated within speaker meaning, raises the question of whether the responsibility for conveying stereotypes lies with the individual or the collective.

The primary goal of the present paper is to shed light on the unconscious mechanisms which reveal biases in discourse, which means that I set aside the question of accountability. However, one might wonder whether the fact that TCP does not account for all the complexity of IB may be an argument in favour of Minimalism, especially for utterances such as (1), since the notion of literal meaning does not appeal to speaker intentions. I would strongly argue against this, as it would come down to deciding upfront which theory of meaning is adequate based on the outcomes one wishes to attain (i.e. making people accountable for the stereotypes they unknowingly convey). Therefore, rather than being an argument in favour of Minimalism, I view the analysis of IB as supportive of interactional accounts (see e.g. Elder & Haugh 2018 for an interactional analysis of hints), and more generally, of the view that the addressee's perspective should be given more weight in accounts of utterance meaning (see e.g. Hansen & Terkourafi 2023). Approaching IB from a conversation-analytic point of view may yield useful insights into the many ways stereotypes can insidiously surface in discourse.

Lastly, given the complexity of the phenomenon, one might wonder whether the processes underlying IB as it occurs in discourse resist experimental testing. After all, even finding adequate experimental paradigms for studying utterance interpretation is challenging (Carston 2007: 42; Recanati 2003: 14–15). One possible route to escape this conundrum is to only focus on the addressee. Indeed, instead of explicit statements focused on self-reflective judgement, as in the studies mentioned in Subsection 2.1, one could assess participants' interpretation of utterances involving stereotypes (with varying degrees of explicitness), so as to see if their interpretation correlates with the biases revealed in IAT. That said, clarifying the distinction between automatic and unconscious mechanisms, as well as establishing whether spontaneous processes can be consciously accessible – and therefore, inferential, according to TCP – will be crucial.

7 Conclusion

This paper has demonstrated how separate disciplines – in this case, philosophy and linguistics – can benefit from one another despite having distinct research goals. By applying a linguistic lens to the phenomenon of implicit bias, this work has shown the importance of taking the context, co-text, and addressee's perspective into account when trying to pinpoint how conversational participants (unwillingly) reveal stereotypes in discourse. The analysis in Section 5 also leads to the conclusion that the bias cannot, as such, be considered as part of speaker meaning. There is at most an overlap between IB and TCP processes, due to the fact that stereotypes can play a role in both.

References

Borg, E. 2004. *Minimal semantics*. Clarendon. doi:10.1093/0199270252.001.0001.

Borg, E. & P. J. Connolly. 2022. Exploring linguistic liability. In E. Lepore & D. Sosa (eds.), *Oxford studies in philosophy of language, volume 2*, 1–26. Oxford University Press.

Camp, E. 2018. Insinuation, common ground, and the conversational record. In D. Fogal, D. W. Harris & M. Moss (eds.), *New work on speech acts*, 40–66. Oxford University Press.

Carston, R. 2004. Explicature and semantics. In S. Davis & B. S. Gillon (eds.), *Semantics*, 817–845. Oxford University Press. doi:10.1093/oso/9780195136975.003.0040.

Carston, R. 2007. How many pragmatic systems are there? In M. J. Frápolli (ed.), *Saying, meaning and referring: Essays on françois recanati's philosophy of language*, Palgrave-Macmillan.

Carston, R. 2013. Word meaning, what is said, and explicature.

De Houwer, J. 2006. What are implicit measures and why are we using them? In R. Wiers & A. Stacy (eds.), *Handbook of implicit cognition and addiction*, 11–28. SAGE Publications, Inc. doi:10.4135/9781412976237.n2.

Elder, C.-H. & M. Haugh. 2018. The interactional achievement of speaker meaning: Toward a formal account of conversational inference. *Intercultural Pragmatics* 15(5). 593–625. doi:10.1515/ip-2018-0021.

Fazio, R. H. & M. A. Olson. 2003. Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology* 54(1). 297–327. doi:10.1146/annurev.psych.54.101601.145225.

Fricker, M. 2007. *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press. doi:10.1093/acprof:oso/9780198237907.001.0001.

Hahn, A., C. M. Judd, H. K. Hirsh & I. V. Blair. 2014. Awareness of implicit attitudes. *Journal of Experimental Psychology: General* 143(3). 1369–1392. doi:10.1037/a0035028.

Hansen, M.-B. M. & M. Terkourafi. 2023. We need to talk about hearer's meaning! *Journal of Pragmatics* 208. 99–114. doi:10.1016/j.pragma.2023.02.015.

Holroyd, J. 2015. Implicit bias, awareness and imperfect cognitions. *Consciousness and Cognition* 33. 511–523. doi:10.1016/j.concog.2014.08.024.

Holroyd, J. & J. Sweetman. 2016. The heterogeneity of implicit bias. In M. Brownstein & J. Saul (eds.), *Implicit bias and philosophy, volume 1*, 80–103. Oxford University Press. doi:10.1093/acprof:oso/9780198713241.003.0004.

Jaszczolt, K. M. 2006. Meaning merger: Pragmatic inference, defaults, and compositionality. *Intercultural Pragmatics* 3(2). doi:10.1515/IP.2006.012.

Jaszczolt, K. M. 2011. Default meanings, salient meanings, and automatic processing. In K. Jaszczolt & K. Allan (eds.), *Salience and defaults in utterance processing*, 11–35. De Gruyter Mouton. doi:10.1515/9783110270679.

Jaszczolt, K. M. 2015. Default semantics. In B. Heine & H. Narrog (eds.), *The oxford handbook of linguistic analysis*, 743–770. Oxford University Press. doi:10.1093/oxfordhb/9780199677078.013.0009.

Levinson, S. C. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. The MIT Press. doi:10.7551/mitpress/5526.001.0001.

Machery, E., L. Faucher & D. R. Kelly. 2010. On the alleged inadequacies of psychological explanations of racism. *Monist* 93(2). 228–254. doi:10.5840/monist201093214.

Mercier, H. & D. Sperber. 2017. *The enigma of reason: A new theory of human understanding*. Allen Lane.

Recanati, F. 2002. Does linguistic communication rest on inference? *Mind & Language* 17(1-2). 105–126. doi:10.1111/1468-0017.00191.

Recanati, F. 2003. *Literal meaning*. Cambridge University Press 1st edn. doi:10.1017/CBO9780511615382.

Recanati, F. 2007. Recanati's reply to carston. In M. J. Frápolli (ed.), *Saying, meaning and referring: Essays on françois recanati's philosophy of language*, Palgrave-Macmillan.

Saul, J. 2013. Scepticism and implicit bias. *Disputatio* 5(37). 243–263. doi:10.2478/disp-2013-0019.

Wilson, D. & D. Sperber. 2012. *Meaning and relevance*. Cambridge University Press.

Appendices

Example	Variants of (2)	Biased individual (i)	Unconscious process?		
(2)	I like my black friend Martin.	Speaker	Speaker	Modifying the noun ‘friend’ with ‘black’ appears unnecessary. However, more context is needed to rule out the possibility of referent disambiguation.	Yes, provided it is not a case of referent disambiguation.
(2b)	I like my black friend.	Speaker	Speaker	Similar to (2); without any additional context, (2b) lends itself less easily to the referent disambiguation reading. It seems to convey the idea that liking one’s black friend is somewhat special. That reading is emphasized by the fact that the friend in question is not given a proper name.	Yes
(2c)	A: I like my black friend Martin. B: Oh, I didn’t know he was called Martin.	Speaker$^?$ Addressee$^?$	Speaker$^?$ Addressee$^?$	Too little context: (2c) could either demonstrate that the fact that Martin is black does not matter to the interactants, or that none of them picked up the (potential) stereotype. Compare with (2d).	Unclear
(2d)	A: I like my black friend Martin. B: Oh, I didn’t know he was also called Martin.	/	/	Compared to (2)-(2b)-(2c), the disambiguation reading is stronger.	/
(2e)	A: I like my black friend. B: Yeah, he’s a nice chap.	Speaker Addressee	Speaker Addressee	Through the combination of elements of (2b) and (2c), (2e) exemplifies a clearer case of a stereotype being conveyed unconsciously and not being picked up by the addressee.	Yes

Table 1 Mechanisms at play and role of speaker/addressee in the revelation of stereotypes in examples based on utterance (2) ‘I like my black friend Martin.’ Underlined = Anaphoric reference, ? = Cannot be clearly revealed without additional turns

Example	Variants of (3)	Biased individual (i)	Revealer (ii)	Mechanism (iii)	Unconscious process?
(3)	You won't be happy living in this neighbourhood.	Speaker	/	More context and co-text needed.	Unclear
(3b)	A: You won't be happy living in this neighbourhood. B: Maybe you're right. I'd prefer my children to grow up with native speakers of English.	Speaker Addressee	Addressee	Unclear for the speaker: more context and co-text would be needed. The addressee's turn shows that a stereotype is affecting her interpretation of the utterance.	Yes (for addressee)
(3c)	A_{t1}: You won't be happy living in this neighbourhood. B: Maybe you're right. I'd prefer my children to grow up with native speakers of English. A_{t2}: Oh, I didn't mean to suggest that. I just meant that it would take you forever to go to work because of the traffic in that area.	Speaker Addressee	Addressee	Same process as (3b) – co-text helps to reveal the stereotype. A's denial of the stereotype in $_{t2}$ may be insincere, which is why (3c) is no evidence of the original speaker's (potential) bias.	Yes (for addressee)
(3d)	A_{t1}: You won't be happy living in this neighbourhood. B: Maybe you're right. I'd prefer my children to grow up with native speakers of English. A_{t2}: Yes, that's what I thought.	Speaker, Addressee	Speaker, Addressee	A's response to B in $_{t2}$ indicates that the stereotype pertaining to 'IMMIGRANT' = 'UNDESIRABLE' pervades through the exchange and that the interactants do not see it as problematic.	Yes, although it is harder to determine for speaker A as the stereotype may have been intentionally (and indirectly) conveyed.

Table 2 Mechanisms at play and role of speaker/addressee in the revelation of stereotypes in examples based on utterance (3) 'You won't be happy living in this neighbourhood.' **?** = Cannot be clearly revealed without additional turns, $_{t1}$ = First turn, $_{t2}$ = Second turn (same speaker as $_{t1}$)

Pamela Gitani
University of Cambridge
pgitani@gmail.com